数字人在证券客户创新服务模式方面的应用研究

1、课题背景

1.1 数字人简要介绍

数字人是通过建模、动作捕捉或 AI 等科技手段，制作出具有外貌特征和行为模式的虚拟形象，并通过显示设备呈现出来。数字人创造的价值主要是打破物理的空间限制，提供了更多沉浸感、参与感和互动感。

从概念上来说，数字人的范畴包含虚拟人，虚拟人的范畴包含虚拟数字人，三者概念存在细微差别。数字人强调角色存在于数字世界，数字人的身份设定可以是按照现实世界中的人物进行设定，外观也可以完全一致，按照真人还原制作的数字人也可以称为数字孪生；虚拟人的身份是虚构的且现实世界中不存在，虚拟人没有现实世界中的身体，它是通过计算机图形学技术进行虚拟制作的，通过显示设备呈现出来；虚拟数字人强调虚拟身份和数字化制作手段，存在于非物理世界中，由计算机图形学、图形渲染、动作捕捉、深度学习、语音合成等计算机手段创造及使用的产物。

从表现形式上看，数字人又可分为 2D 仿真数字人和 3D 建模数字人，如图 1 所示。目前，3D 建模数字人精美度高，但过高的建模成本和制作周期，导致其在商业化应用上存在一定难度，通常被用

作品牌大使等场景。但对于市场空间更大的直播带货、教育、客服、营销等场景来说，难以普及。尤其在 AIGC 发展迅速的现在，2D 仿真数字人的制作门槛、周期和成本远远低于 3D 建模数字人。并且，在市场空间较大的直播带货、娱乐主播、客服、营销活动、资讯播报和游戏 NPC 场景下，2D 仿真数字人已经足够满足当下的需求，爆发在即。

图 1 2D 数字人（左）和 3D 数字人（右）

本课题旨在自研 2D 仿真数字人技术，按照真人形象和音色还原制作“数字员工”、“数字投顾”等，并在公司内探索潜在的应用场景。

1.2 数字人应用场景

数字人的应用场景主要包括娱乐、偶像、代言人、企业数字化转型、体育、金融等垂直领域。未来数字人将在第一产业农业和第二产业工业领域中（如生产领域、销售领域、售后服务领域等）更多地被使用。本课题研究范围主要聚焦于金融行业，更具体地为，数字人在证券客户服务中的应用。

随着金融行业数字化转型的不断深入，客户的需求日益多样化和个性化，传统的金融服务模式已经无法满足客户的需求，需要借助数字人技术来提供更加智能、高效、个性化的“智慧”服务。诸多金融领域公司与第三方科技公司合作，打造技术能力应用较强的虚拟数字人员工。中信建投与腾讯云智能联合推出的“数字员工”，应用在了证券开户场景，如图 2 所示。

图 2 中信建投“数字员工”开户指引界面，形象与真人无异

本课题期望能自研并结合数字人技术，将文字、动画等传统的内容输出形式升级为“真人”形象讲解，有效提升内容吸引力与可信度。

1.3 数字人行业动态

随着元宇宙产业在全球迅速发展及 ChatGPT 的亮相，生成式 AI 算法的突破被市场点燃了规模化应用的热情。另一方面得益于构建虚拟数字人所依赖的 CG、动捕、AI 等技术逐渐成熟。2020 年至 2022 年期间全球数字人企业数量不断取得新突破从 12 家增至 44 家同比增长 2.7 倍。目前市场投融资最大的一笔融资来自 2021 年韩国元宇宙虚拟社交平台 ZEPETO 融资金额达 12.1 亿元。

为促进数字经济发展，加强数字中国建设整体布局，我国在 2019 至 2022 年间，出台多项政策鼓励支持相关产业发展，特别是 2021 年 10 月广电总局发布的《广播电视和网络视听“十四五”科技发展规划》中指出：“要推动虚拟主播、动画手语广泛应用于新闻播报、天气预报、综艺科教等节目生产，创新节目形态，提高制播效率和智能化水平”，首次明确地鼓励和支持数字人的发展。

近年来数字人竞争激烈，但真正拥有原创知识产权的机构并不多。据不完全统计，截至 2021 年底，中国机构在国内共申请了 1322 项数字人专利，其中高校申请超 200 项，互联网巨头申请超 110 项；共计 58 家机构获专利授权，企业涉及科技巨头、高校、数字人领域企业及银行。截至 2022 年底，成立仅 6 年的追一科技表现亮眼，其专利申请专利数量达 67 项。百度专利获批授权数量最多，目前共有 82 项专利数量。

2、数字人算法自研

对于多数金融领域公司来说，数字人技术往往通过与第三方供应商合作来获取，如前述中信建投与腾讯云智能联合推出的“数字员工”等。通过此种方式，金融公司可制作有限数量的数字人形象，若要做大面积推广，就会大大抬高推广成本。据悉，多数数字人技术供应商的收费标准，按单一形象和视频制作时长来收费，每增加额外的形象，或制作超额时长的视频，都会产生额外的费用。这对于大多数金融公司来说，要做“数字员工”、“数字投顾”等的大面积推广，高额的成本是无法接受的。

为实现系统技术不受外部影响，自主可控，并在大面积推广数字人形象的同时有效控制成本，本课题自研数字人算法，实现数字人对真人形象和音色的复刻。正如之前所提到的，2D 仿真数字人技术，可按照真人形象和音色进行还原。因此，本课题自研 2D 数字人技术，主要包含两部分，一是对真人音色的复刻，二是对真人形象的仿真，二者相结合，完成对真人的真实还原。

2.1 数字人语音复刻

数字人语音复刻，实际上由名为 TTS（Text To Speech，文字转语音）的语音合成技术实现。基于深度学习的语音合成技术，流水线包含文本前端（Text Frontend）、声学模型（Acoustic Model） 和 声码器（Vocoder） 三个主要模块：

- 文本前端模块将原始文本转换为字符/音素。

- 声学模型将字符/音素转换为声学特征，如线性频谱图、mel 频谱图、LPC 特征等。
- 声码器将声学特征转换为波形。

语音合成的流程图如图 3 所示。

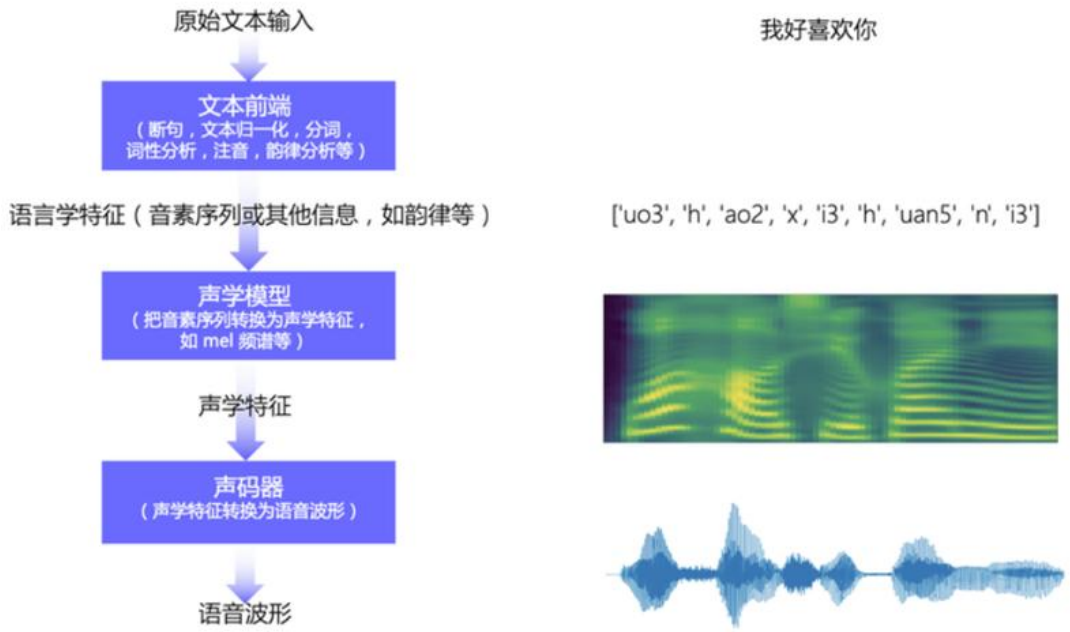


图 3 语音合成流程图

文本前端模块主要包含： 分段（Text Segmentation）、文本正则化（Text Normalization, TN）、分词（Word Segmentation）、词性标注（Part-of-Speech, PoS）、韵律预测（Prosody）和字音转换（Grapheme-to-Phoneme，G2P）等。其中最重要的模块是 文本正则化 模块和 字音转换模块。各模块输出示例：

- Text: 全国一共有 112 所 211 高校
- Text Normalization：全国一共有一百一十二所二一一高校
- Word Segmentation：全国/一共/有/一百一十二/所/二一一/高校/

◆ G2P（注意此句中“一”的读音）：

- quan2 guo2 yi2 gong4 you3 yi4 bai3 yi1 shi2 er4 suo3 er4 yao1 yao1 gao1 xiao4（可以进一步把声母和韵母分开）
- q uan2 g uo2 y i2 g ong4 y ou3 y i4 b ai3 y i1 sh i2 er4 s uo3 er4 y ao1 y ao1 g ao1 x iao4（把音调和声韵母分开）
- q uan g uo y i g ong y ou y i b ai y i sh i er s uo er y ao y ao g ao x iao
- 0 2 0 2 0 2 0 4 0 3 ...

◆ Prosody (prosodic words #1, prosodic phrases #2, intonation phrases #3, sentence #4)：全国#2 一共有#2 一百#1 一十二所#2 二一一#1 高校#（分词的结果一般是固定的，但是不同人习惯不同，可能有不同的韵律）

声学模型将字符/音素转换为声学特征，如线性频谱图、mel 频谱图、LPC 特征等。声学特征以“帧”为单位，一般一帧是 10ms 左右，一个音素一般对应 5~20 帧左右。声学模型需要解决的是“不等长序列间的映射问题”，“不等长”是指，同一个人发不同音素的持续时间不同，同一个人在不同时刻说同一句话的语速可能不同，对应各个音素的持续时间不同，不同人说话的特色不同，对应各个音素的持续时间不同。示例如下：

```
# 卡尔普陪外孙玩滑梯
```

```
000001|baker_corpus|sil 20 k 12 a2 4 er2 10 p 12 u3 12 p 9
ei2 9 uai4 15 s 11 uen1 12 uan2 14 h 10 ua2 11 t 15 i1 16
sil 20
```

声学模型主要分为自回归模型和非自回归模型。非自回归模型不存在预测上的依赖关系，预测时间快。本课题所用的基础模型即非自回归模型 FastSpeech2，如图 4 所示。

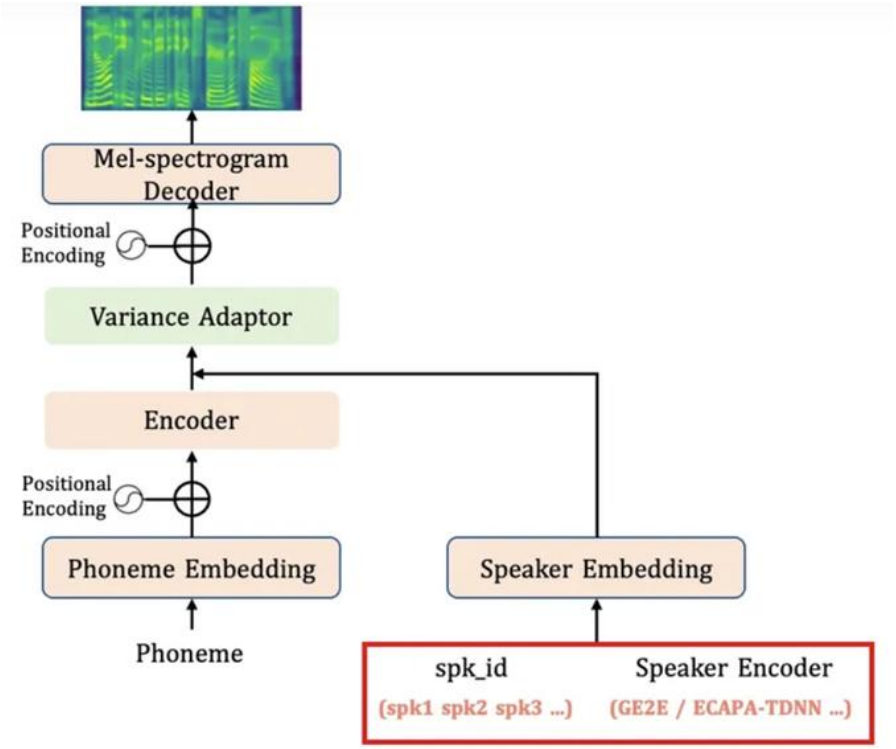


图 4 基于 FastSpeech2 的多说话人语音合成模型

声码器将声学特征转换为波形，需要解决的是“信息缺失的补全问题”。信息缺失是指，在音频波形转换为频谱图时，存在相位信息的缺失；在频谱图转换为 mel 频谱图时，存在频域压缩导致的信息缺失。假设音频的采样率是 16kHz，即 1s 的音频有 16000 个采样点，一帧的音频有 10ms，则 1s 中包含 100 帧，每一帧有 160 个采样点。声码器的作用就是将一个频谱帧变成音频波形的 160 个采样

点，所以声码器中一般会包含上采样模块。本课题采用的声码器是HiFiGAN，能够将声学模型产生的频谱转换为高质量的音频，这种声码器采用生成对抗网络（Generative Adversial Networks，GAN）作为基础生成模型，相比于之前相近的 MelGAN，贡献点主要在：

- 引入了多周期判别器（Multi-Period Discriminator，MPD）。HiFiGAN 同时拥有多尺度判别器（Multi-Scale Discriminator，MSD）和多周期判别器，目标就是尽可能增强 GAN 判别器甄别合成或真实音频的能力。
- 生成器中提出了多感受野融合模块。WaveNet 为了增大感受野，叠加带洞卷积，逐样本点生成，音质确实很好，但是也使得模型较大，推理速度较慢。HiFiGAN 则提出了一种残差结构，交替使用带洞卷积和普通卷积增大感受野，保证合成音质的同时，提高推理速度。

本课题数据采集及语音克隆流程如下：

1. 数据采集：需要克隆个人音色的投顾或直播路演人员，根据“中文音色克隆专用文档”+“英文音色克隆专用文档”，录制个人语音音频。
2. 数据处理成训练数据：对录制好的音频去杂质，自动切割，调整音量等，形成训练数据。
3. 声学模型训练：根据上述的模型结构进行各人员的专有音色声学模型训练。
4. 文本前端处理：针对中文多音字、英文多音字处理，对金融

专有名词的特殊读音进行处理等。

5. 语音预测：根据以上得到的文本前端处理模块及声学模型，对需要生成的具体内容进行语音生成，可得到特定音色的中英文语音音频结果。

2.2 数字人形象仿真

2D 数字人形象仿真，准确的来说，是基于数字人音色算法生成的人物播报音频结果，完成数字人“对嘴型”的过程。因此，这部分技术也可称为 2D 数字人唇形算法，该算法基于真实人物拍摄的视频，只改变人物嘴部区域动作，保留人物其余特征如面部容貌和肢体动作等，生成带有不同音频的数字人播报视频。

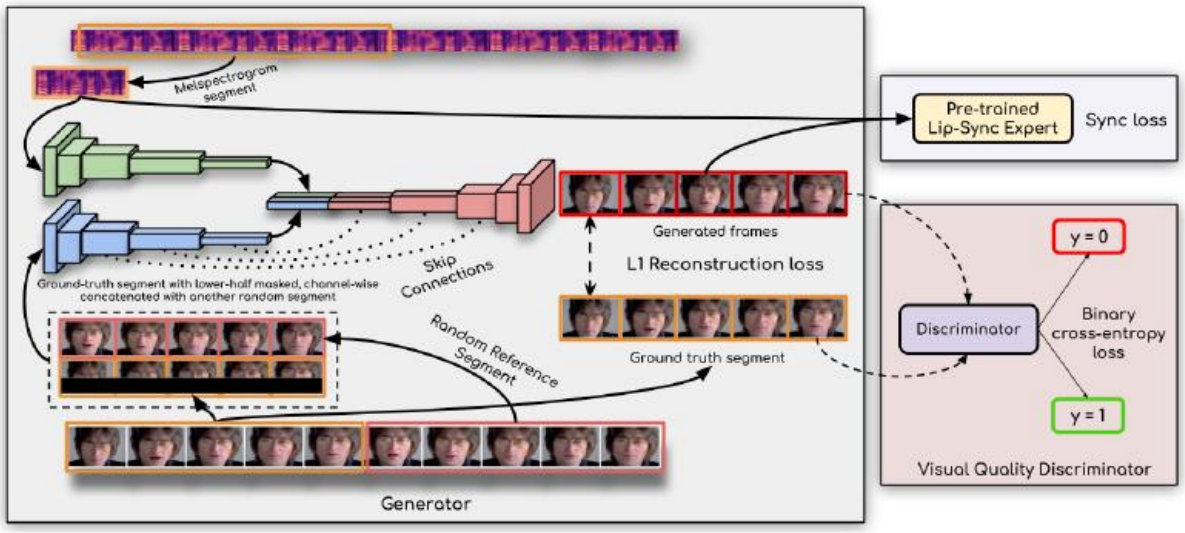


图 5 wav2lip 网络结构

在研究初期，我们调研、尝试了诸多 2D 数字人唇形算法。其中，比较经典且效果表现较好的算法是 wav2lip 算法。该算法是一个基于 GAN 的唇形动作迁移 AI 算法，网络结构如图 5 所示，算法基于真实人物播报视频进行训练，从而实现生成的视频人物口型与输

入语音同步。wav2lip 不仅可以基于静态图像来输出与目标语音匹配的唇形同步视频，还可以直接将动态的视频进行唇形转换，输出与输入语音匹配的视频，也即是前面说到的“对口型”。

但在测试过程中发现，原始 wav2lip 算法基于英文数据训练，在中文语音输入上表现不佳，唇形音频匹配度较差，其次，原始 wav2lip 算法基于低分辨率视频数据训练，导致生成的数字人视频分辨率较低，难以满足分辨率要求高的场景。为解决所述问题，我们着手收集中文且高分辨率视频数据，重新训练 wav2lip 算法。

wav2lip 模型的训练分为两个阶段，第一阶段是音频和口型同步判别器预训练，第二阶段是 GAN 网络训练。第一阶段预训练完毕后，在 GAN 训练过程中保持冻结。具体来说，wav2lip 的训练流程如下：首先，提取音频特征，将音频特征与人脸图像进行配对，形成一个音频-图像对，然后训练音频和口型同步判别器。接下来，wav2lip 使用 GAN 来学习音频-图像对之间的映射关系。生成器网络负责生成逼真的嘴唇动作，而判别器网络则负责评估生成的嘴唇动作的一致性和真实性，通过不断的训练和反馈，生成器网络逐渐学习到如何根据音频特征生成与之匹配的嘴唇动作。

我们收集了来自互联网的 720p（1280*720）视频数据约 25h，处理后可用数据约为 14h，以及来自公司培训中心的 1080p（1920*1080）视频数据约 10h，处理后可用数据约为 6h，最终总共可用数据约为 20h。除此之外，为了让 wav2lip 能适应高分辨率输入数据，我们还对 wav2lip 算法进行了改进，再基于自收集处理数据进

行训练。重新训练后测试结果显示，数字人唇形音频匹配度大幅提升，但生成视频清晰度提升有限。

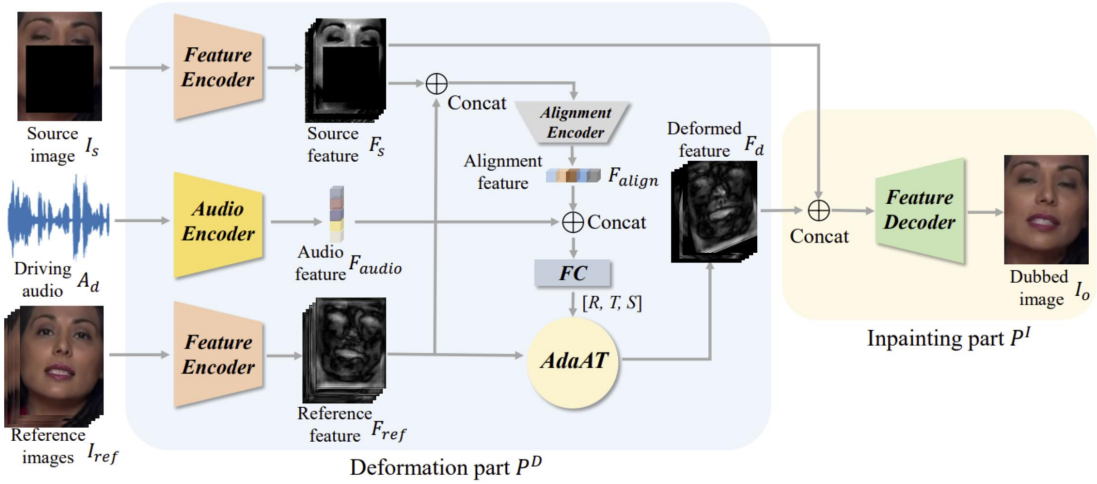


图 6 DINet 网络结构

经过不断尝试，我们最终将基础网络结构选型为 DINet（Deformation Inpainting Network）。为了解决生成视频清晰度问题，DINet 作者提出了一个应用于高分辨率人脸视觉配音的形变修复网络。该作者指出，前人工作依赖于多个上采样层直接从隐向量生成像素值，而 DINet 模型对参考图片的特征图采用空间形变的方式，更好地保留了高频纹理细节。具体地，DINet 模型由一个形变模块和一个修复模块组成，结构如图 6 所示。在形变模块中，为了与输入的驱动音频和原图像中的头部姿态对齐，对五张参考人脸图像采用自适应空间形变，从而在每一帧中产生编码着嘴部形状信息的形变特征图。在修复模块中，为了生成人脸视觉配音效果，一个特征解码器负责将形变后的特征图和来自原特征图的其它属性（头部姿态和脸部表情）自适应融合到嘴部。最后，DINet 实现了有着丰富纹理细节的人脸唇形驱动效果。

经过测试发现，原始 DINet 同样由于基于英文数据训练，在中文语音输入上表现不佳，但是输出结果清晰度较 wav2lip 有较大提升，因此我们打算对 DINet 进行重训练。除此之外，我们发现，原始 DINet 使用的音频特征更适合于英文音频，因此，我们选取了更适合中文的音频特征，并改造了 DINet 的网络结构以适应新特征的输入。最后，我们基于自收集的中文视频数据，完成了改造后 DINet 的训练。重训练后测试结果显示，数字人唇形音频匹配度较好，且生成视频清晰度较之前提升明显。

2.3 数字人算法流程总结

本自研数字人算法的流程可大致总结如下图 7：

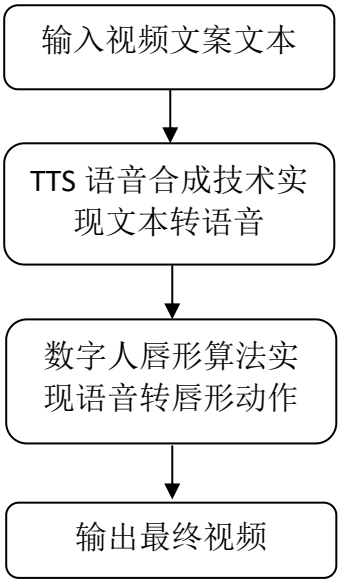


图 7 算法流程图

对于本自研数字人算法流程，接收一段视频文案文本，首先经过 TTS 语音合成技术实现文本转语音功能，得到一个音频结果，比如一个音频文件。接着，以前述所得音频结果和一段提前录制好的

真人出镜播报视频，共同作为数字人唇形算法输入，完成数字人“对嘴型”，得到一个视频结果，比如一个视频文件。最后，基于前述所得的音频结果和视频结果，即可合成得到最终的数字人播报视频，音频与人物唇形匹配度好，且保留了真人的形象和音色特征。

3、成果展示

3.1 数字人算法

为实现系统技术不受外部影响，自主可控，并在大面积推广数字人形象的同时有效控制成本，本课题完成了数字人算法自研，实现数字人对真人形象和音色的复刻。

为验证本课题自研数字人算法的效果，我们首先基于公司内部多位同事，制作了数字人播报 demo，如图 8 所示。

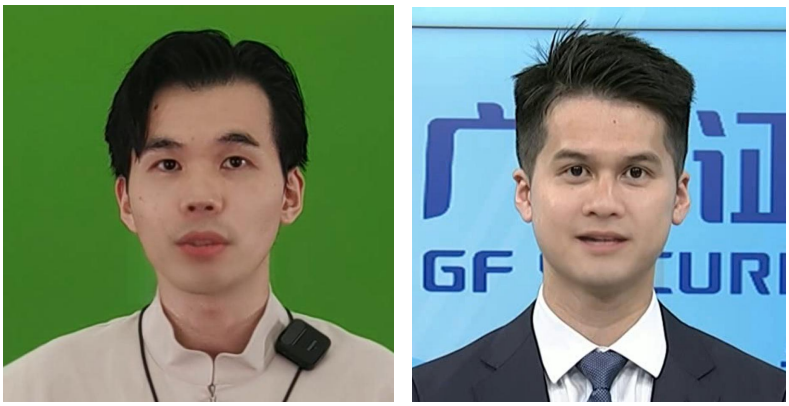


图 8 基于我司不同同事形象制作的数字人 demo

其次，我们邀请了三家外部厂商，基于同一位同事形象制作数字人播报 demo，并在公司内部发起了数字人效果盲评问卷。问卷评比参与人员为来自公司各个部门不同职级的同事，共计 59 人参与评比。问卷从整体自然度、唇形自然度、语音自然度、唇形语音匹配度等四个维度进行强制排名，并根据排名进行计分，问卷截图如图 9 所示。


数字人demo效果评审问卷

本调查问卷由 信息技术部[redacted]发出

特邀请您作为评价专家对3个外部厂商和信息技术部自研数字人效果进行评价，预计总耗时2-3分钟。感谢您的支持，谢谢！
请您在办公网环境下，通过电脑端观看以下视频：
数字人视频1
数字人视频2
数字人视频3
数字人视频4
如有任何疑问，可联系[redacted]进行咨询，谢谢！

1、 针对以上4个demo视频，在数字人整体自然度方面，您心目中的排名是（请勿在多个名次下选择同一个视频） * （每行最少选择1项）

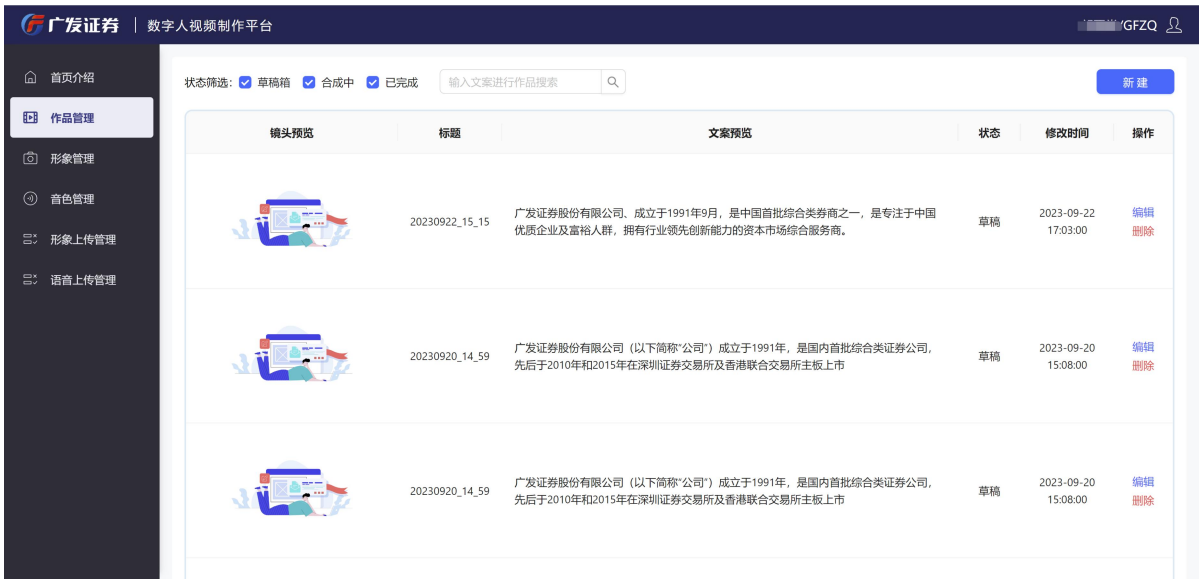
	数字人视频1	数字人视频2	数字人视频3	数字人视频4
第一名				
第二名				
第三名				

图 9 数字人 demo 效果评审问卷截图

问卷结果显示，本课题自研数字人算法，在四个维度上均排名第三名，且与第四名有较大分差，基于这一结果，我们总结如下：问卷结果说明本课题自研数字人效果与市场一流存在差距，但已经能战胜某些科技公司，效果属于可接受范围内。在综合考虑数字人制作成本、渲染成本、定制开发等因素后，本课题自研数字人算法更适合规模化应用（如制作数以百计的个性化数字人）和定制化场景（如对接内部信息系统、对接行情资讯等）。

3.2 数字人视频制作平台

本课题在除自研数字人算法外，还配套开发了一个供公司内部使用的数字人视频制作编辑平台，平台提供了数字人形象制作和管理、视频文案编辑、视频分镜、视频元素（如背景、音乐、特效）插入等功能，可供用户在平台进行数字人视频的制作及合成，如图10所示。



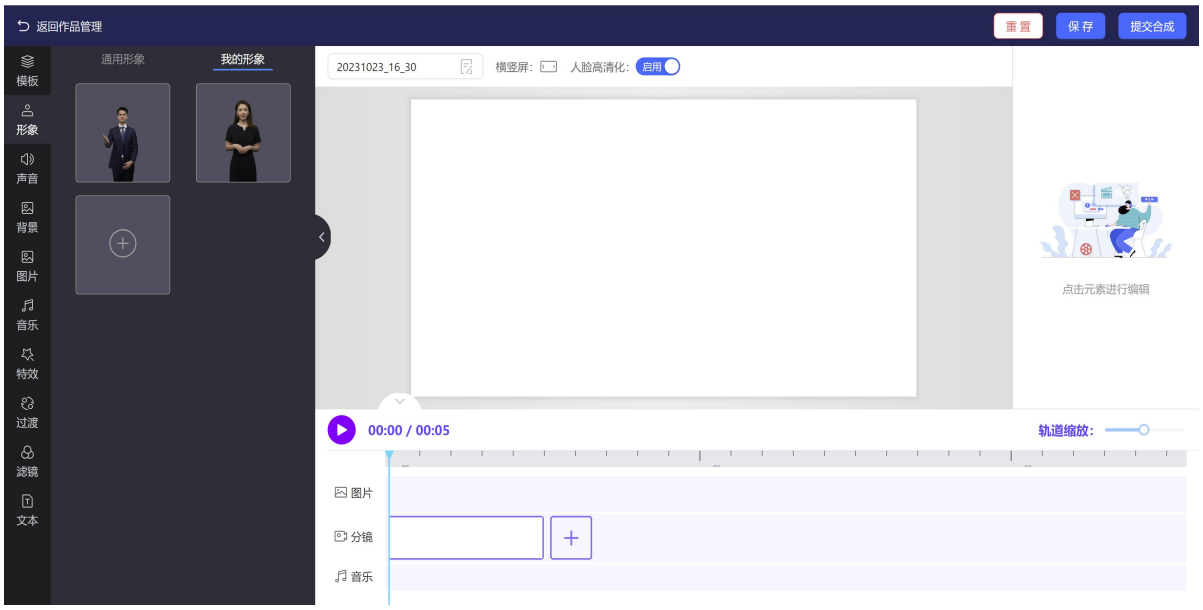


图 10 本课题自研数字人视频制作平台

平台具体功能介绍如下：

1. 视频制作平台：平台包含首页、作品管理、形象管理、语音管理四个模块。其中，首页展示了与用户视频作品制作相关的通知信息。

2. 形象管理：用户在此模块可增加或删除数字人物形象。用户可通过上传真人形象视频来增加可选的数字人物形象。

3. 音色管理：用户在此模块可增加或删除数字人物音色。用户可通过上传真人音色音频来增加可选的数字人物音色。

4. 作品管理：用户在此模块可查看数字人播报视频作品的相关信息，并进行数字人播报视频制作。制作视频时，用户需要选择数字人物形象和音色，可选择预置的或上传的各类素材（包括图片、背景、音乐），编辑视频内容、规格和文案，再提交作品合成。

5. 形象上传管理：此模块为管理员使用模块，管理在此处理用

户在形象管理模块上传的形象物料，并上传处理结果。

6. 语音上传管理：此模块为管理员使用模块，管理在此处理用户在音色管理模块上传的语音物料，并上传处理结果。

3.3 数字人应用信创部署

随着信创产业布局的全面铺开，“信创+智慧”的组合方式也不再陌生。信创产业作为战略性新兴产业，国家不断出台相关政策对行业的发展进行支持，国产化进程稳步推进。因此，各行业智慧产业不断发展的同时，引进信创支持是时代发展和政策支持下的必然结果。

公司积极响应国家号召，已经完成了各类信创产品的选型工作，并已建立起信创容器云平台、信创网络平台、信创数据库平台等的信创基础设施。目前，我司新一代信创容器云平台基于国产X86、ARM 架构芯片，部署鲲鹏服务器、海光服务器合计超 400 台，为业务应用提供全信创资源，完成了全信创基础软硬件的国产化改造，夯实了业务信息系统安全底座。平台支撑办公、中台、交易等几大基础类别业务，实现了数十个业务应用的正常运行，推动信创工作加速快跑，为后续持续国产应用上线提供良好基础和经验。

本课题自研数字人算法及视频编辑平台，形成数字人视频制作系统，部署于我司新一代信创容器云平台，部署架构图如图 11 所

示。本系统基于所述信创基础设施及资源，逐步推进，完成部署。

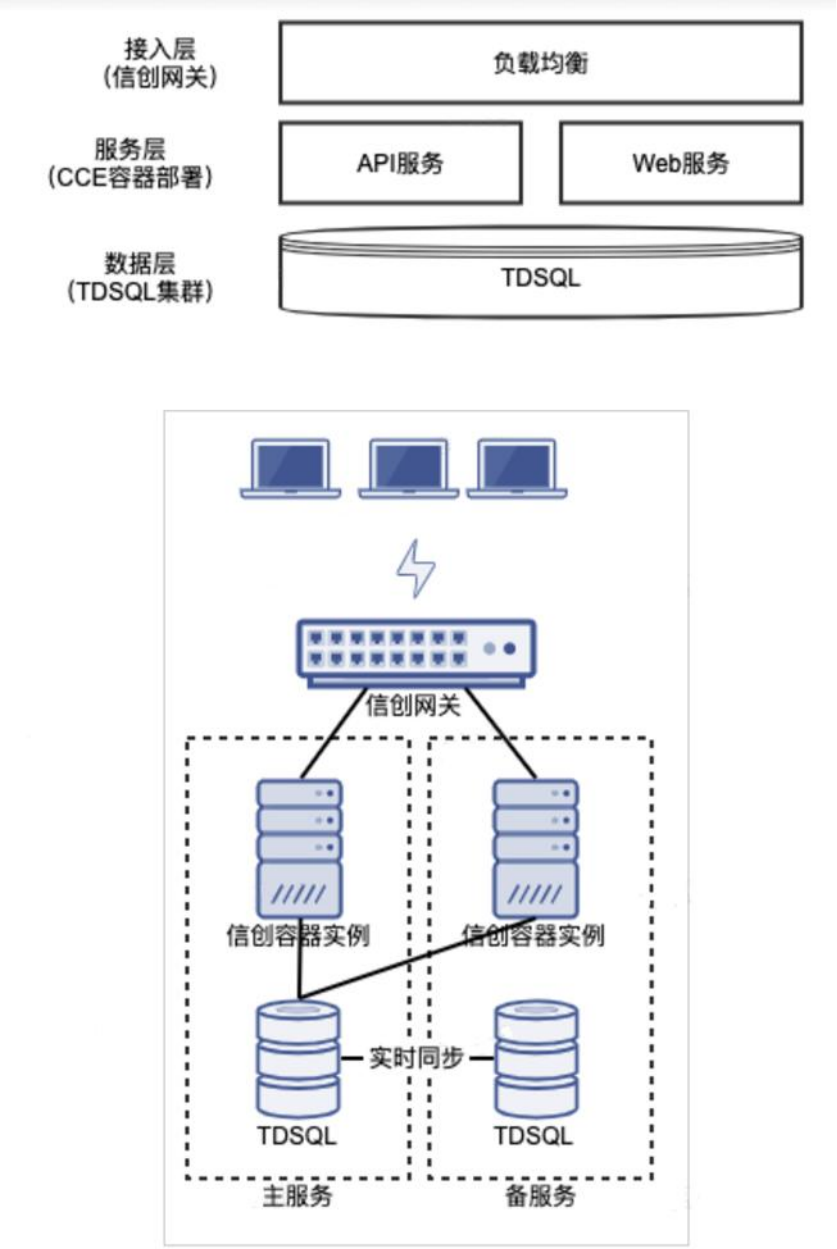


图 11 本课题数字人视频制作系统部署架构图

3.4 数字人应用实践

基于自研数字人技术，目前已成功制作了公司两位明星投顾老师的数字人形象，并应用在了公司运营管理部的企微收盘点评中，效果如下图 12 所示。

图 12 本课题自研数字人技术在企微收盘点评中的应用

在企微收盘点评的场景中，数字人播报视频挂载在收盘点评文

章页面，用户可选择点击观看，如图 13 所示。

图 13 收盘点评界面

上述界面统计了数字人视频点击量，如图 14 所示，统计时间范围从 2023 年 8 月 18 日至 2023 年 11 月 4 日，除去假期和周末时间，视频点击总量约为 3.89 万次。

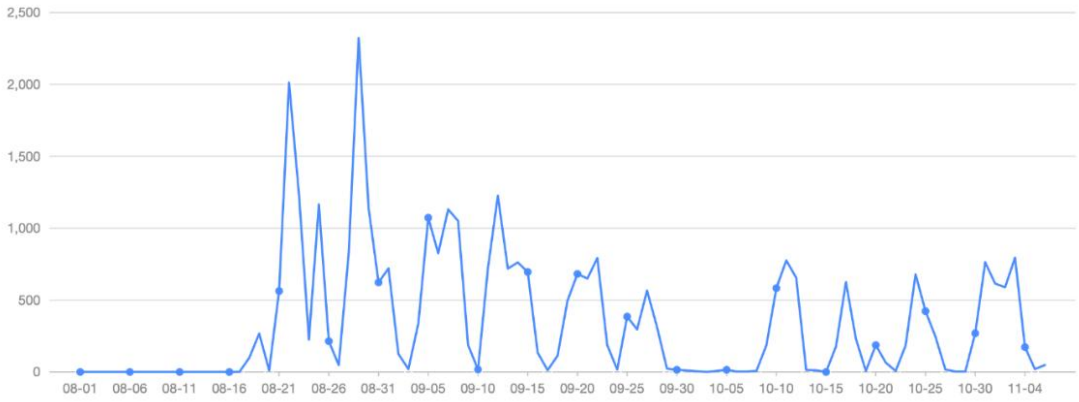


图 14 企微收盘点评中数字人视频点击量统计

除此企微收盘点评的场景之外，我们还拟将自研数字人算法应用到账户分析的场景中，如图 15 所示。在该场景中，可针对不同用户，生成个性化的数字人账户分析视频。如图 15，气泡元素里的数据，是由后台通过接口动态获取到的用户数据。再结合数字人播报视频合成，形成最终视频。

图 15 数字人账户分析

随着数字人视频制作系统上线、迭代及稳定运行，我们将探索数字人在更多场景下的应用。

4、课题研究总结

为实现系统技术不受外部影响，自主可控，并在大面积推广数字人形象的同时有效控制成本，本课题完成了数字人算法自研，并

配套开发了数字人视频编辑平台，形成了数字人视频制作系统，并部署在公司新一代信创容器云平台。本课题自研数字人算法，与市场一流存在差距，但已经能战胜某些科技公司，效果属于可接受范围内。在综合考虑数字人制作成本、渲染成本、定制开发等因素后，本课题自研数字人算法在规模化应用场景（如制作数以百计的个性化数字人）和定制化场景（如对接内部信息系统、对接行情资讯等）更具优势。目前数字人视频制作系统已在部分场景展开试用，且未出较大异常情况，随着未来的不断迭代及稳定运行，我们将探索数字人在更多场景下的应用。