论文题目：以实时声音情感驱动的虚拟形象

编写者：Jun Hanaizumi

单目相机动捕技术的发展已非常成熟，如谷歌开发的MediaPipe可从移动端相机实时捕捉面部、肢体、双手等特征点位置，该工具贡献了虚拟形象应用的发展。虚拟形象的使用场景主要为线上会议、直播、元宇宙内的娱乐活动等。

目前的虚拟形象随动技术通常让虚拟模型直接模仿捕捉的人体姿态，该方法直观且易用，但会被拍摄范围、用户的肢体活动自由度在可用动作上受到限制。特别在电脑上，用户因坐姿使用，基本上无法让虚拟形象做躯干以下的动作，难以模仿站姿的虚拟交流场景。目前，这些需求的解决方式为通过按键输入的方式播放用户设置的人物动画，但在交流途中按键去控制动作带来了操作不便的问题。因此，本研究希望从人的声音带有的情感作为动作控制的输入方案。

人声情感控制动作的优势为，人声中的情感至少可被神经网络分类成24种语义上的不同种类，具有充分的潜力调用多样化的动作。此外，人声情感是帧单位的实时检测，只要用户正在说话就可以自动播放对应情感的人物动作，这种随声驱动的方式提供更流畅的使用体验。

当前的人声情感识别技术中，较为成熟的为HumeAI的基于24情感维度空间的人声情感识别深度网络。该网络的构建中使用了包括中英文在内的5种语言的语料库，并且包含不同文化下的非语言语气词（Vocal Bursts）的感情标签，在感情分类数量、模型泛化程度与对非语言语气词的支持上另HumeAI在同行内的研究中尤为突出。HumeAI提供Python SDK，能调用人声感情识别模块，可实时检测语音片段的情感可信度，可以输出离散的48种情感标签。

预计完成的本项目系统使用流程里，用户先要在各感情标签上绑定人物动画文件，这些动画文件可在Mixamo等网站找到充分的情感动作可供下载。然后，用户启动本项目的动捕系统，通过设备相机开始捕捉用户姿态，此时用户可以尝试进行说话。本项目会采集用户的语音中的语调、语速、强弱等声学特征，并映射至感情标签上，如果该标签上绑定了人物动画，虚拟形象将会跟随动捕姿态的同时播放人物动画。虚拟形象跟随动捕姿态还是人物动画的规则可在本项目内的用户偏好设定，例如双手不在拍摄范围内时可以优先人物动画内的动作。

本项目实现方法为使用已完成的基于MediaPipe Hands, MediaPipe Faces的随动虚拟形象Unity项目，添加Python端的人声情感识别模块，该模块通过HumeAI的PythonSDK实现。Python端模块将使用设备麦克风采集人声，并把各情感可信度列表以Socket方式发送至Unity，Unity上驱动的虚拟形象会在接受到绑定过动画的人声情感，可信度超过阈值同时不在播放其它动画时，将播放对应动画。

该项目展示将使用站姿的全身虚拟形象，通过坐姿的电脑用户控制，相机与麦克风都使用电脑的内置设备。为方便展示，本项目默认在主要的感情上绑定了对应的情感动作。测试环节预计召集多名志愿者展开，本项目系统收集反馈延迟、识别精度评价，用于成果总结。