语音文字识别与说话人识别

李磊 陶孟骐 王学勤 老智昊

第 19 组

目　　录

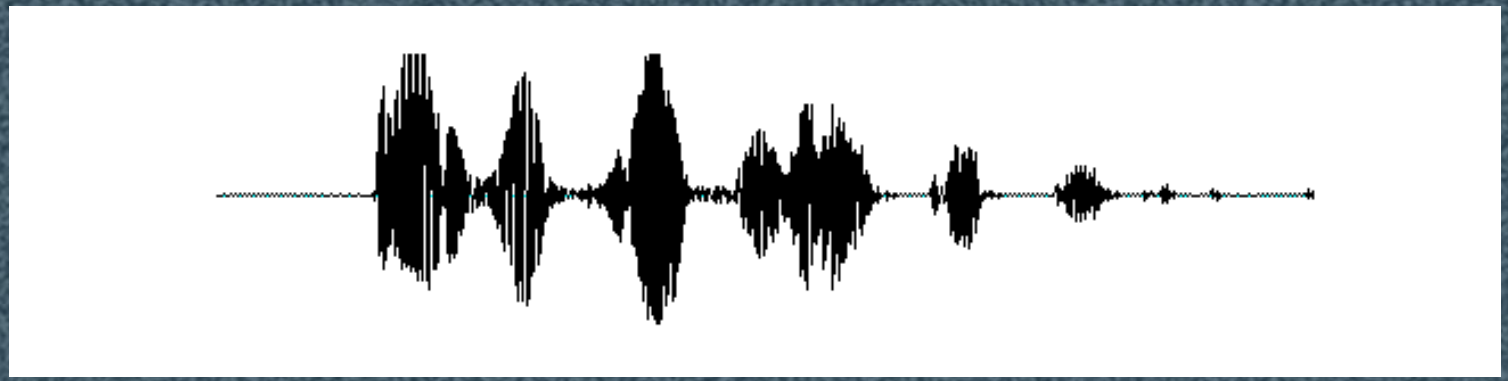
01

Part one

语音识别

语 音 识 别

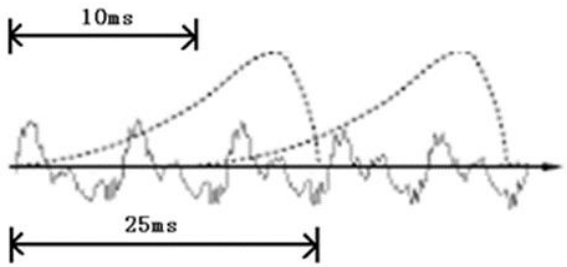
首先，我们知道声音实际上是一种波。常见的mp3等格式都是压缩格式，必须转成非压缩的纯波形文件来处理，比如Windows PCM文件，也就是俗称的wav文件。wav文件里存储的除了一个文件头以外，就是声音波形的一个个点了。下图是一个波形的示例。



在开始语音识别之前，有时需要把首尾端的静音切除，降低对后续步骤造成的干扰。这个静音切除的操作一股称为VAD，需要用到信号处理的一些技术。

语 音 识 别

要对声音进行分析，需要对声音分帧，也就是把声音切开成一小段一小段，每小段称为一帧。分帧操作一般不是简单的切开，而是使用移动窗函数来实现，这里不详述。帧与帧之间一般是有交叠的，就像下图这样：

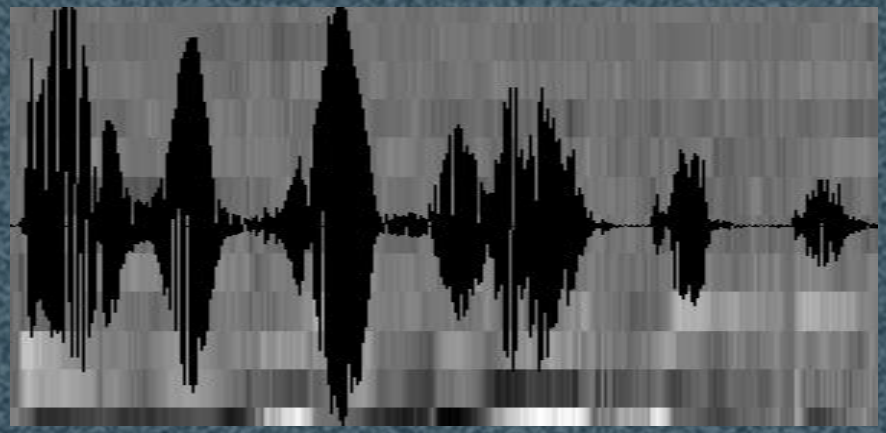


图中，每帧的长度为25毫秒，每两帧之间有25-10=15毫秒的交叠。我们称为以帧长25ms、帧移10ms分帧。

分帧后，语音就交成了很多小段。但波形在时域上几乎没有描述能力，因此必须将波形作交换。常见的一种交换方法是提取MFCC特征，根据人耳的生理特性，把每一帧波形交成一个多维向量，可以简单地理解为这个向量包含了这帧语音的内容信息。

语 音 识 别

至此，声音就成了一个12行(假设声学特征是12维）、N列的一个矩阵，称之为观察序列，这里N为总帧数。观察序列如下图所示，图中，每一帧都用一个12维的向量表示，色块的颜色深浅表示向量值的大小。

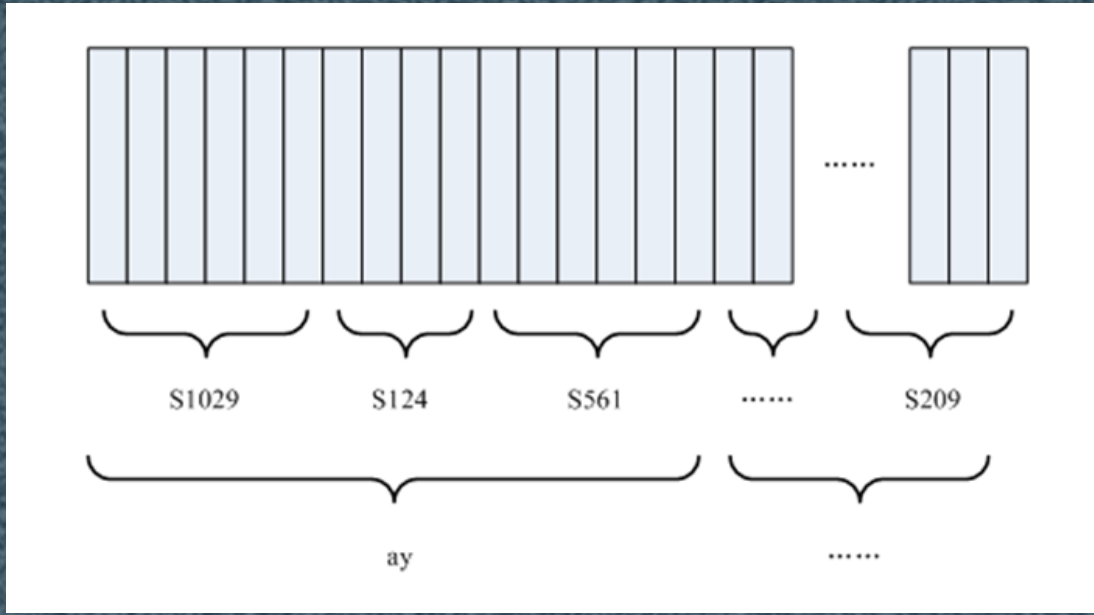


接下来就要介绍怎样把这个矩阵变成文本了。首先要介绍两个概念：

1. 音素：单词的发音由音素构成。对英语，一种常用的音素集是卡内基梅隆大学的一套由39个音素构成的音素集。汉语一般直接用全部声母和韵母作为音素集，另外汉语识别还分有调无调，不详述。

2.状态：这里理解成比音素更细致的语音单位就行啦。通常把一个音素划分成了个状态。

语 音 识 别

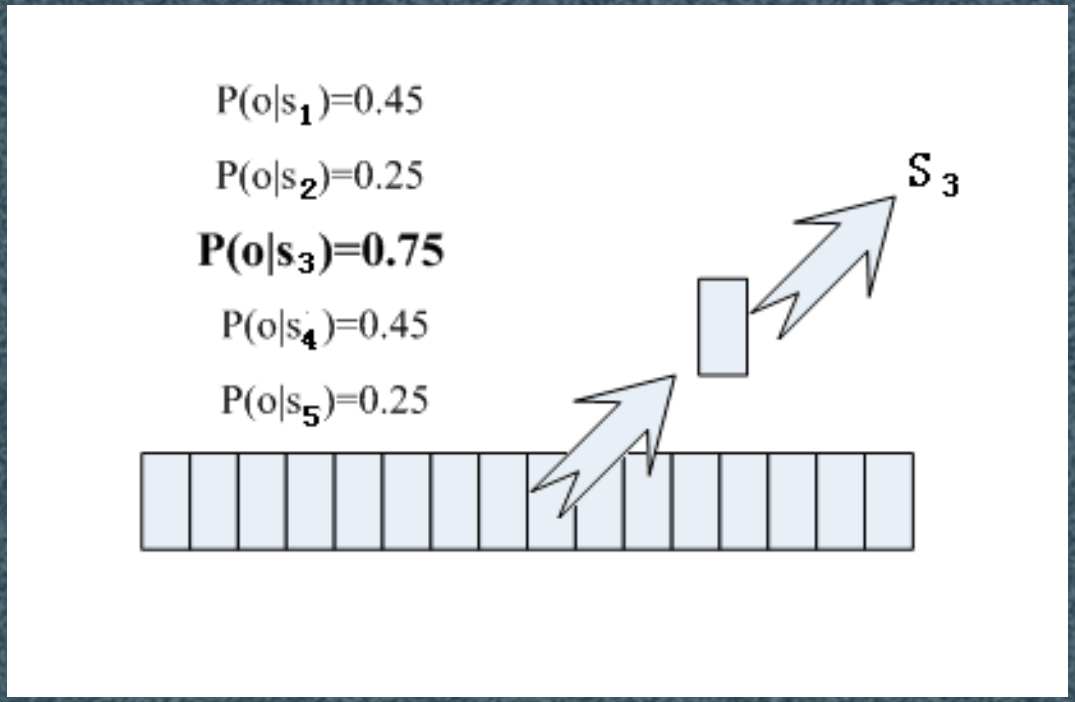


语音识别是怎么工作的呢？实际上一点都不神秘，无非是：

1. 把帧识别成状态（难点）。
2. 把状态组合成音素。
3. 把音素组合成单词。

图中，每个小竖条代表一帧，若干帧语音对应一个状态，每三个状态组合成一个音素，若干个音素组合成一个单词。也就是说，只要知道每帧语音对应哪个状态了，语音识别的结果也就出来了。

语 音 识 别



那每帧音素对应哪个状态呢？有个容易想到的办法，看某帧对应哪个状态的概率最大，那这帧就属于哪个状态。比如下面的示意图，这帧在状态S3上的条件概率最大，因此就猜这帧属于状态S3。

语 音 识 别

那这些用到的概率从哪里读取呢？有个叫“声学模型”的东西，里面存了一大堆参数，通过这些参数，就可以知道帧和状态对应的概率。获取这一大堆参数的方法叫做“训练”，需要使用巨大数量的语音数据，训练的方法比较繁琐

但这样做有一个问题：每一帧都会得到一个状态号，最后整个语音就会得到一堆乱七人糟的状态号。假设语音有1000帧，每帧对应1个状态，每3个状态组合成一个音素，那么大概会组合成300个音素，但这段语音其实根本没有这么多音素。如果真这么做，得到的状态号可能根本无法组合成音素。实际上，相邻顿的状态应该大多数都是相同的才合理，因为每帧很短。

语 音 识 别

解决这个问题的常用方法就是使用隐马尔可夫模型 (Hidden Markov Model, HMM)。这东西听起来好像很高深的样子，实际上用起来很简单：
第一步，构建一个状态网络。
第二步，从状态网络中寻找与声音最匹配的路径。

这样就把结果限制在预先设定的网络中，避免了刚才说到的问题，当然也带来一个局限，比如你设定的网络里只包含了“今天晴天”和“今天下雨”两个句子的状态路径，那么不管说些什么，识别出的结果必然是这两个句子中的一句。

那如果想识别任意文本呢？把这个网络搭得足够大，包含任意文本的路径就可以了。但这个网络越大，想要达到比较好的识别准确率就越难。所以要根据实际任务的需求，合理选择网络大小和结构。

语 音 识 别

搭建状态网络，是由单词级网络展开成音素网络，再展开成状态网络。语音识别过程其实就是在状态网络中搜索一条最佳路径，语音对应这条路径的概率最大，这称之为“解码”。路径搜索的算法是一种动态规划剪枝的算法，称之为Viterbi算法，用于寻找全局最优路径。

这里所说的累积概率，由三部分构成，分别是：

1.观察概率：每帧和每个状态对应的概率

2.转移概率：每个状态转移到自身或转移到下个状态的概率

3. 语言概率：根据语言统计规律得到的概率

其中，前两种概率从言学模型中获取，最后一种概率从语言模型中获取。语言模型是使用大量的文本训练出来的，可以利用某门语言本身的统计规律来帮助提升识别正确率，语言模型很重要，如果不使用语言模型，当状态网络较大时，识别出的结果基本是一园乱麻。这样基本上语音识别过程就完成了。

发 展 历 史

1

1952年贝尔研究所Davis等人研究成功了世界上第一个能识别10个英文数字发音的实验系统。

2

1960年英国的Denes等人研究成功了第一个计算机语音识别系统。

3

大规模的语音识别 研究是在进入了70年代以后，在小词汇量、孤立词的识别方面取得了实质性的进展。

4

进入80年代以后，研究的重点逐渐转向大词汇量、非特定人连续语音识别。在研究思路上也发生了重大变化，即由传统的基于标准模板匹配的技术思路开始转向基于统计模型（HMM）的技术思路。

5

进入90年代以后，在语音识别的系统框架方面并没有什么重大突破。但是，在语音识别技术的应用及产品化方面出现了很大的进展。

基 本 模 型

隐马尔可夫模型

现代通用语音识别系统基于隐马尔可夫模型。这些是输出符号或数量序列的统计模型。HMM用于语音识别，因为语音信号可以被视为分段静止信号或短时静止信号。在短时间尺度中，语音可以近似为静止过程。语音可以被认为是许多随机目的的马尔可夫模型。

基于动态时间规整

动态时间扭曲是一种历史上用于语音识别的方法，但现在已经被更成功的基于HMM的方法取代

神经网络

在20世纪80年代后期，神经网络在ASR中成为一种有吸引力的声学建模方法。从那时起，神经网络已被用于语音识别的许多方面，例如音素分类，孤立词识别，视听语音识别，视听说话人识别和说话者适应。

02

Part two

说话人识别

说话人识别目录

说话人识别简介

说话人识别又称声纹识别（Voiceprint Recognition,VPR），顾名思义，即通过声音来识别出来“谁在说话”，是根据语音信号中的说话人个性信息来识别说话人身份的一项生物特征识别技术。

声纹识别的理论基础是每一个声音都具有独特的特征，通过该特征能将不同人的声音进行有效的区分

说话人识别发展史

知识积累阶段

20世纪30-50年代

模板匹配阶段

20世纪50-60年代

模式识别阶段

20世纪60-80年代

统计模型阶段

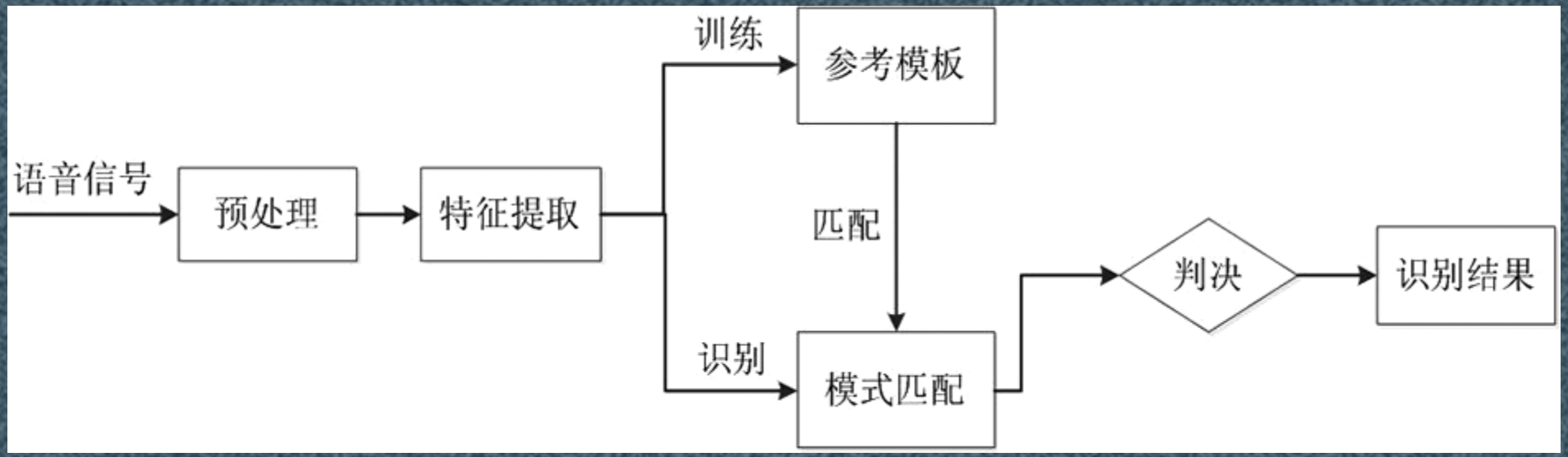
1980-2000年

机器和深度学习

2000-2010

2011年至今

说话人识别系统框图

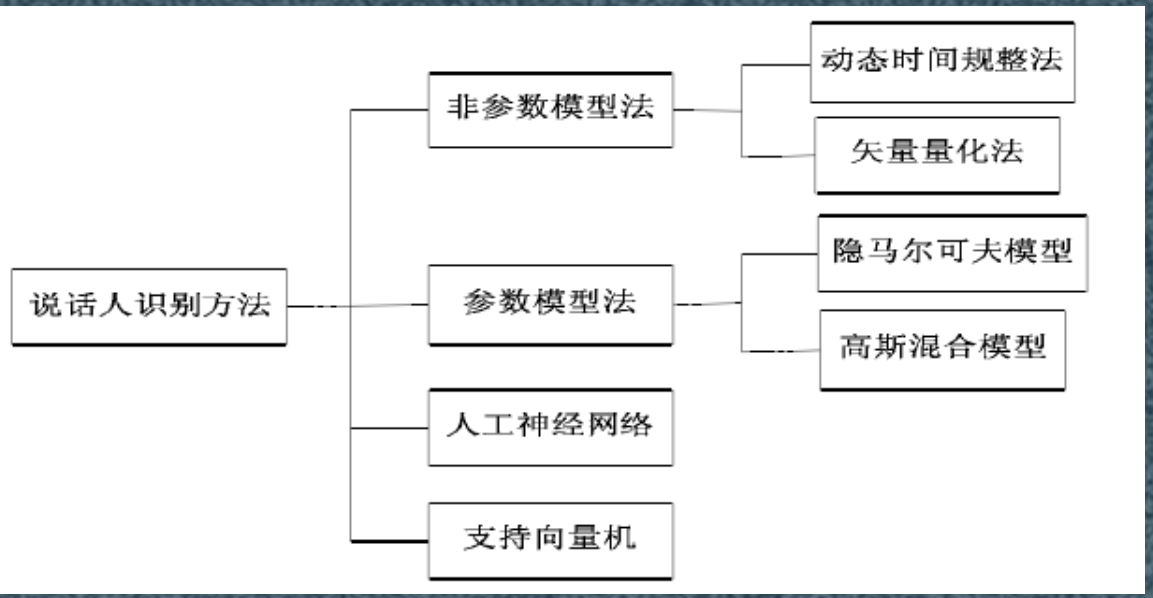


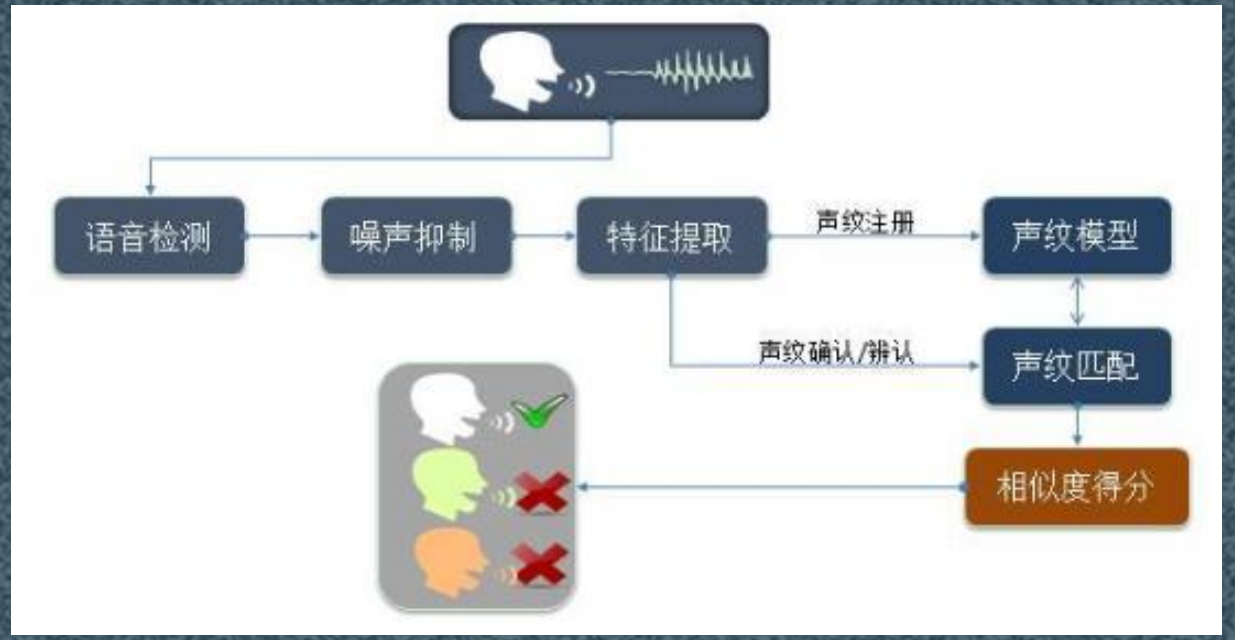
说话人识别分类

说话人识别任务根据识别方式的不同， 可以分为三类：

说话人确认
说话人鉴别
说话人探测跟踪

说话人识别原理


说话人识别方法
非参数模型法
参数模型法
人工神经网络
支持向量机
动态时间规整法
矢量量化法
隐马尔可夫模型
高斯混合模型


语音检测
噪声抑制
特征提取
声纹注册
声纹模型
声纹确认/辨认
声纹匹配
相似度得分

说话人识别模式匹配方法

1.概率统计

2.动态时间规整

3.最近邻方法

4.矢量量化

5. VQ聚类方法

6.隐马尔可夫模型

7. 人工神经网络

概 率 统 计

语音中说话人信息在短时内较为平稳，通过对稳态特征如基音、声门增益、低阶反射系数的统计分析，可以利用均值、方差等统计量和概率密度函数进行分类判决。其优点是不用对特征参量在时域上进行规整，比较适合文本无关的说话人识别。

动态时间规整

说话人信息不仅有稳定因素，而且有时变因素（语速、语调、重音和韵律）。将识别模板与参考模板进行时间对比，按照某种距离测定得出两模板间的相似程度。常用的方法是基于最近邻原则的动态时间规整DTW。

最邻近方法

训练时保留所有特征矢量，识别时对每个矢量都找到训练矢量中最近的K个，据此进行识别，通常模型存储和相似计算的量都很大；

矢 量 量 化

矢量量化最早是基于聚类分析的数据压缩编码技术。矢量量化就是将若干个标量数据组构成一个矢量，然后在矢量空间给以整体量化，从而压缩了数据而不损失多少信息。Helms首次将其用于声纹识别，把每个人的特定文本编成码本，识别时将测试文本按此码本进行编码，以量化产生的失真度作为判决标准。这种方法的识别精度较高，且判断速度快。

VQ聚类方法

VQ聚类方法(如LBG，K-均值)：效果比较好，算法复杂度也不高，和HMM方法配合起来可以收到更好的效果；

隐马尔可夫模型

隐马尔可夫模型是一种基于转移概率和传输概率的随机模型，它把语音看成由可观察到的符号序列组成的随机过程，符号序列则是发声系统状态序列的输出。在使用HMM识别时，为每个说话人建立发声模型，通过训练得到状态转移概率矩阵和符号输出概率矩阵。

HMM不需要时间规整，可节约判决时的计算时间和存储量，目前被广泛应用在工业领域，缺点是训练时计算量较大。

人 工 神 经 网 络

人工神经网络在某种程度上模拟生物的感知特性，它是一种分布式并行处理结构的网络模型，具有自组织和自学习能力、很强的复杂分类边界区分能力以及对不完全信息的鲁棒性，其性能近似理想的分类器。缺点是训练时间长，动态时间规整能力弱，网络规模随说话人数目增加时可能大到难以训练的程度。

说话人识别模型基础

GMM-UBM（混合高斯-通用背景模型）

联合因子分析

说话人矢量因子（Identity-Vector，I-Vector）

TVM-I-Vector

信道补偿算法

LDA线性鉴别

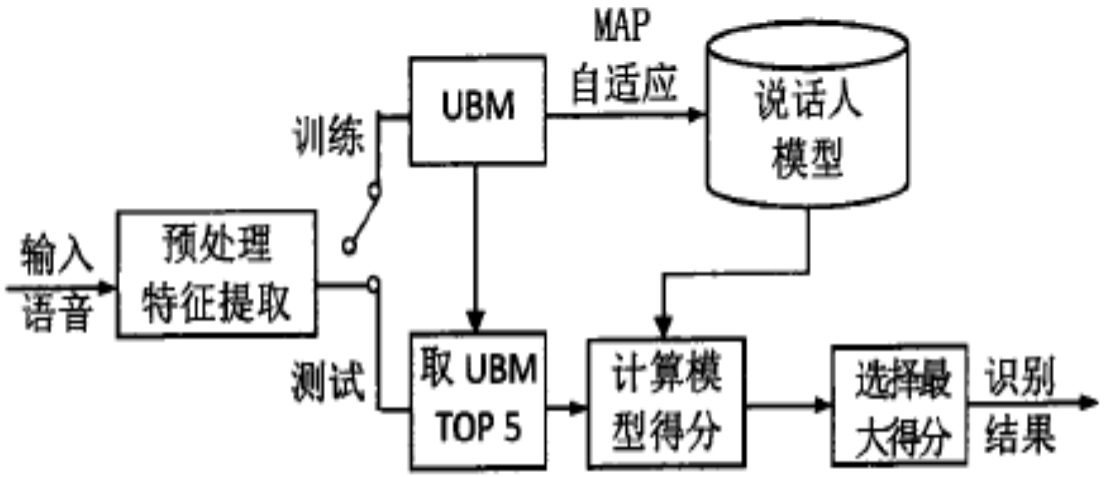


图 4.3 基于 GMM-UBM 的说话人识别系统流程图

Fig.4.3 System flow chart of speaker recognition based on GMM-UBM.

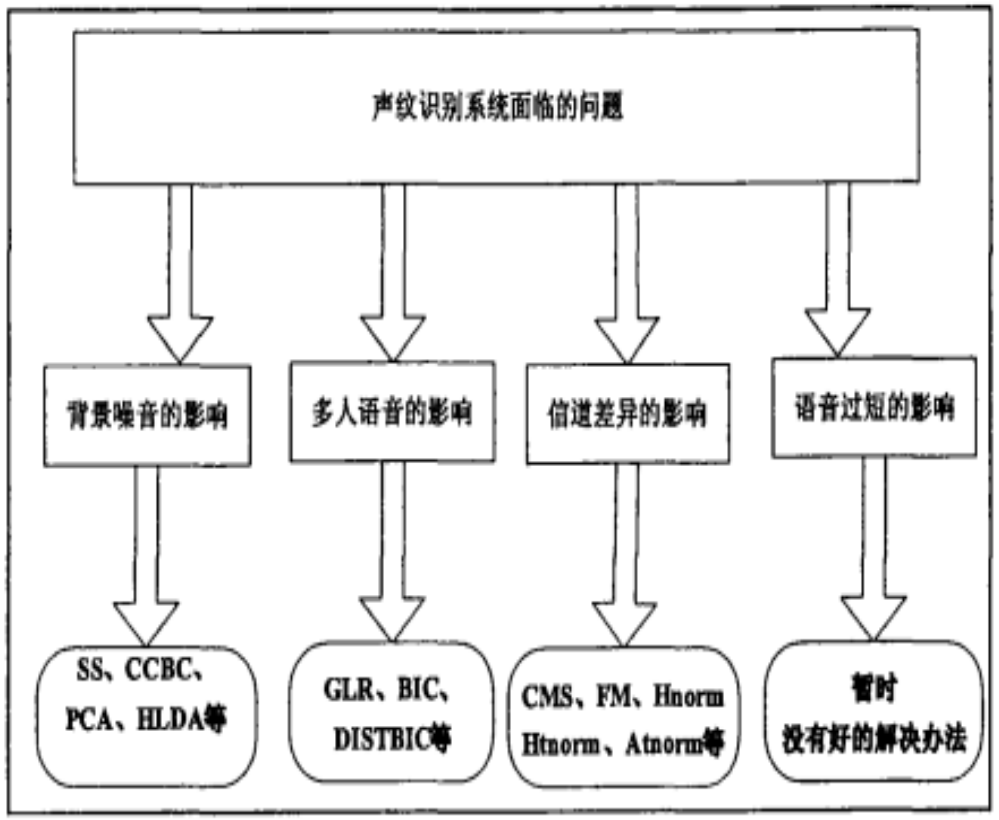
应 用 场 景

网络视听内容监管-------审查监管

公安司法鉴定----------协助取证

军事领域----------军事保密、指令确认、情报侦听

应 用 难 点


声纹识别系统面临的问题
背景噪音的影响
多人语音的影响
信道差异的影响
语音过短的影响
SS、CCBC、
PCA、HLDA等
GLR、BIC、
DISTBIC等
CMS、FM、Hnorm
Htnorm、Atnorm等
暂时
没有好的解决办法

Part three

DEMO

03


Microsoft Windows [版本 10.0.22000.493]
(c) Microsoft Corporation。保留所有权利。
C:\Users\lzh\Desktop\speech-demo-master\speech-demo-master\rest-api-asr\python>conda activate ai
(ai) C:\Users\lzh\Desktop\speech-demo-master\speech-demo-master\rest-api-asr\python>python asr_json.py
1920X1080
超格式录屏大师
https://luping.hgs.cn/
媒体播放器
主页
最近使用的文件
0:00:03
0:00:01
播放 (Ctrl+P)
16k

谢谢大家

李磊 陶孟骐 王学勤 老智昊

第 1 9 组