嘴唇识别

视听识别（AVR）被认为是一种**解决音频损坏时**语音识别任务的方法，也是一种用于多说话人场景中说话人验证的视觉识别方法。

**AVR系统的方法是利用从一个模态中提取的信息，通过对缺失信息的补充来提高另一模态的识别能力。其根本问题是找到音频流和视频流之间的对应关系**

**我们建议使用一个耦合的三维卷积神经网络（3D-CNN）架构，它可以将两种模式映射到一个表示空间中，使用所学习的多模态特征来评估视听流的对应性。**提出的架构**将空间和时间信息结合起来，有效地找到不同模式的时间信息之间的相关性。**

通过使用相对较小的网络结构和较小的训练数据集，我们提出的方法优于现有的类似的基于三维CNN的特征表示的视听匹配方法。

基于常识的推论是，由于人类通常能够**将听到的声音与给定的一组嘴唇运动进行匹配**，因此嘴唇运动与由语音特征表示的听到的声音高度相关。然而，视觉嘴唇运动及其相应的音频流仍然可以具有不可忽略的不相关信息。决策融合已被证明是有效的，其中，**最后的决定是通过融合不同模式的统计独立决定，强调不同模式之间的不相关特征**[4]。然而，在早期阶段的数据融合，显示出更大的前景，因为它创建了一个联合表示基于跨模态相关性的两种模态之间的关系。

然而，**目前还缺乏同时融合时空视听信息来解决音频流和视频流是否匹配的根本问题的研究**。例如，在多说话人场景中，如果可以找到连接音频和视频的特征，则说话人的嘴唇运动可以由音频流确定，反之亦然。

为了解决这个问题，我们建议使用最近被用于**动作识别、场景理解和说话人验证**的**三维卷积神经网络模型**，并证明了它的应用前景

**三维CNNs同时从空间和时间两个维度提取特征**，**从而在相邻帧中捕获和连接运动信息**。我们使用3D CNNs从输入帧生成独立的信息通道。所有相关信息通道的组合创建最终的特征表示。本文研究的重点是实**现两个不完全相同的三维cnn进行音视频匹配**（第五节）。我们的目标是设计**非线性映射**，学习非线性嵌入空间之间的相应的音频和视频流使用一个简单的方法

方法核心：

•一种新的**耦合3D CNN架构**，**它同时提取空间和时间信息，与输入空间相比，其设计具有显著的降维，并且针对区分匹配和非匹配视听流进行了优化。**

•网络相对较小，其优点是易于培训和快速测试。

•与传统的MFCC相比，使用了不同类型的语音特征来表示音频流，这提供了更具前景的结果。

•提出了一种以输出特征空间距离为准则的自适应在线配对选择方法，用于选择主要贡献配对，以加快收敛速度，防止过拟合。据我们所知，这是首次尝试使用三维卷积神经网络进行视听匹配，其中建立了时空特征之间的桥梁，以建立视听模式之间的公共特征空间。