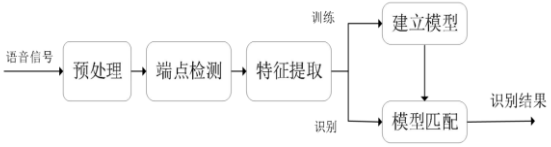
声纹识别模型综述

声纹识别任务包括

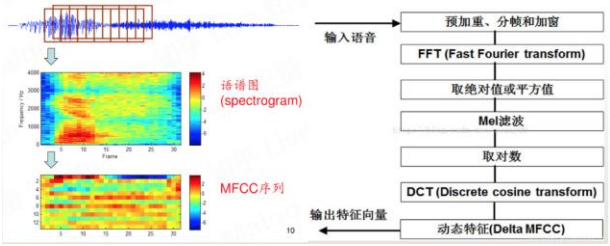
- 说话人辨认：在多人语音中，辨认某一段语音是谁说的；
- 说话人确认：在一段语音和一个人的情况下，确认它是不是他说的；
- 文本无关：对每个人的发音内容没有要求；
- 文本相关：要求训练数据与测试数据的内容相同或前者包含后者；
- 开集辨认：假定待识别说话人可以在集合外；
- 闭集辨认：假定待识别说话人在集合内；

本工程实践的应用实际当属于**文本无关的开集说话人辨认**任务。

声纹识别的基本流程：

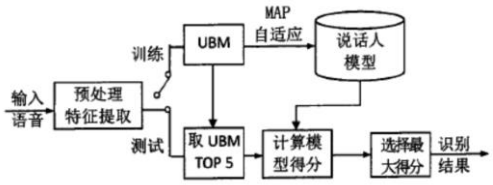


特征提取：
经典方法是 MFCC，还包括语音端点检测、预加重、分帧、加窗和提取声纹模型参数几个步骤。



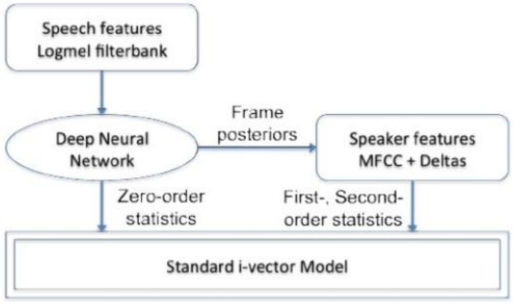
- **GMM_UBM（高斯混合模型-通用背景模型，对单纯 GMM 的改进，2000 年提出）：**

GMM 仍然是常用的简单模型之一，而 UBM 通过最大后验估计(Maximum A Posterior, MAP)的算法对模型参数进行估计，避免过拟合的发生、减少了训练时对说话人语音数据量的需求。

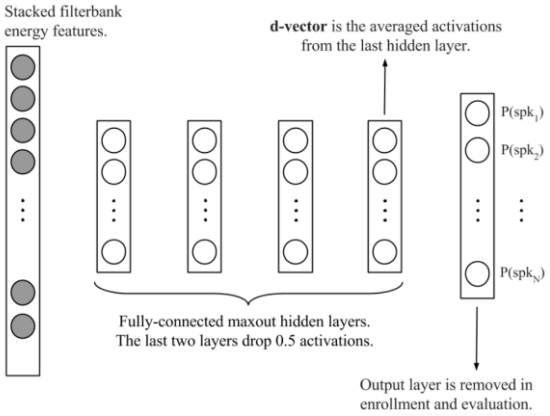


其缺陷在于仍需要较多参数和训练数据、不抗噪声和设备干扰。

- **i-vector 及其延申（2009 年）**
  将说话人的声纹信息和信道信息映射到低维 i-vector 上，利用其空间方向区分性，使用各种聚类方法实现说话人识别，典型的如 DNN/i-vector、TVM-i-vector（Total Variability Modeling 全局差异空间建模）



- **d-vector （《Deep neural networks for small footprint text-dependent speaker verification》，2014 年提出）**
  深度网络的特征提取层（隐藏层）输出帧级别的说话人特征，将其以合并平均的方式得到句子级别的表示，这种语篇级的表示即深度说话人向量，简称 d-vector。计算两个 d-vectors 之间的余弦距离，得到判决打分。类似主流的概率统计模型 i-vector，可以通过引入一些正则化方法（线性判别分析 LDA、概率线性判别分析 PLDA 等），以提高 d-vector 的说话人区分性。



- **x-vector Embedding（《X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION》，2018 年提出）**
  使用数据增广来提高深度神经网络（DNN）embedding 对于说话人识别的性能。经过训练以区分说话者的 DNN 将可变长度的语料映射到我们称为 x-vector 的固定维度 embedding。

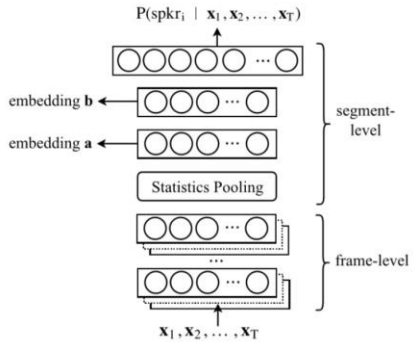


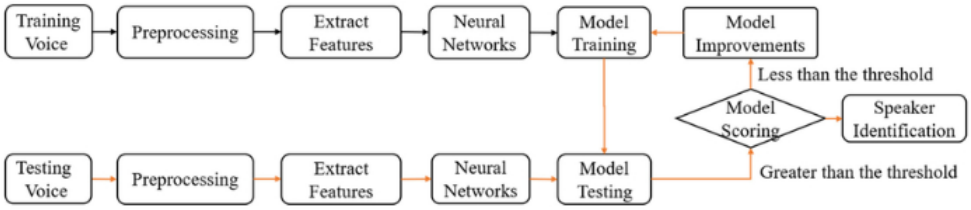
Figure 1: *Diagram of the DNN. Segment-level embeddings (e.g.,* **a** *or* **b***) can be extracted from any layer of the network after the statistics pooling layer.*

有点事有效且训练稳定，缺点是

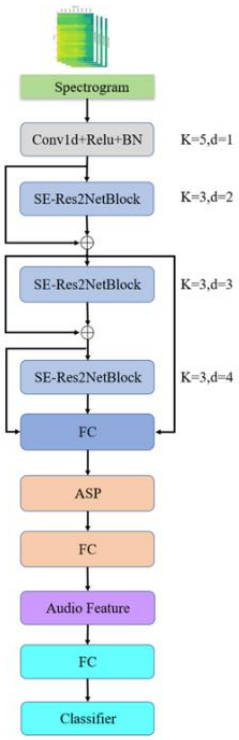
- 训练标签需要精确到每句话对应的说话人身份；
- 使用 Softmax 不一定是最优的损失函数；
- 输出层随着说话人数量增大而变大，因而不适应与变化较多的开集任务；

- **ECAPA-TDNN 及其改进（2023 年提出，该网络能够应对更复杂的声学环境和说话人识别任务，但也更复杂）**

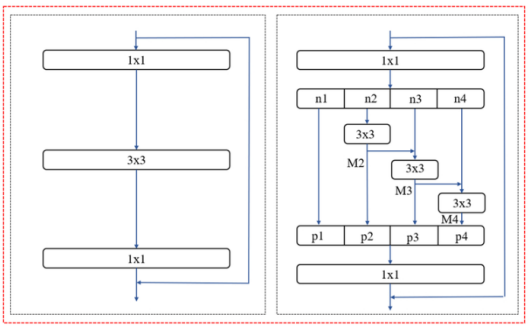
整体结构



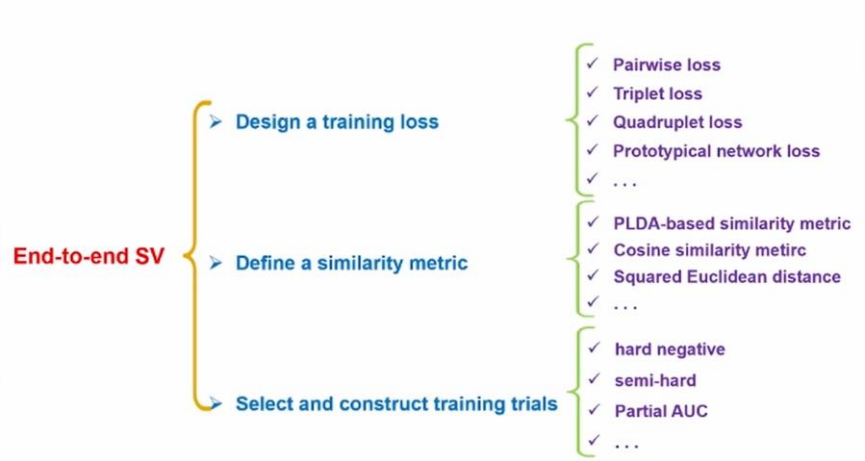
识别模型（神经网络）



ECAPA-TDNN 本质上是一个残差网络，其主要结构是 Res2Net。层残差的连接使 Res2Net 拥有比 ResNet 更大的感知野和更多感知野的组合。相比之下，并行分支结构使 Res2Net 拥有比 ResNet 更少的参数。Res2Net 内部的残差块结构和 ResNet 残差块结构如下。左 ResNet 残差块结构，右图显了 Res2Net 残差块的结构。



- **End-to-end（2019 年至今的多种模型，关键点集中在损失函数确认、定义相似度指标、选择和构建训练方法）**



有学者提出了基于 LSTM 神经网络的说话人识别模型，对比三元组、TE2E、GE2E 三个损失函数，起的了较好的准确率。

- 端到端的总结:
- 非端到端的分类损失需要引入减小类内方差的 margin
- 端到端确认损失引入类中心学习可以增加训练稳定性、提高性能