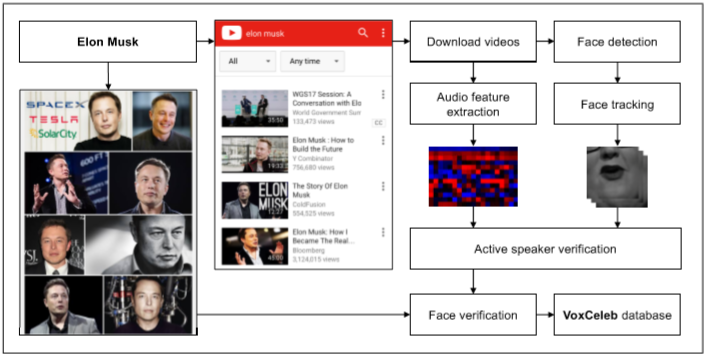
# 声纹识别

## 语音处理

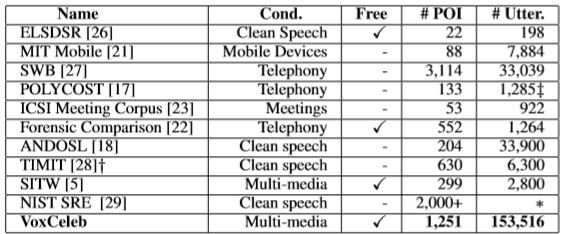
### VoxCeleb

* 提出一个全自动和可扩展的管道，以创建大规模的“现实世界”说话者识别数据集。通过使用视觉主动说话人识别和面部验证，我们的方法完全避免了对人类注释的需求。我们使用这种方法来整理VoxCeleb，这是一个大规模的数据集，具有针对一千名演讲者的数百种语音。
* 研究通过很少的预处理就可以直接从原始音频文件中提取出的深CNN声谱图训练方法的不同体系结构和技术，并使用传统的最新方法将我们在新数据集上的结果进行比较。

### 数据集如何产生



### 现有的说话人识别数据集



### 输入处理

* 频率 <-- 采样率

Fn=(n−1)∗Fs/N

* Fs(采样率)为16KHz，表示一秒钟内采样16000个点;

N，表示对多少个点做FFT,如果一帧里面的点的个数小于N就会zero-padding 到N的长度;

* 每个点对应一个频率点，某一点n（n从1开始）表示的频率为Fn=(n−1)∗Fs/N，第一个点（n=1，Fn等于0）表示直流信号，最后一个点N的下一个点（实际上这个点是不存在的）表示采样频率Fs。
* n\_fft = 512 (25ms\*16000)

hop\_length = 256 (10ms\*16000)

* 512 x 300 中的 512 ？

## 模型

### 高斯混合模型（GMM）

* 缺点

依赖于音频输入的低维表示，例如梅尔频率倒谱系数（MFCC）。但是，不仅MFCC的性能在现实世界的噪声中会迅速降低[11，12]，而且通过仅关注短帧的整个频谱包络，MFCC可能会缺少扬声器区分特征（例如\*\*信息）。

* 转变

深层CNN领域

要求：现实条件下获得的大型数据集

### CNN架构

* 我们的目标是从需要传统手工功能的技术过渡到可以选择说话者识别任务所需功能的CNN架构。这使我们可以最大程度地减少音频数据的预处理，从而避免在处理过程中丢失有价值的信息。输入功能。为了保持一致性，首先将所有音频以16kHz采样率转换为单通道16位流。然后，使用宽度为25ms，步长为10ms的汉明窗以滑动窗口的方式生成频谱图。这样，语音3秒钟的频谱图大小为512 x 300。在频谱的每个频率段上均值和方差归一化。如表7所示，这种归一化至关重要，可将分类准确率提高近10％，而没有使用其他语音特定的预处理（例如，静音消除，语音活动检测或清音消除）。这些短时幅频谱图被用作CNN的输入。建筑。由于在封闭集合下的说话人识别可以被视为多类分类问题，因此我们基于VGG-M [37] CNN的体系结构，以图像数据的良好分类性能而著称，并进行了修改以适应声谱图输入。尺寸为9×8的完全连接的fc6层（在两个尺寸上均受支撑）被两层取代-完全连接的9×1层（在频域中受支撑）和具有支撑1×n的平均池层，其中n取决于输入音段的长度（例如3秒）段，n = 8）。这使得网络对于时间位置不变，但对于频率不变，并且同时保持了与原始连接层相同的输出尺寸。这也将参数数量从VGG-M中的319M减少到我们网络中的67M，这有助于避免过度拟合。完整的CNN架构在表4中指定。

### ResNet