Q：如何将声纹识别与口型识别相匹配？

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，**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**。

**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**

**有些学者的初步探索：**

**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**

**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**

**—》**设计**非线性映射**

**—》与输入空间相比，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**

Q：嘴唇部分的实现

**OpenCV的接口可以实现人脸定位。**

**cvHaarDetectObjects是opencv1中的函数，opencv2中人脸检测使用的是 detectMultiScale函数。它可以检测出图片中所有的人脸，并将人脸用vector保存各个人脸的坐标、大小（用矩形表示），函数由分类器对象调用：**

**Q：ivector**

因为从UBM模型自适应到每个说话人的GMM模型时，只改变均值，对于权重和协方差不做任何调整，所以说话人的信息大部分都蕴含在GMM的均值里面。GMM均值矢量中，除了绝大部分的说话人信息之外，也包含了信道信息。联合因子分析(Joint Factor Analysis, JFA)可以对**说话人差异和信道差异**分别建模，从而可以很好的对信道差异进行补偿，提高系统表现。但由于JFA需要大量不同通道的训练语料，获取困难，并且计算复杂，所以难以投入实际使用。由Dehak提出的，**基于I-Vector因子分析技术**，提出了全新的解决方法。**JFA方法是对说话人差异空间以与信道差异空间分别建模，而基于I-Vector的方法是对全局差异进行建模**，将其二者作为一个整体进行建模，这样处理放宽了对训练语料的限制，并且计算简单，性能也相当。

Q：GMM-UBM

**GMM(高斯混合模型)。它将空间分布的概率密度用多个高斯概率密度函数的加权来拟合，可以平滑的逼近任意形状的概率密度函数，**并且是一个易于处理的参数模型，具备对实际数据极强的表征力。但反过来，GMM规模越庞大，表征力越强，其负面效应也会越明显：参数规模也会等比例的膨胀，需要更多的数据来驱动GMM的参数训练才能得到一个更加通用（或称泛化）的GMM模型。

GMM-UBM模型最重要的优势就**是通过MAP算法对模型参数进行估计**，避免了过拟合的发生，同时我们**不必调整目标用户GMM的所有参数（权重，均值，方差）只需要对各个高斯成分的均值参数进行估计**，就能实现最好的识别性能。