一种说话人分割方法

王东 李蓝天

背景

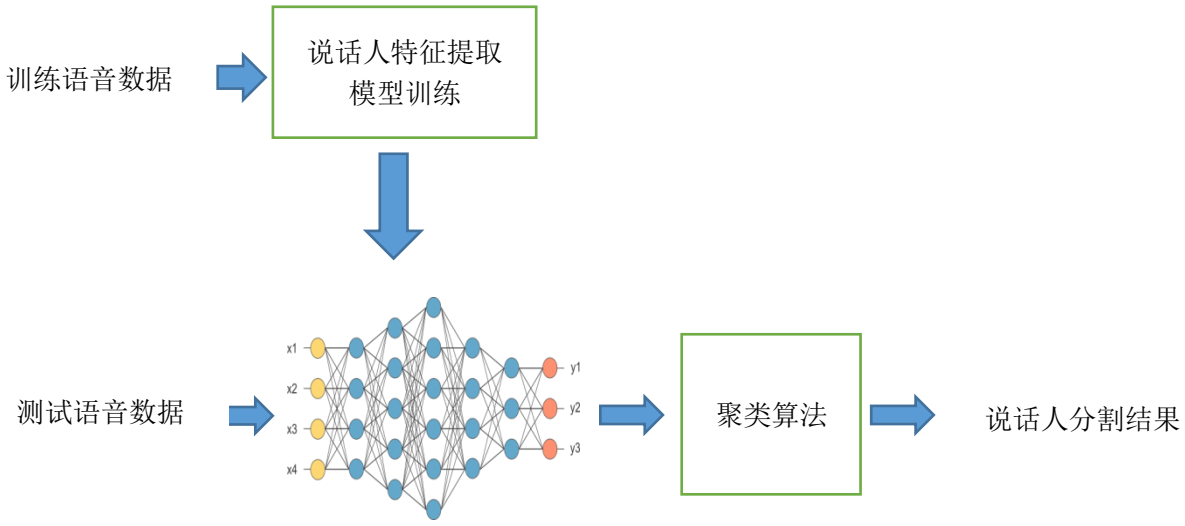
说话人分割是将一段声音信号按说话人进行时序分割的技术，如将客服电话录音分成客服和客户声音，将会议录音分成不同说话片段。说话人分割对通话质量检测、场景分析、语音识别自适应等都有重要实用价值。说话人分割可分为（1）确定性说话人分割和（2）非确定性说话人分割。确定性说话人分割已知说话人数，并将语音信号按此人数进行分割，如在电话语音中，说话人数确定为两个，因此只需将每一个语音帧分到两个说话人中的一个即可；非确定性说话人分割中，语音中的说话人个数未知，需要对说话人个数进行判断，再将语音帧分到相应说话人中。确定说话人分割有时也称说话人切分（Speaker Segmentation），非确定说话人分割一般称为说话人分析（Speaker Diarization）。

传统说话人分割方法分为两种：一种通过检测说话人切换点（Speaker Turn），将语音信号切分成仅包含一个说话人发音的句子，再通过聚类方法将属于同一个人的句子聚成一类；另一种方法是模型法，首先对每个说话人训练单独模型，再依某一语音帧对每个模型的概率将该语音帧分到某个说话人模型。这两种方法的显著缺陷是他们都基于初级声学特征（如 MFCC），因此说话人信息混杂在说话内容信息之中，故而分割稳定性不强，对参数选择敏感。

发明内容

本发明提出一种基于深度说话人特征与聚类算法相结合的说话人分割方法，对说话人切分和说话人分析都有良好效果。该发明的基本思路是：利用深度神经网络从初级声学特征中提取和说话人相关的特征，由于该特征可以有效表示说话人属性，用一个简单的聚类算法即可实现分割，因而极大降低模型复杂度，提高说话人分割系统的正确率和时间解析度。

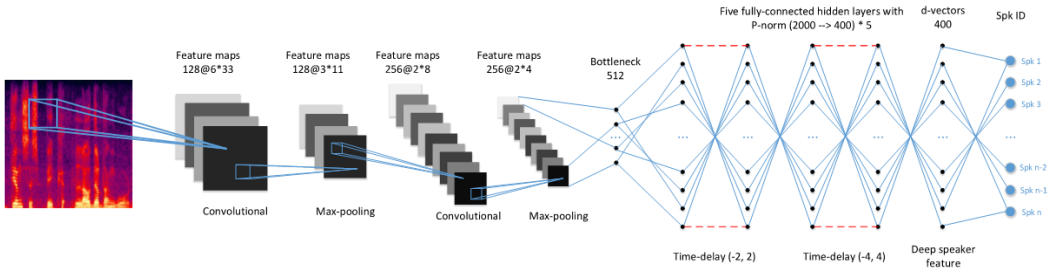
该发明所提出的说话人分割系统分为训练和运行两部分。在训练阶段，我们用一组说话人语音数据训练一个对说话人特征提取的深度神经网络，该网络可以在短时语音帧上（0.3 秒左右）提取说话人特征；在运行阶段，利用该网络对测试语音每一个短时语音帧提取说话人特征，并基于这些特征进行帧级别的聚类，从而完成说话人分割任务。注意，基于初级声学特征是无法通过聚类实现说话人分割的，因为每个特征间的距离并不能代表说话人之间的距离，而是包含了说话人内容、信道等多种信息之间的总体距离。



说话人特征提取网络

模型训练过程

本发明中，说话人特征提取模型为深度神经网络，包括 N 层卷积层，M 层延时层，K 层全连接层。网络的输入结点对应一个初级语音特征帧，输出结点对应训练集中所有说话人，训练的目标是交叉熵。经过训练以后，给定一个初级语音特征帧，经过网络映射之后，即得到该帧属于每个目标说话人的概率。我们取最后一个隐藏层的输出为说话人特征。一个该特征提取网络的实现特例如下图所示。

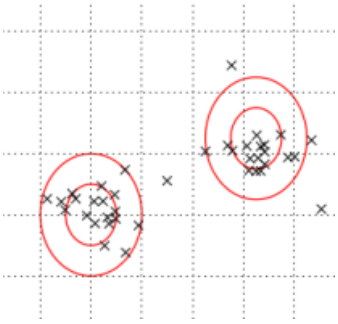


说话人分割过程

基于提取的说话人特征序列，可利用各种聚类算法实现说话人分割。下面以 EM 算法作为实现特例对分割过程做说明。

EM 算法假设每一个说话人的说话人特征符合高斯分布，因此将可将整个说话人特征序列用一个包含 K 个高斯的高斯混合模型来建模，其中 K 为说话人个数。对该高斯混合模型进行随

机初始化，利用 EM 算法进行迭代训练，最终收敛到一个最大似然解，收敛后每个高斯模型即代表一个说话人。一个包含两个说话人的高斯混合模型如下图所示。



基于该模型，判断每一个语音帧 x_t 对应的说话人特征对每个高斯成分 M_i 的条件概率 $P(x_t|M_i)$，取最大的 M_i 所对应的说话人为 x_t 所属的说话人。

对于说话人切分任务，语音中包含的说话人个数已知，则可直接利用上述 EM 方法得到包含两个高斯成份的混合模型，进而对每一个语音帧进行切分。对于说话人分析任务，无法事先确知说话人个数，可从 1 开始逐渐增加说话人的个数 K，并利用贝叶斯信息准则（BIC）选择最合适的 K 作为对说话人个数的估计，并依混合数为 K 的高斯混合模型对语音帧进行分割。

发明优势

本发明所提方法利用说话人特征信息实现基于聚类的说话人分割，解决了基于初级特征分割中的不确定性问题，分割模型简单，可靠性高。同时，由于分割在帧层次上进行，时间解析度高。