**多人对话场景音频分离**

曾昕 于鹏 刘昱

**背景**

国内外进展

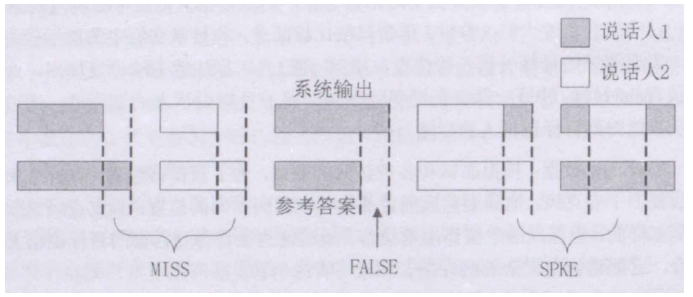
自 20 世纪 90 年代以来，针对不同的应用需求背景，说话人分割聚类己经在不同的说话情境下进行了研究。最开始的研究主要是针对双人对话电话数据：到了 1990 年代末期，研究重点转向了多人的广播电视数据，2002 年以来，会议场景下的分割聚类研究成为了热点。根据分割和聚类之间结合方式的不同，目前主流的说话人分割聚类系统可以分为两类：先进行分割再聚类，即先对输入的语音进行变化点检测，将语音切分成只包含一个说话人的语音片段，然后再将同一个人的片段聚为一类；分割和聚类同时进行，每次聚类之后，根据当前聚类结果建立的隐马尔科夫模型重新进行分割。这两类系统有各自的优缺点，前者系统简单，但每次聚类的错误会累积，后者会对每次聚类的错误进行修正，最终的聚类结果通常要比前者好，但是需要花费更多计算时间。

任务

针对一段包含客服与顾客对话的录音，区分出不同人员，并将不同人员的对话抽离、分轨。举例，如果是一通客户（记为 A）与客服（记为 B）的通话，需要能够从对话中分离出两个对话文件。其中，A.mp3 中只有客户的声音，B.mp3 中，只有客服的声音。

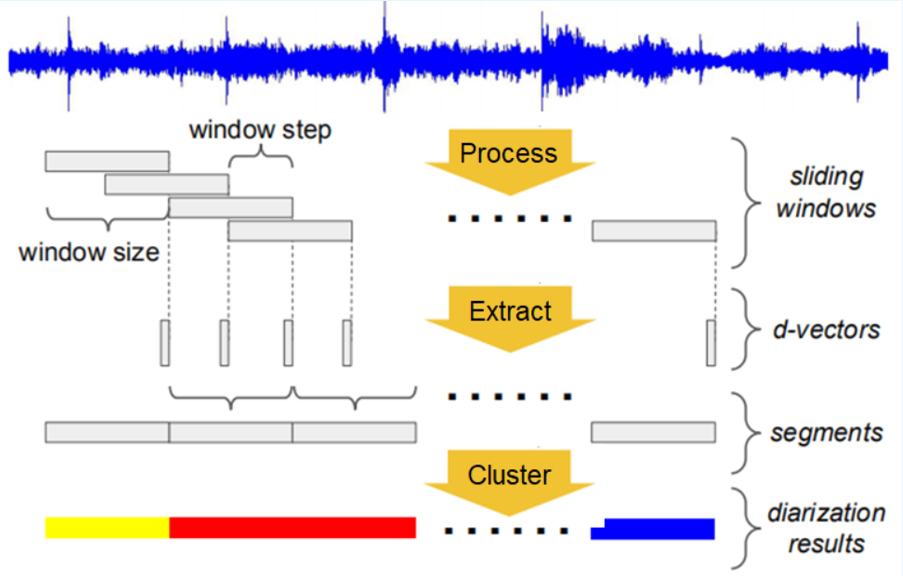
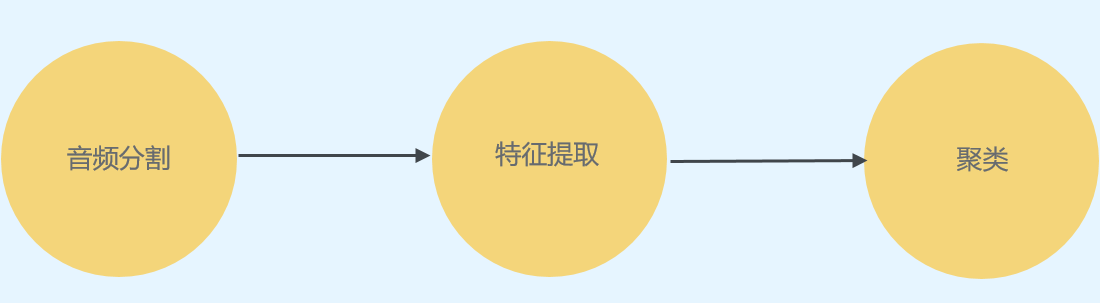
评价指标

DER(Diarization Error Rate)是NIST用于说话人聚类分割的主要评价指标，表示一段时间内某段语音不能正确归类到正确的说话人身上。DER讲错误分为三类：MISS、FALSE ALARM、ERROR（分别指少一段、多一段、错误分段）。

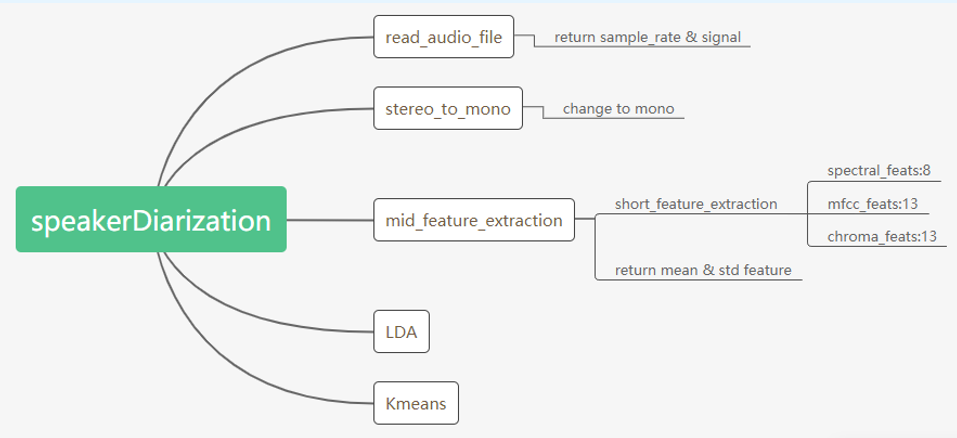


**原理**

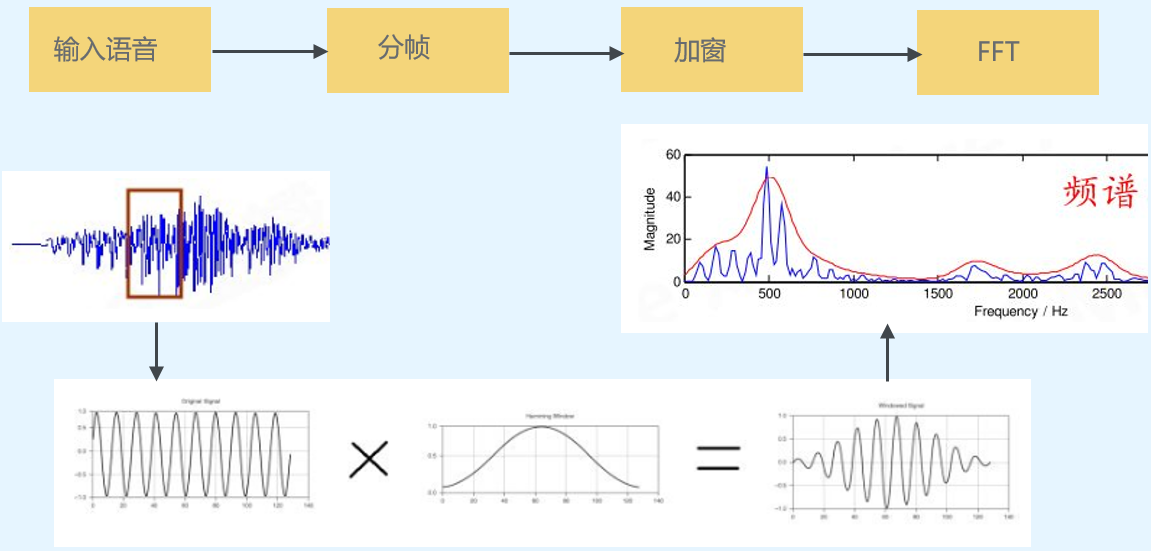
核心步骤



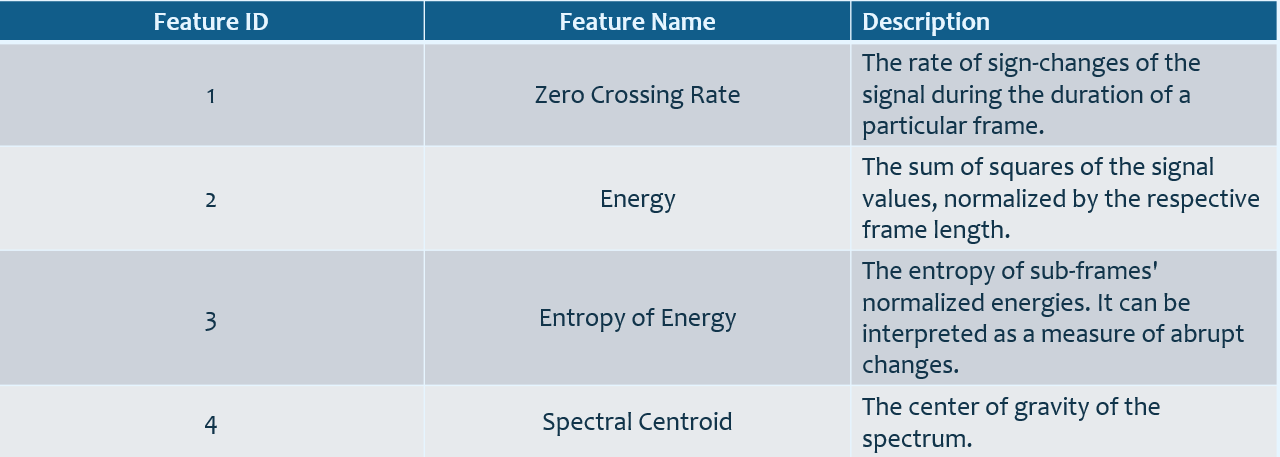
网络架构



音频分割



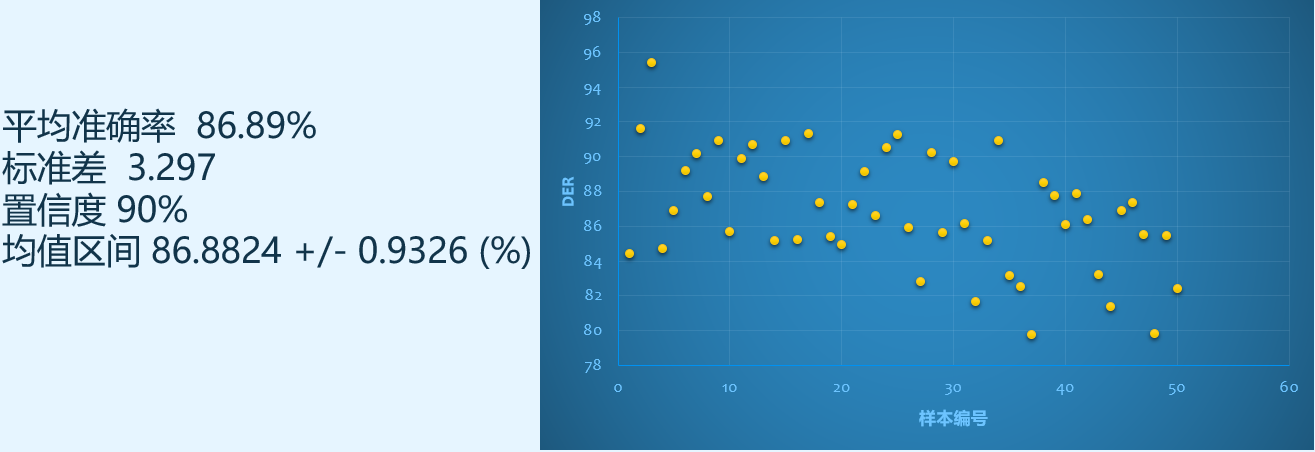
特征提取



聚类

K-means算法。K-means 算法的基本思想是，将制定样本划分为 K 个簇，用一个聚类的中心来代表一个簇，这个中心就是各个簇中所有数据样本的均值（means）。1） 任意选择K个对象作为初始的簇中心；2） 分别计算数据集中每个元素与所选簇的中心计算距离（一般采用欧式距离），根据最近邻原则，将元素划分到相应的簇中；3） 计算每个簇中对象的平均值，更新簇的中心；4） 重复上面的步骤，直至更新的簇的中心与原簇的中心的差值在预定范围内，或者达到预设的迭代次数；5） 输出K个簇中心。

**结果**



**参考文献**

1. Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mans- field, and Ignacio Lopz Moreno, “Speaker diarization with lstm,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5239–5243.
2. Zhang A , Wang Q . FULLY SUPERVISED SPEAKER DIARIZATION[J]. 2018.
3. 李芝峰. 基于深度学习的多说话人语音分离技术研究[D].辽宁大学,2019.
4. 朱唯鑫. 多人对话场景下的说话人分割聚类研究[D].中国科学技术大学,2017.
5. Neural building blocks for speaker diarization: speech activity detection, speaker change detection, speaker embedding：https://github.com/pyannote/pyannote-audio
6. Python Audio Analysis Library: Feature Extraction, Classification, Segmentation and Applications：https://github.com/tyiannak/pyAudioAnalysis