说话人/语种识别

李军锋

中科院语言声学与内容理解重点实验室
中国科学院声学研究所

提纲

- 概述
- 说话人识别原理和系统
- 系统鲁棒性问题

P1 概述

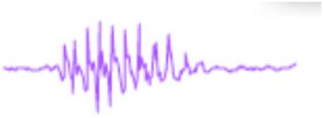
相关概念

- 语音：人们之间进行沟通所使用的最基本的方式
- 语音信号：包含了一些对人类交流非常有用的信息，从语音的产生方面来讲，语音信号包含了语言学方面的信息(语义和语种)以及说话人方面的信息。因此通过对语音信号进行研究与处理，提取出人们感兴趣的信息就具有非常重要的意义。
- 自动语音识别技术：其目的是通过机器，利用相应的算法从语音信号中自动提取出人们需要的、具有实际意义的信息。

P1 概述

目的：从语音中自动提取出人们感兴趣的、具有实际意义的信息

语音信号



语音识别 → 语音内容：“你好吗？”

语种识别 → 语种信息：汉语

说话人识别 → 说话人身份：李雷

P1 概述

- 说话人识别定义
  - 给定语音信号，自动识别其说话人身份的技术

- 说话人识别技术是生物识别技术的一种。生物识别技术是指通过生理或者行为特征对人的身份进行识别。用来进行生物识别的特征应该符合以下特点：
  - 每个人都应该拥有该特征
  - 该特征对每个人都具有明显的区分性
  - 该特征在一定的时期内固定不变
  - 该特征易于采集。

P1 概述

- 目前，说话人识别是生物识别研究中的一大热点，原因有下：
  - 语音的收集方便、自然，具有比较低的用户侵犯性
  - 采集设备比较简单
  - 获取成本低廉，使用简单；
  - 无处不在的语音通信网络，如固定电话、移动通信和互联网等提供了大量的音频数据，为说话人识别技术的研究和应用提供了良好的条件。

P1 概述

说话人识别的应用

- **信息领域。**声纹识别技术可以在呼叫中心（Call Center）应用中为注册的常客户提供友好的个性化服务。
- **银行、证券。**鉴于密码的安全性不高，可以用声纹识别技术对电话银行、远程炒股等业务中的用户身份进行确认，为了提供安全性，还可以采取一些其他措施，如密码和声纹双保险，如随机提示文本用文本相关的声纹识别技术进行身份确认（随机提示文本保证无法用事先录好的声音去假冒），甚至可以把交易时的声音录下来以备查询。
- **公安司法。**对于各种电话勒索、绑架、电话人身攻击等案件，声纹辨认技术可以在一段录音中查找出嫌疑人或缩小侦查范围；声纹确认技术还可以在法庭上提供身份确认的旁证。
- **军队和国防。**声纹辨认技术可以察觉电话交谈过程中是否有关键说话人出现，继而对交谈的内容进行跟踪（战场环境监听）；在通过电话发出军事指令时，可以对发出命令的人的身份进行确认（敌我指挥员鉴别）。
- **保安和证件防伪。**如机密场所的门禁系统。

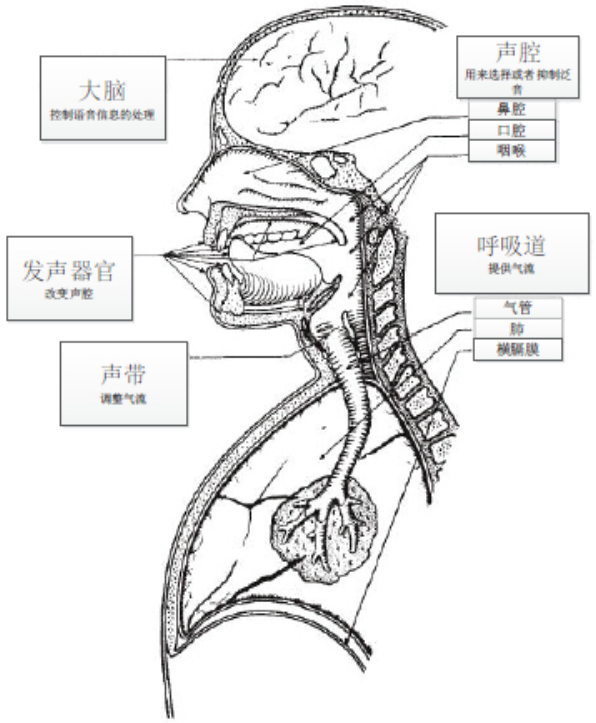
P1 概述

- 识别任务分类
  - 说话人识别任务根据实际应用的类型不同，可以分为：
    - 说话人辨认
    - 说话人确认
  - 根据测试语音是否一定由当前的多个目标说话人所说，可以继续被分为
    - 闭集模式
    - 开集模式
  - 说话人识别任务从文本相关性角度进行分类，可以被分成:
    - 文本无关
    - 文本相关

P1 概述

研究历史及现状

- 最早的说话人识别研究可以追溯到半个世纪之前。
- 1962年，贝尔实验室的Lawrence Kersta在《自然》杂志上发表了基于语谱图(voiceprint)的说话人识别方法。文中对人体发声器官功能进行了总结在此基础上，提出了使用语谱图作为一个人的关键生物特征(类似指纹的作用)进行说话人识别的思想。



P1 概述

- 从这之后到1980年的二十年间内，说话人的研究开始逐步系统化。
- 1963年，贝尔实验室的Pruzansky搭建了第一个自动说话人识别系统。
- 在后面若干年中，框架不断被修缮：
  - 谱信息的表征被改进
  - 表征说话人的特征中引入了韵律信息;
  - 引入时间上的动态信息建模、统计建模等
- 这段时期，文本相关说话人识别和文本无关说话人识别都有一定的发展：
  - 文本无关说话人识别在这个时期处于刚起步的状态，系统性能较差，有一些新特征被提出。如瞬时谱方差矩阵、线性预测系数(linear prediction coefficients,LPC)等。这段时间还出现了特征规整技术，如倒谱均值减(cepstral mean subtraction, CMS)、高阶差分等。
  - 文本相关的说话人识别在事先给定的文本下，当录音环境较好时，系统识别的效果可以非常精确。

P1 概述

- 80年代到90年代这段时间里，说话人识别的工作重点从之前较为随意的模版、规则匹配转移到基于统计模型的严格的概率计算上来。
- 语音识别领域得到广泛应用的统计模型HMM(Hidden Markov Model)在文本相关的说话人识别任务中被应用起来。这里的文本相关任务主要针对一串特定的数字或者短语进行说话人识别。
- 同一时期，HMM 也被用在文本无关的说话人识别任务中。Pritz在声学特征空间上使用了5状态各态遍历的HMM模型。随后Tishby将Pritz方法中的状态数扩展到了8个状态。
- 这期间HMM单状态的混合数一般在2到8。同时非参数统计模型矢量量化也被引入说话人识别中。训练集中的说话人的特征被映射到一个矩阵中，这个矩阵被称为VQ 编码字典。测试语句通过编码字典进行编码变换到一个距离打分上，系统通过这个打分进行决策。

P1 概述

- 进入90年代后， 说话人识别系统的鲁棒性成为了研究的重中之重。Matsui对比了说话人识别中矢量量化以及各态遍历的HMM模型的性能，得出这两种系统在训练数据足够多的时候是非常鲁棒的。
- Reynolds经过研究发现，HMM中模型中真正影响性能的是单状态的高斯数而不是状态数，从而提出了单状态的HMM：高斯混合模型-通用背景模型(Gaussian Mixture Model-Universal Background Model, GMM-UBM)。GMM相较HMM保留了混合高斯分布，去掉了状态跳转，简单有效，具有较好的鲁棒性，迅速成为文本无关的说话人识别任务中的主流技术。
- 从1996年开始，为了评估说话人识别技术的进展及水平，美国国家标准技术局(National Institute of Standard and Technology, NIST) 开始举办公开的说话人识别评测(Speaker Recognition Evaluation, SRE)。
- NIST说话人评测的技术水平代表了当今说话人识别研究领域的最高水平，历年的评测积累了大量各种信道下的语音数据。为说话人识别技术的研究提供了一个良好的平台，在该项评测的推动下，说话人识别逐年快速进步。

P1 概述

- 2000年之后，说话人识别技术逐渐向实用方向发展，复杂信道下说话人识别变成了研究的重点。
- 得分规整技术如ZNORM和TNORM等在这时段被提出。
- 此外信道补偿算法如说话人模型合成Speaker Model Synthesis, SMS)、特征映射(Feature Mapping)和特征平滑(Feature Warping)等被提出。模式识别中的新工具支持向量机 (Support Vector Machine, SVM)也被引入说话人中。最早SVM使用的是广义线性判别式序列核函数，但是该算法缺乏理论的支持。
- Campell提出了基于KL距离核函数的高斯混合模型-支撑向量机系统(GMM-SVM)。该系统中输入支持向量机的是由GMM的各个高斯分量的均值向量拼接而成高维向量。为了消除信道干扰，Campell基于GMM-SVM系统提出了扰动属性投影Nuissance Attribute Projection, NAP)。

P1 概述

- 随着因子分析技术被引入说话人识别中，复杂信道情况下的说话人识别性能有了很大提升。
- 基于高斯超向量，Kenny在高维空间提出了联合因子分析算法(Joint Factor Analysis，JFA)。
- Dehak和Kenny对说话人空间建模进行了更合理的假设，就有了总变化因子分析(Front-End Factor Analysis，FEFA)。为了利用这部分的信息，在总变化因子分析的后端，使用了类内协方差规整(Within-Class Covariance Normalization，WCCN)以及线性判别分析(Linear Discriminant Analysis，LDA)对总变化因子进行类内以及类间信息的补充。
- 通过更深入地分析总变化因子的特点，并对其进行低维因子分析建模，概率线性判别分析模型（Probabilistic Linear Discriminant Analysis，PLDA）被引入说话人识别中。这种低维的因子分析模型能够在去除信道影响的基础上，更好地学习说话人类内及类间的信息，从而达到更好地表征说话人的作用。

P1 概述

- 目前的说话人识别研究工作针对的还是以下两个影响说话人识别性能的重要因素：
  - 信道差异:通话信道调制和语音录制设备的差异统称为信道差异，目前是严重阻碍说话人识别实际应用的最重要的因素之一。语音录制设备在实现语音声信号到电信号的转变过程中，也会造成语音信号的失真。
  - 背景噪声。现实语音通话中带有背景噪声是一个无法避免的现象，不但影响着通话质量，也给说话人识别带来了难题。在实际应用中，噪声的引入严重影响了待识别语音的质量，从而导致说话人识别的性能急剧下降。

P1 概述

通用数据库

- 数据库对任何模式识别任务来说都是十分重要的，说话人识别任务也是一样。数据库是说话人识别研究的基础。
- NIST说话人评测对目前说话人识别语音数据库的积累起到了非常积极的作用。1996年开始，NIST几乎每隔两年就举行一次说话人识别评测，这期间各单位也积累了大量的训练和测试数据。
- 这些语音数据在信道方面涵盖了日常生活常用的信道类型:包括电话信道、麦克风信道访谈、麦克风信道电话录音等；语种方面也从最初的只有英文数据，到后面扩展到了十几种语言和方言。同一个人的多次录音的时间间隔也扩大到了40-50天，信道也不尽相同。训练和测试语音的长度和信道类型基本涵盖了实际中说话人识别应用的大部分情况。

P1 概述

- SRE2002以及之前:
  - 数据主要来源于SwitchBoard数据库。
- SRE2004:
  - 数据相对于SwitchBoard数据信道情况就要复杂很多，说话人测试和训练语句中引入了多语种。语种有汉语、英语、俄语、阿拉伯语以及法语五个语种。
- SRE 2005：
  - 数据信道情况更为复杂，加入了很多语种，但95%为英语数据。同时还引入了麦克风信道的数据。

P1 概述

- SRE2006：
  - 数据集的设计也考虑了信道复杂度，多语种的问题，其中包括了更多的双语种甚至多语种的说话人，相比2004年的五个语种，SRE2006引入了二十多个语种。
- SRE2008：
  - 信道复杂度，多语种等关注点的难度有所增加，最大的变化在于访谈麦克风信道数据的引入，并将其纳入核心测试集。
- SRE2010：
  - 同样提供访谈麦克风信道数据，并且数据全部为英文数据。
- SRE2012：
  - 语音时长发生了变化，数据为300秒、100秒、30秒三种。各信道的数据部分被加入了噪声，还有部分数据录制时自带背景噪声。

P1 概述

识别性能评价指标

- 在说话人识别过程中，需要将每一个测试对(trial)进行“是同一个说话人”和“不是同一个说话人”的判决。每个识别对由一句目标说话人语音和一句测试说话人语音构成，当答案“不是同一个说话人”的识别对判决为“是同一个说话人”时，称为“虚警”；当答案“是同一个说话人”的识别对判决为“不是同一个说话人”的时候，称为“漏检”。
- NIST说话人识别评测以及目前说话人识别研究中常用一些性能指标如下：
  - 等错率(Equal Error Rate, EER)
  - 检测代价(Detection Cost Function, DCF)
  - 检错平衡曲线(Detection Error Trade-off, DET曲线)

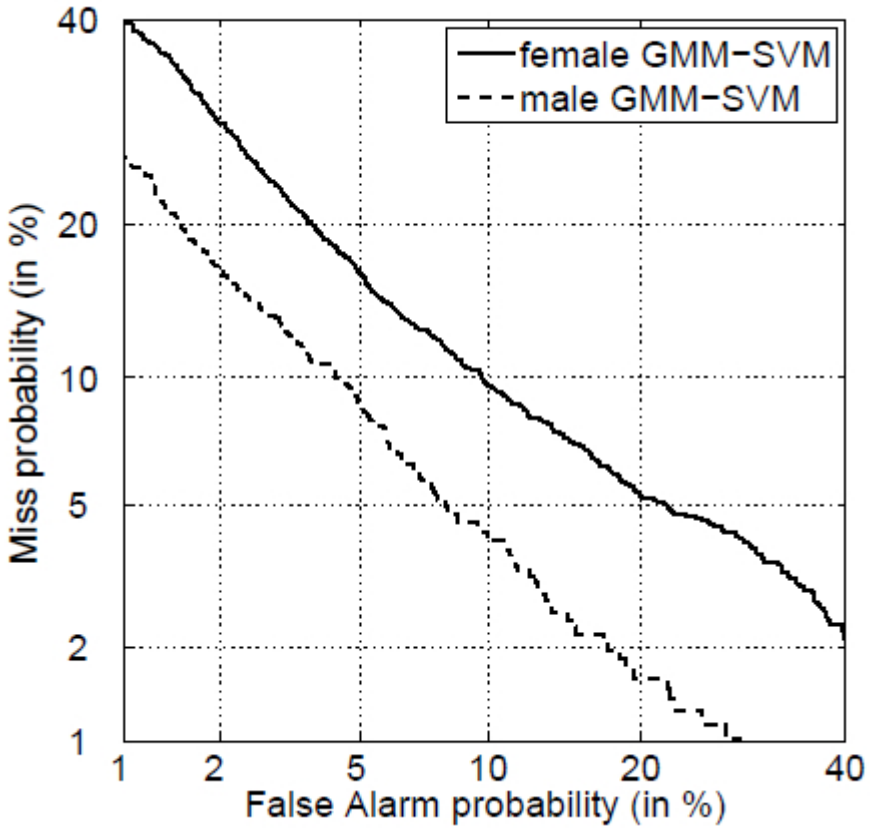
P1 概述

- 等错率
  - 随着判决门限的变化，说话人识别系统的虚警率和漏检率会向相反方向发生变化，当系统的虚警率和漏检率相等时候，这时的虚警率或者漏检率被称为等错率。等错率是一个简单但有效的衡量系统性能的评价指标，该指标认为虚警和漏检都是同等代价的。
- 检测代价
  - 检测代价通常是关于虚警率和漏检率的函数，可以很好的表示系统的性能。SRE2010的检测代价函数定义如下:

$$C_{DCF\_2010} = C_{Miss} \times P_{Target} \times P_{Miss|Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target})$$

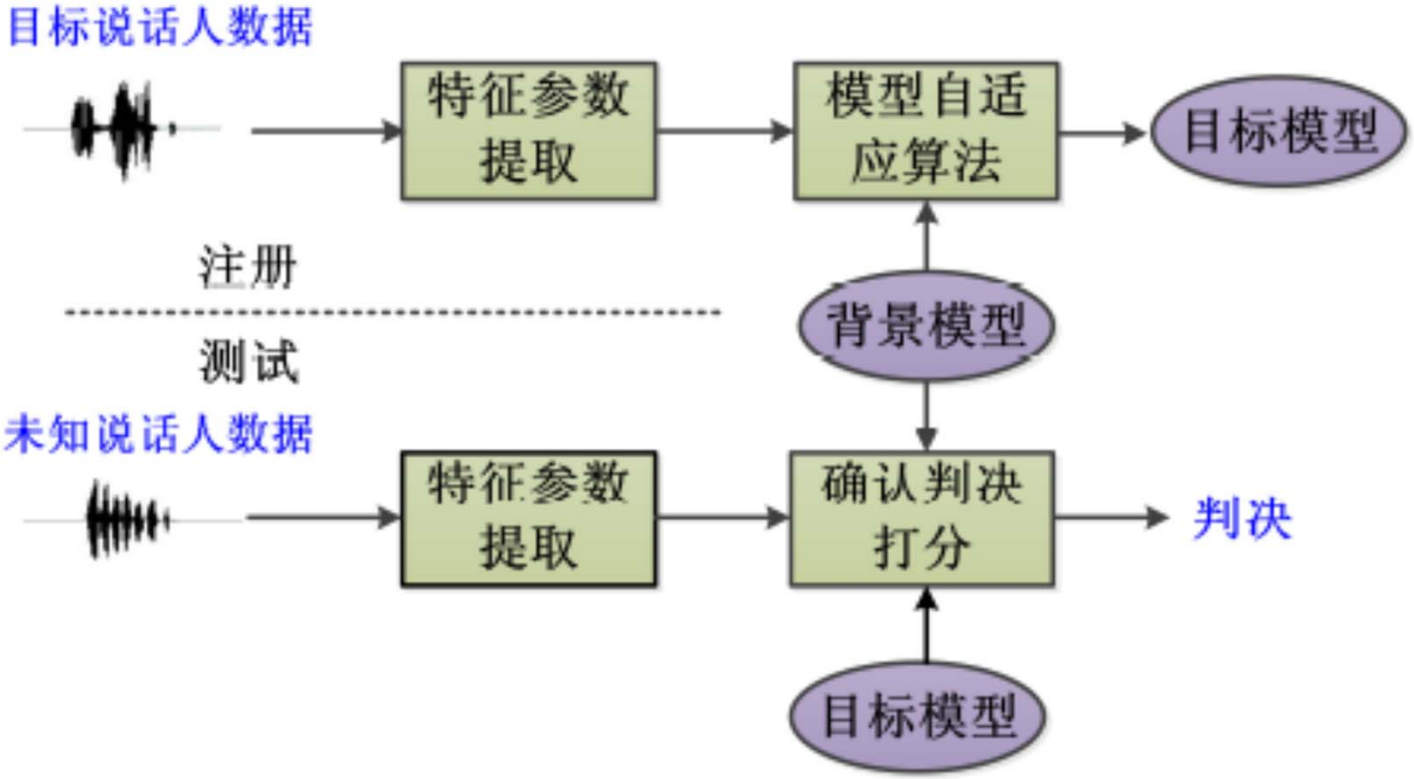
P1 概述

- 检错平衡曲线
  - 虚警率和漏检率之间的相互对应关系可以在二维平面上描绘成一条曲线来表示系统的性能，图中的曲线即为检错平衡曲线，即DET曲线。当不同系统的DET曲线描绘在同一张图上的时候，可以全面地看出不同系统之间的性能差异。



P2 说话人识别原理和系统

GMM-UBM说话人识别



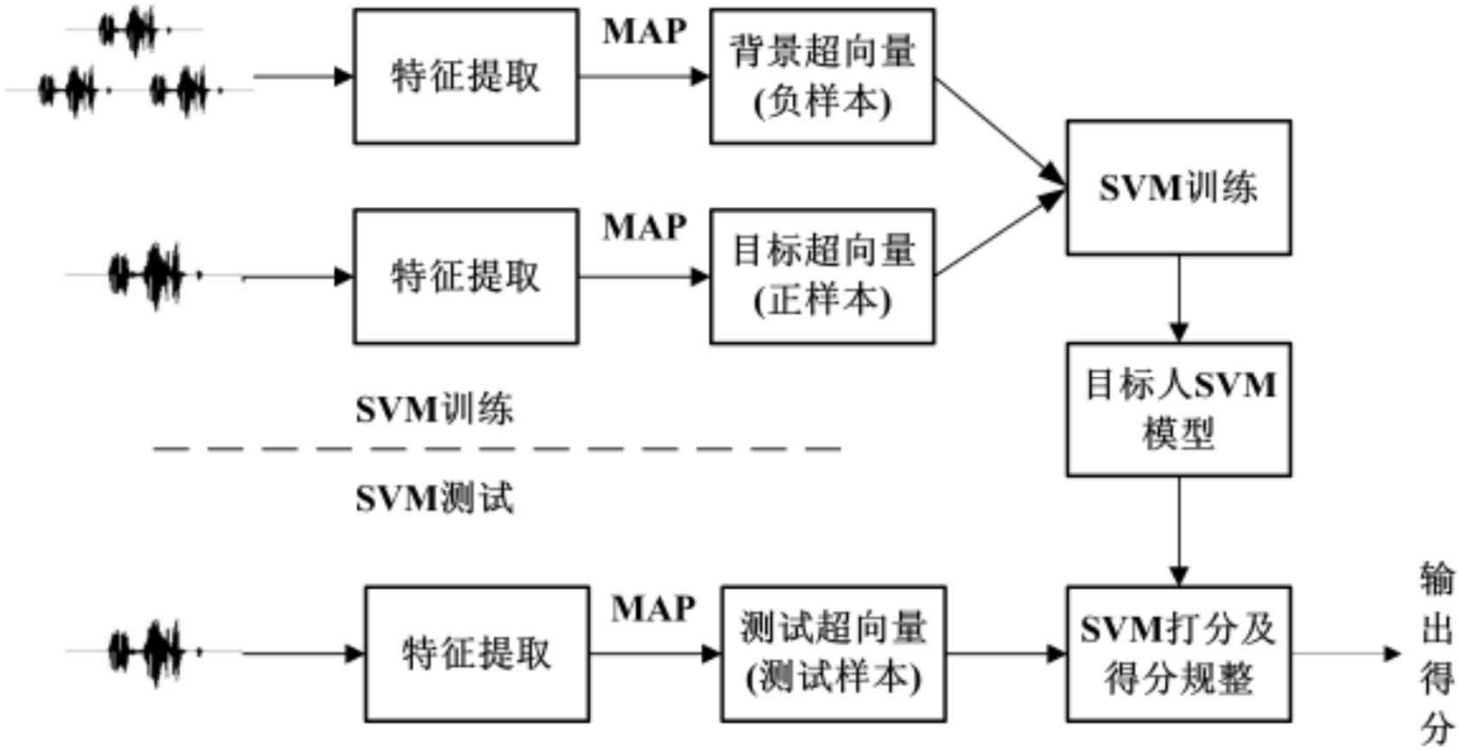
P2 说话人识别原理和系统

- 声学层特征
  - MFCC、LPCC、PLP
- 韵律层特征
  - 基频、能量以及它们的动态特征
- 音素层特征、词法特征、其它特征
- 特征后处理
  - 倒谱高阶差分、倒谱均值减、倒谱方差规整
  - Feature Warping

    MFCC 18基本倒谱系数+1阶差分+feature warping

P2 说话人识别原理和系统

GMM-SVM说话人识别



P3 系统鲁棒性问题

- 问题提出：如何消除信道影响，说话人影响？
  - 特征层面补偿
    - 倒谱均值减
    - 倒谱方差规整
    - Feature Warping
    - Rrasta 滤波
  - 模型层面补偿
    - 因子分析

P3 系统鲁棒性问题

- 倒谱均值减(Cepstral Mean Sub)
  - 在特征参数中减去一个偏移量，去除信道的影响，只能去除线性加性信道噪声
  - 偏移量可以在整段语音上估计，取整段语音的均值作为偏移量

$$\bar{\mu} = \frac{1}{T}\sum_{i=1}^{T}\bar{x}_i$$

  - 对倒谱做CMS为：

$$X = (\bar{x}_1 - \bar{\mu}, \bar{x}_2 - \bar{\mu}, \ldots, \bar{x}_T - \bar{\mu})$$

P3 系统鲁棒性问题

- 倒谱方差规整(Cepstral Variance Normal)
  - 把语音的特征矢量规整到0均值，1方差的分布内
  - 计算一段语音MFCC特征序列 $X=(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_T)$ 的方差：

$$\vec{\sigma}=\sqrt{\frac{1}{T}\sum_{i=1}^{T}(\vec{x}_i-\vec{\mu}_i)^2}$$

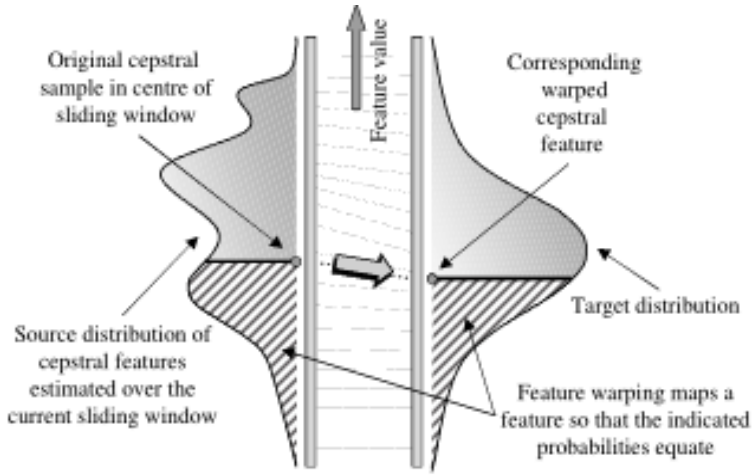
  - 对整段语音做减均值除以方差

$$\hat{\vec{x}}_i=\frac{(\vec{x}_i-\vec{\mu}_i)}{\vec{\sigma}_i}$$

  - 一般先做CMS再做CVN

P3 系统鲁棒性问题

- 特征弯折(Feature Warping)
  - CMS和CVN可以去除线性噪声，但加性噪声会降低特征参数的方差，从而改变特征参数的分布，使之不满足正态分布，从而严重影响系统性能。特征弯折就是将特征参数映射到标准正态分布，从而增强系统鲁棒性。



P3 系统鲁棒性问题

特征弯折(Feature Warping)

特征弯曲的方法，首先假设特征参数各维之间是相互独立的，特征参数中的每一维元素可独立进行弯曲变换。给定语音的 MFCC 参数序列 $X=(\bar{x}_1,\bar{x}_2,...,\bar{x}_T)$，其中 T 为参数序列的帧数。通过一个长度为 N 的矩形窗截取特征参数序列，将窗中特征参数的第 i 维元素按降序排列，其中窗中心的特征值 x_i 排第 R 位。设 $g(x)$ 为特征弯曲的目标映射函数，则有：

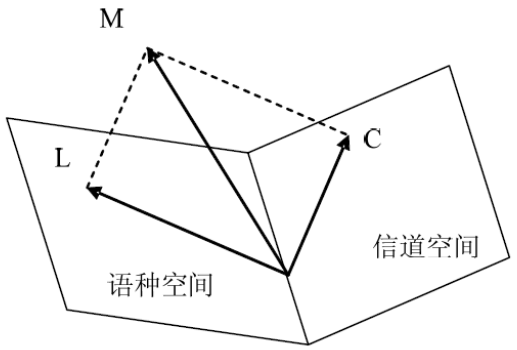
$$\frac{N+0.5-R}{N}=\int_{-\infty}^{m}g(x)dx \qquad (3.46)$$

$$g(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$$

P3 系统鲁棒性问题--因子分析

- 因子分析(Factor Analysis)基本思想
  - 所有语音信息都包含在GMM均值超向量中
  - 在模型估计和得分判决时，不但考虑说话人相关的成分，而且还会考虑信道的成分
  - 说话人空间和信道空间是相互独立的，都可以用一个较低维数的空间来表征
  - 一段语音的超矢量由语种/说话人超矢量和信道超矢量构成

$$M = L + C$$



P3 系统鲁棒性问题--因子分析

- 线性判别分析（LDA）
  - Linear Discriminant Analysis, LDA
  - 线性鉴别分析的基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果，投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性。因此，它是一种有效的特征抽取方法。使用这种方法能够使投影后模式样本的类间散布矩阵最大，并且同时类内散布矩阵最小。就是说，它能够保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离
  - 它和主成分分析（PCA）的区别在于，PCA所求的最优分界面使得原向量在其上的投影最大，而LDA的主要目标在于使得投影后的向量更具区分性。

P3 系统鲁棒性问题--因子分析

- 线性判别分析（LDA）
  - 目标：最大化类间方差，最小化类内方差
  - 类间协方差矩阵和类内协方差矩阵

$$S_b = \sum_{s=1}^{S} (\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^t \qquad S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \bar{w}_s)(w_i^s - \bar{w}_s)^t$$

  - 投影矩阵A的计算

$$S_b v = \lambda S_w v$$

  - 特征映射

$$\varphi(w) = A^t w$$

  - 经过LDA补偿后的余弦核函数

$$k(w_1, w_2) = \frac{(A^t w_1)^t (A^t w_2)}{\sqrt{(A^t w_1)^t (A^t w_1)} \sqrt{(A^t w_2)^t (A^t w_2)}}$$

P4 深度学习

- DNN-ivector

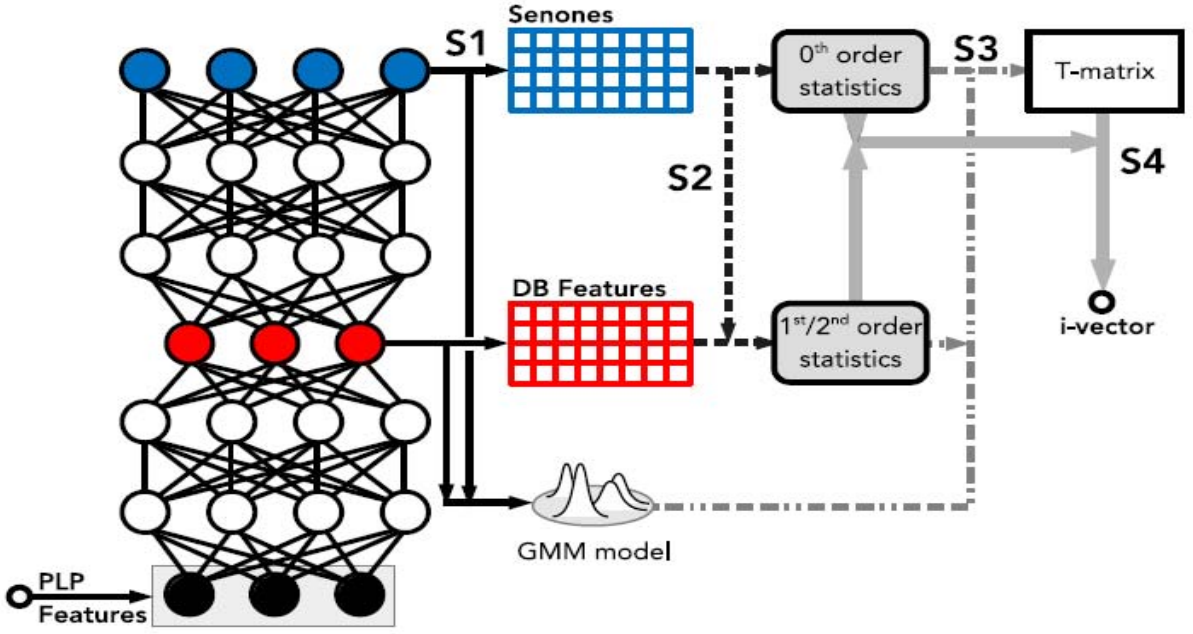
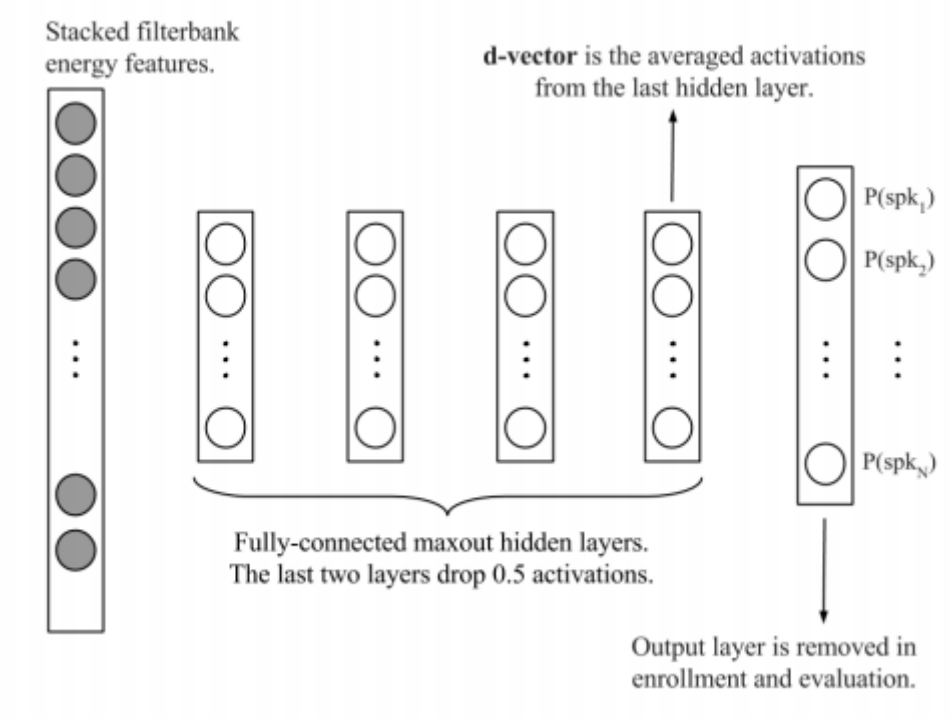


Fig. 3. Structure of the DNN/i-vector system based on senone statistics.

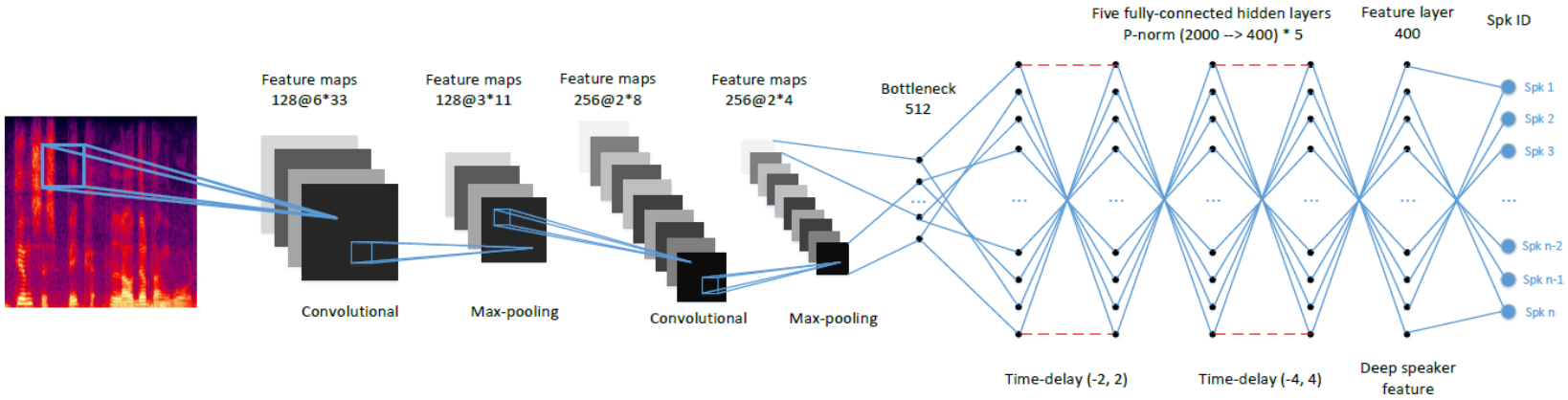
P4 深度学习

- Deep Speaker



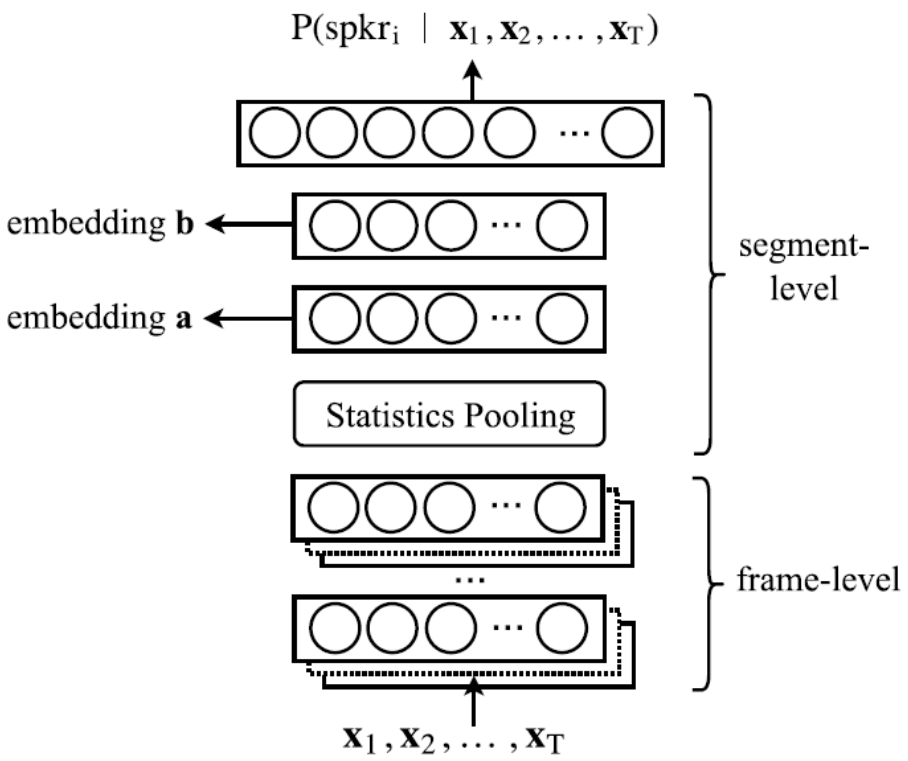
P4 深度学习

- Deep Speaker



P4 深度学习

- Deep Speaker



语种识别
Language Identification (LID)

提纲

- 概述
- 语种识别原理和系统
  - 特征提取
  - GMM系统
  - SVM系统
  - 语法系统
  - 鲁棒性问题
  - 因子分析问题

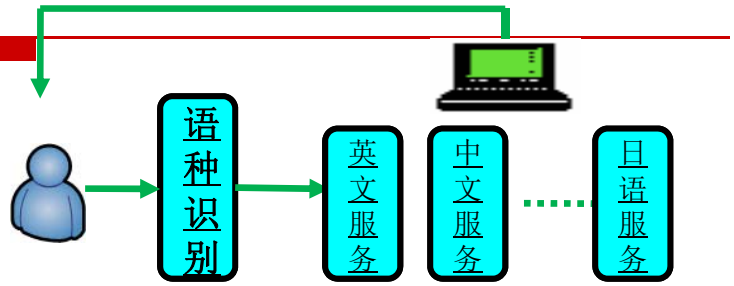
P1 概述—定义

- **语种识别**：称为语种辨识，它是通过分析处理一个语音片段以判断其属于某个语言种类的过程，其本质是语音识别的一个方面。

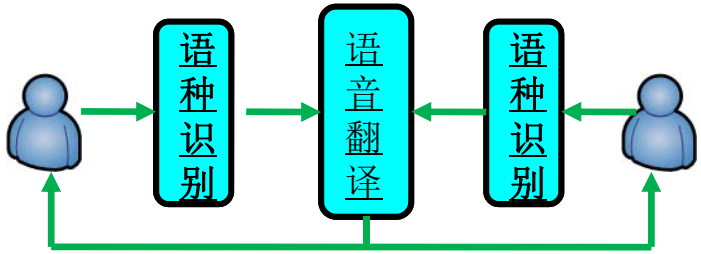
  aims to determine the language identity of a given speech segment

P1 概述—应用

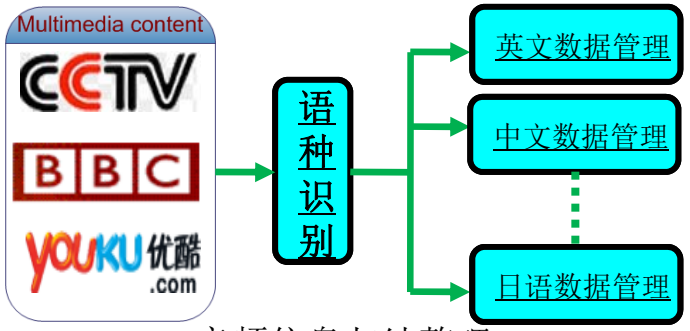
- “人-机”多语种服务系统
  - 自动识别用户语种信息，启动相应自动服务系统
- “人-人”跨语种交互系统
  - 识别对话人语种信息，启动自动语音翻译系统
- 音频信息检索
  - 识别音频文件语种信息，进行音频文件分类整合
- 国家安全领域



“人-机”多语种服务系统



“人-人”跨语种交互系统



音频信息归纳整理

P1 概述—研究现状

- 1980年以前: 开始采用统计学习方法（TI公司、DARPA）
- 1980~1990: 采用LPCC,基频，能量包络等特征和矢量量化方法
- 1990~1994: PRLM技术、 GMM模型被广泛采用（Zissman、VW Zue等）
- 1994~2000： PPRLM框架获得成功（Y Yan、Zissman、Barnard等）
- 21世纪: 采用Lattice-based PPRLM （J Gauvain 、 D Reynolds等
  HMM-ANN架构音子识别器（BUT）
  GMM-MIE, GMM-SVM
  多系统融合技术
- 近两年 ：因子分析技术（Ivector）

- 研究单位
  - MIT, OGI, BUT, QUT, LPT, SRI, LIMSI 等
  - 北京大学，清华大学，信息工程大学，中科大，自动化所，声学所

概述—NIST评测

- NIST ---- National Institute of Standards and Technology
- LID评测：1996, 2003, 2005, 2007, 2009, 2011
- 任务：给定一句测试语音，判断是否包含目标语种

The 2009 NIST language recognition evaluation task is language detection: Given a segment of speech and a language hypothesis (i.e., a target language of interest to be detected), the task is to decide whether that target language was in fact spoken in the given segment (yes or no), based on an automated analysis of the data contained in the segment.

概述—NIST评测

- NIST LRE 2009 目标语种

Table 1: The LRE09 target languages

Amharic	Bosnian	Cantonese
Creole (Haitian)	Croatian	Dari
English (American)	English (Indian)	Farsi
French	Georgian	Hausa
Hindi	Korean	Mandarin
Pashto	Portuguese	Russian
Spanish	Turkish	Ukrainian
Urdu	Vietnamese	

P1 概述—评价指标

- 语种确认(Language Identification):指出测试样本是否为目标语种，“二选一”；
  - DET(Detection Error Tradeoff)曲线
  - 等错率(Equal Error Rate)
- 语种辨识(Language Verification)：指出测试样本是哪个目标语种，“多选一”；
  - 正确率

P1 概述—评价指标

- DET曲线
  - 漏检率

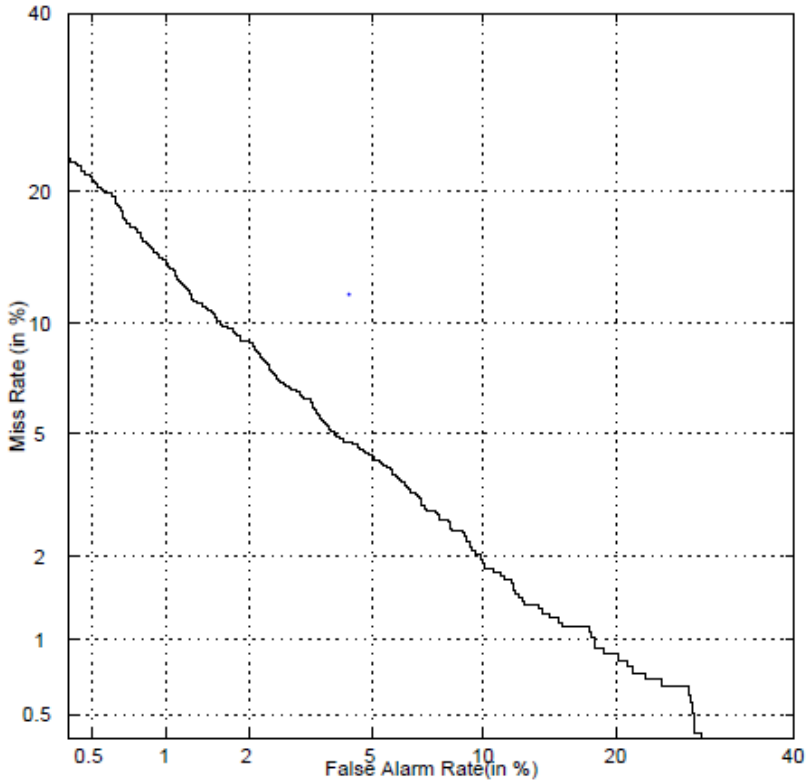
$$\text{漏检率} = \frac{\text{漏检的次数}}{\text{答案为“是”的判决总数}} \times 100\%$$

  - 虚警率

$$\text{虚警率} = \frac{\text{虚警的次数}}{\text{答案为“否”的判决总数}} \times 100\%$$

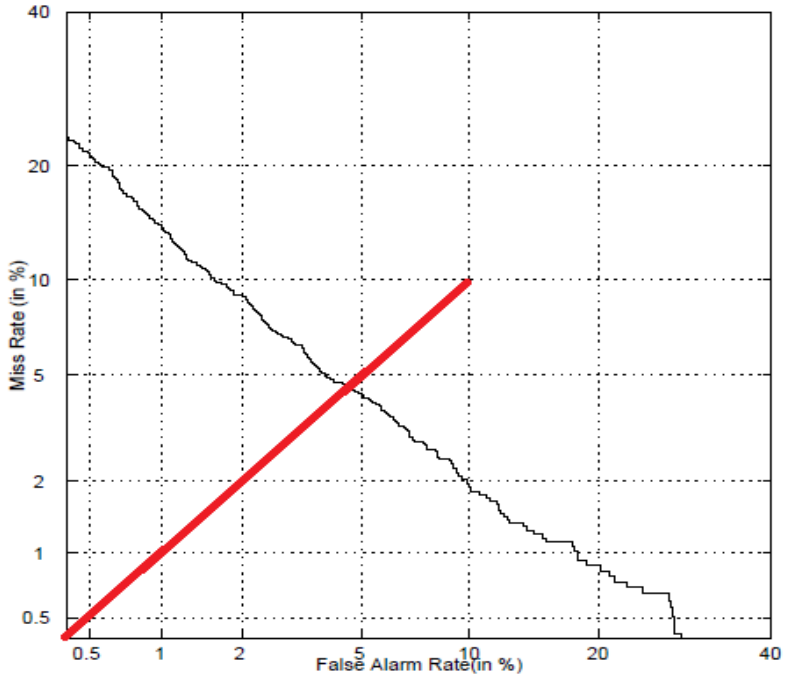
P1 概述—评价指标

- DET曲线



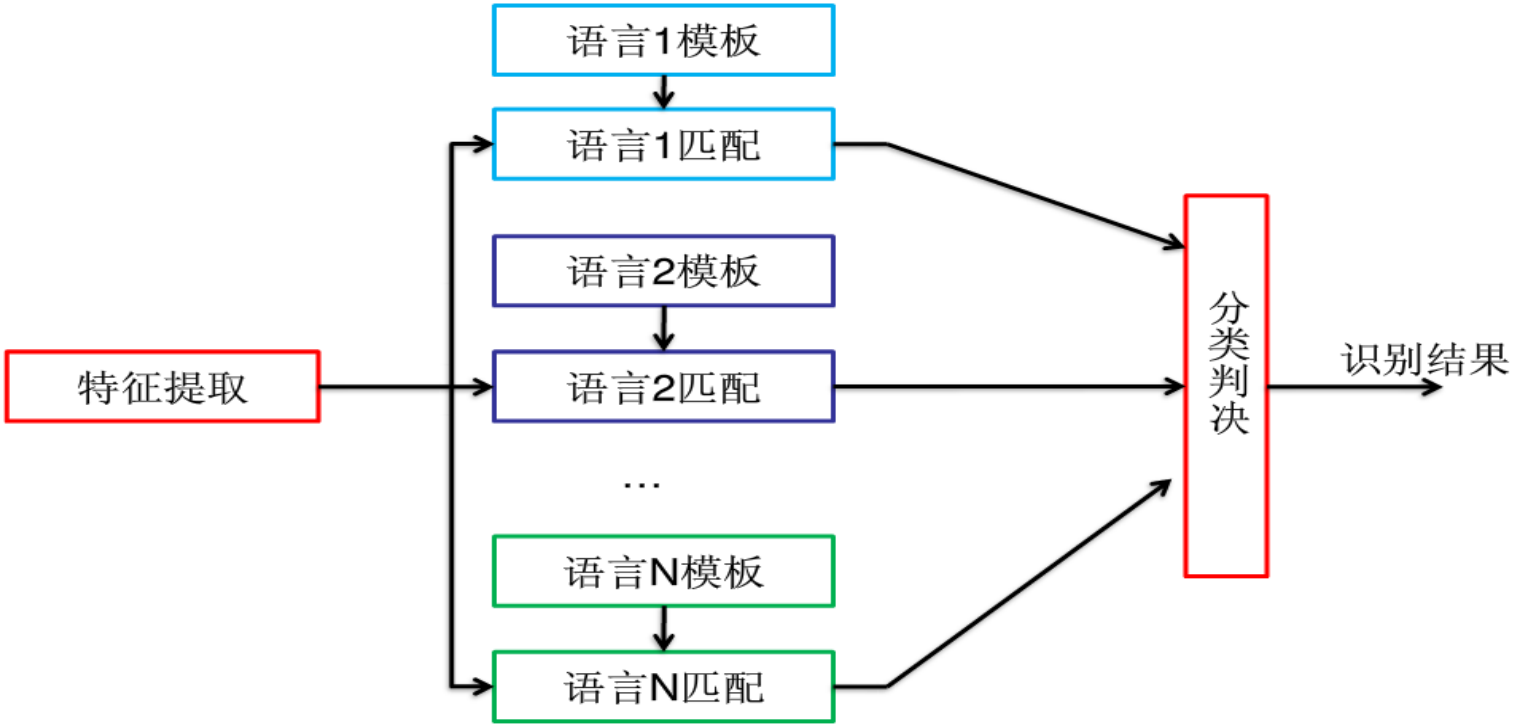
P1 概述—评价指标

- EER : Miss和False相等时， DET曲线上的点对应的值



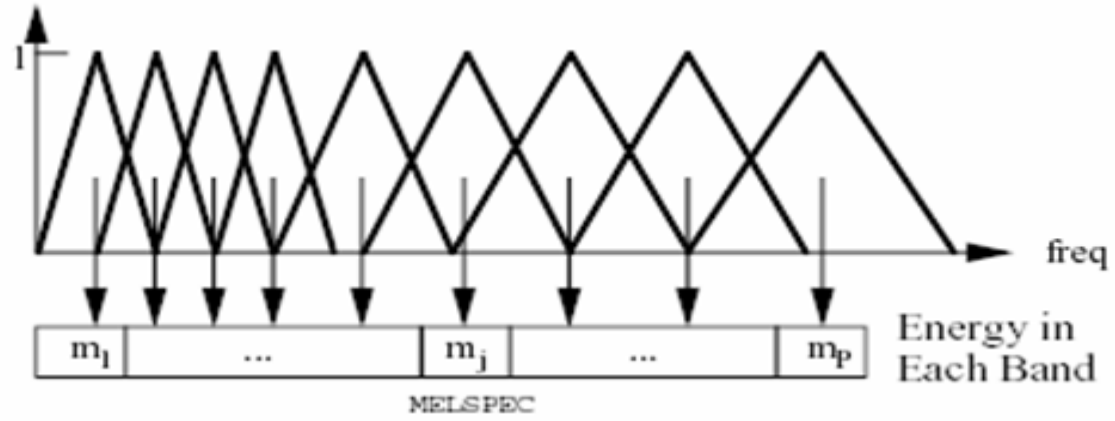
P2 语种识别原理和系统

基本原理框图



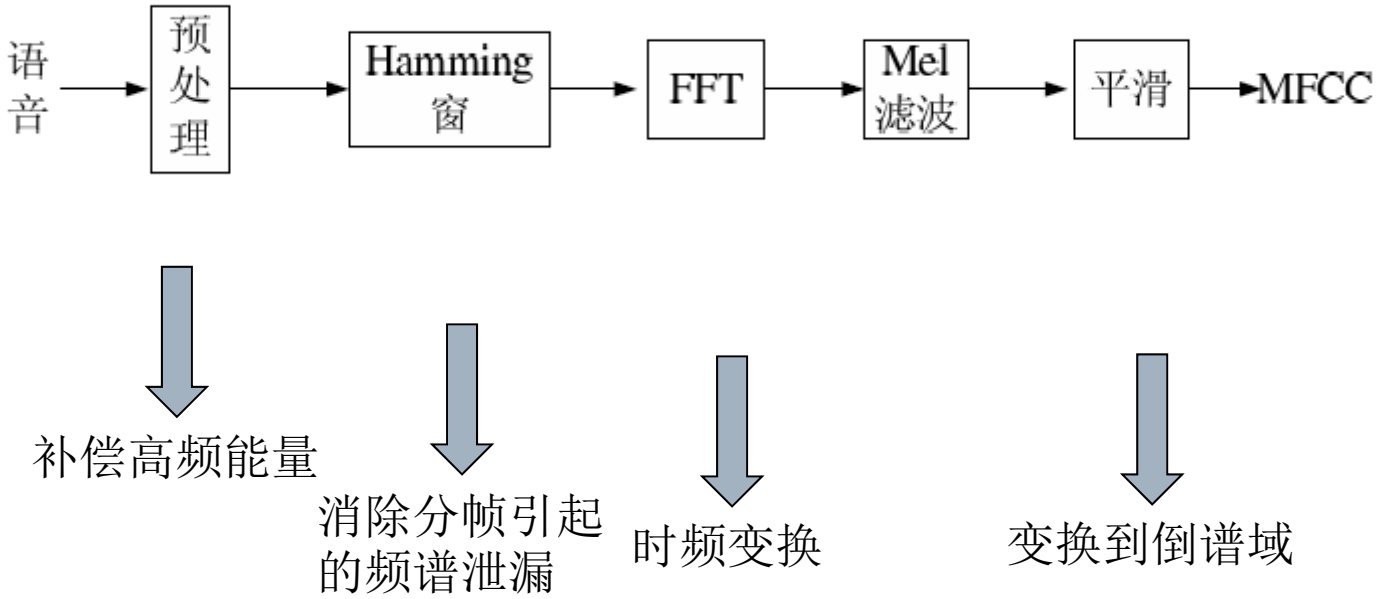
P2.1 特征提取

- 声学特征MSDC（Mel shifted Delta Coefficients）
  - 梅尔倒谱: 美尔倒谱系数（MFCC）模拟人耳对语音的感知，由于人耳对低频部分比较敏感，而对高频部分不敏感，因此，美尔滤波器是一组三角滤波器，其在低频部分的分辨率较高，而高频部分的分辨率较低



P2.1 特征提取

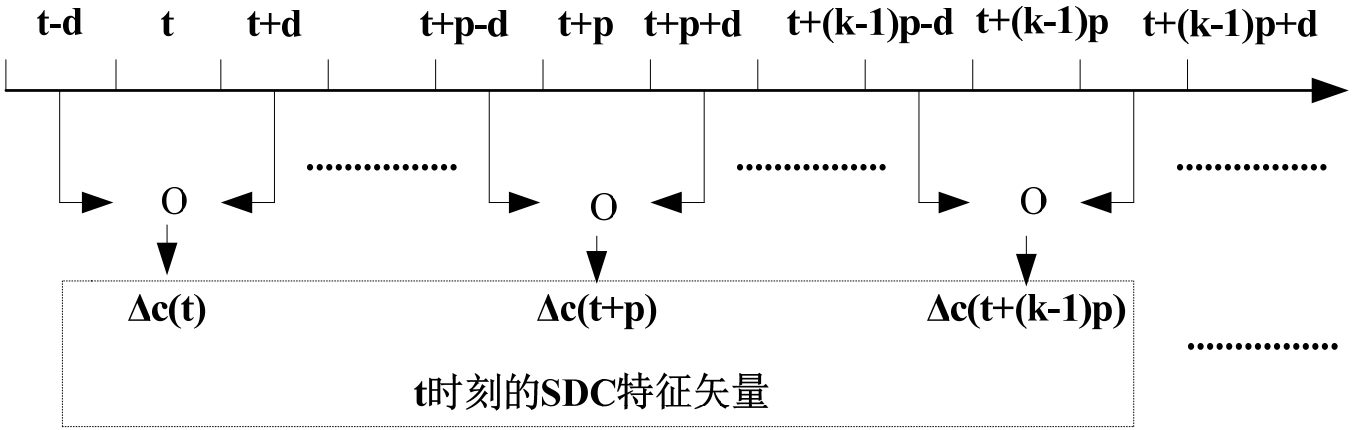
- **声学特征MSDC（Mel shifted Delta Coefficients）**
  - 梅尔倒谱提取算法



P2.1 特征提取

声学特征MSDC（Mel shifted Delta Coefficients）

- 移动查分变换



N代表了计算的倒谱的阶数；d代表了时间的偏移；P代表了模块与模块之间的偏移；k代表了模块的数目。语种识别的系统性能与四个参数设置有直接关系。根据国际上最好系统的经验，我们选择的参数配置为（7，1，3，7）

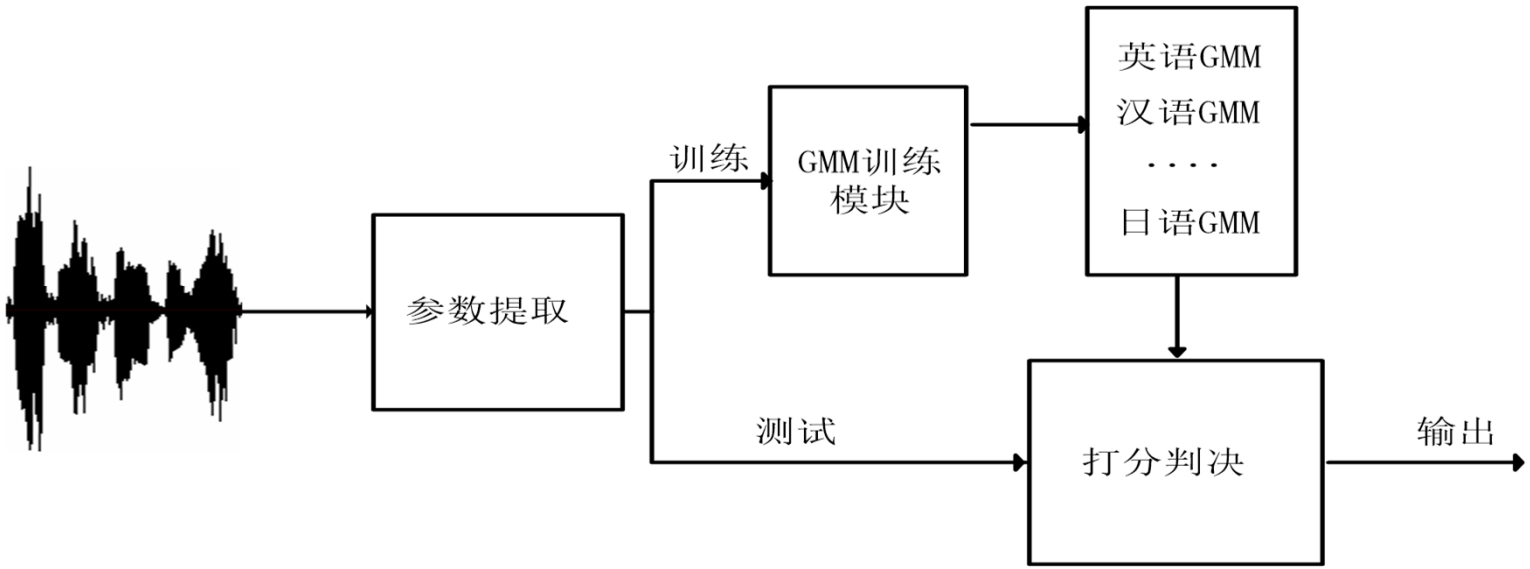
P2.1 特征提取

- 特征比较试验结果
  - 13维MFCC ＋1阶短时差分＋2阶短时差分
  - 7维MFCC＋移动差分变换 SDC(7-1-3-7)

EER (%)	LRE 03	相对下降	LRE 05	相对下降
MFCC 39维	14.2	—	16.9	—
MFCC-SDC 56维	9.3	34.5	11.6	31.4

P2.2 GMM系统框架

- 基本原理框图



P2.2 GMM系统框架

- **GMM原理:**

高斯混合模型可以描述为一个由M个高斯分量组成的高斯混合密度函数,

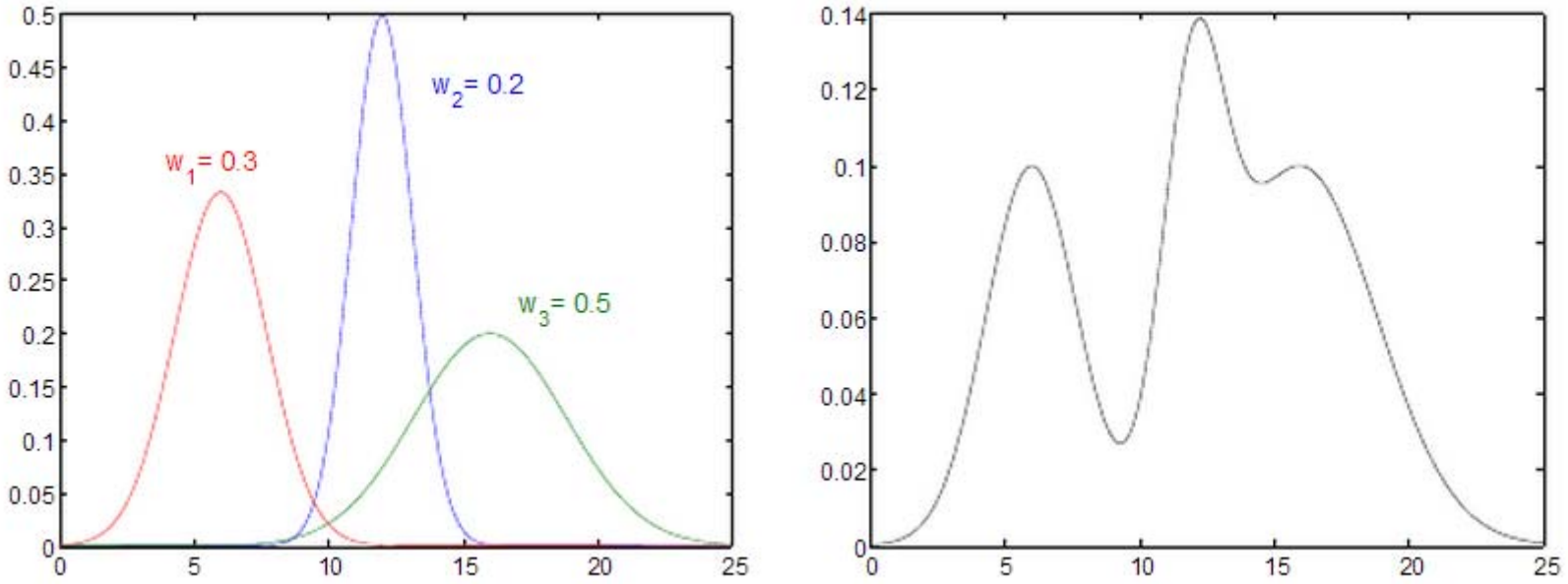
$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x})$$

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)^r \sum_i^{-1} (\vec{x} - \vec{\mu}_i) \right\}$$

$$\sum_{i=1}^{M} w_i = 1$$

P2.2 GMM系统框架

GMM原理



P2.2 GMM系统框架

- **GMM原理:**

一个M分量的高斯混合模型可以表示为：

$$\lambda = \{\omega_i; \vec{\mu}_i; \Sigma_i\} \qquad i = 1,\ldots M$$

对于一段给定的训练语音序列 $X = \{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_T\}$

特征序列对于GMM的似然度，定义为：

$$p(X|\lambda) = \prod_{i=1}^{T} p(\vec{x}_t|\lambda)$$

P2.2 GMM系统框架

- **GMM原理:**

一个M分量的高斯混合模型可以表示为：

$$\lambda = \{\omega_i; \vec{\mu}_i; \Sigma_i\} \qquad i = 1,\ldots M$$

对于一段给定的训练语音序列 $X = \{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_T\}$

特征序列对于GMM的似然度，定义为：

$$p(X|\lambda) = \prod_{i=1}^{T} p(\vec{x}_t|\lambda)$$

P2.2 GMM系统框架

GMM原理:

模型训练采用EM（Expectation Maximum）算法: 使训练出来的模型参数与训练语音的似然度最大，即达到最佳匹配程度。

STP1：初始化模型参数 λ

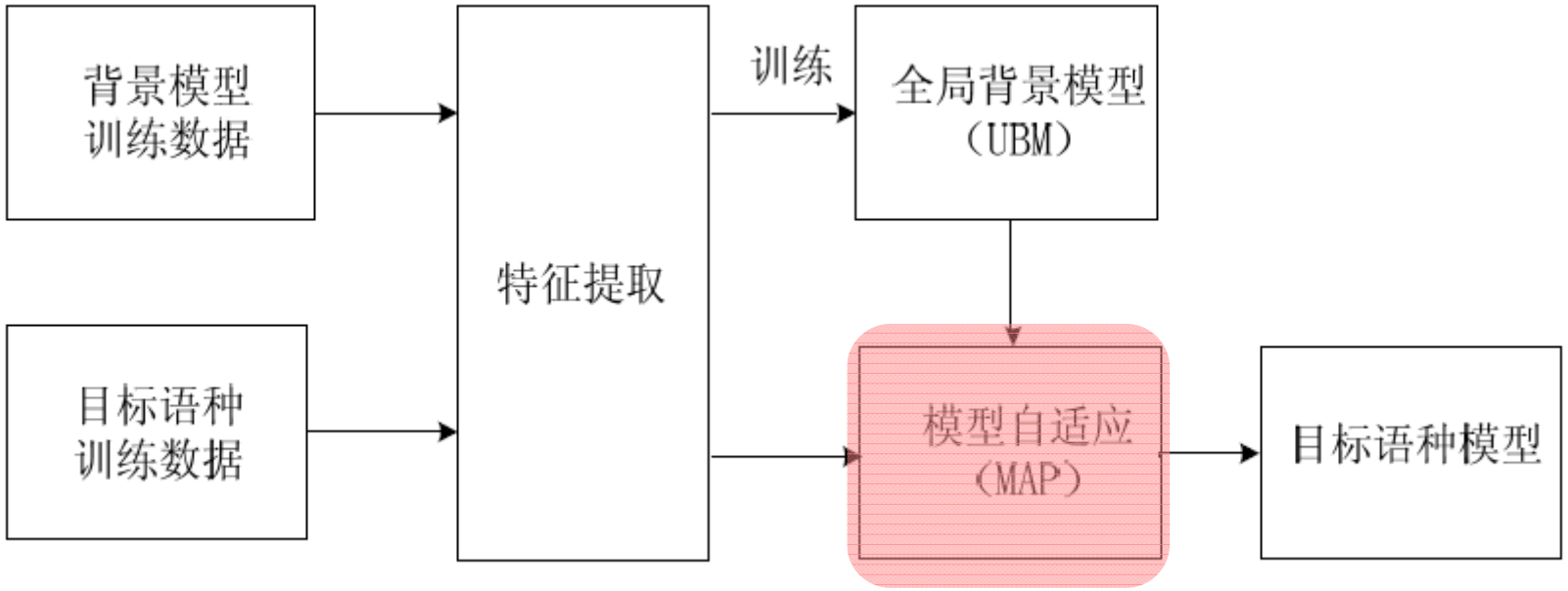
STP2: 更新参数。根据公式，重新估计参数 $\overline{\lambda}$,满足

$$p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$$

STP3：重复以上步骤，直至收敛。

P2.2 GMM系统框架

- 目标语种建模过程



P2.2 GMM系统框架

- 最大后验概率估计(Maximum a posteriori)：可以用较少的数据获得较准确的模型

$$\lambda \to f(\lambda|X) = \frac{f(X|\lambda)g(\lambda)}{\int_{\theta\in\Theta} f(X|\lambda')g(\lambda')d\lambda'}$$

$$\hat{\lambda}_{MAP} = \arg\max_{\theta} \frac{f(X|\lambda)g(\lambda)}{\int_{\theta\in\Theta} f(X|\lambda')g(\lambda')d\lambda'} = \arg\max_{\theta} f(X|\lambda)g(\lambda)$$

- 对比MLE： $\hat{\lambda}_{MLE} = \arg\max_{\lambda} f(X|\lambda)$

P2.2 GMM系统框架

- **训练数据选择要求：**
  - 训练数据包括尽可能多的说话人和语种语音。
  - 针对电话语音的情况，训练数据要覆盖尽可能多的信道。
  - 训练数据要均衡。这里包括各语种、性别、说话人，以及各个信道之间的数据均衡。
  - 训练数据中包括尽可能多的语音现象，即各种音素的语音现象，如浊音，清音鼻音，擦音等。

P2.2 GMM系统框架

- 判决模块：

给定一段测试样本序列 X_r ,和L个目标语种，计算样本序列和某个语种的置信度，描述假设检验问题为：

H_0: X_r 来源于目标语种

以及

H_1: X_r 不是目标语种。

贝叶斯判据如下：

$$\Lambda(X) = \frac{P(X|H_0)}{P(X|H_1)} \begin{cases} \geq threshold & \text{测试语音是目标语种} \\ < threshold & \text{测试语音不是目标语种} \end{cases}$$

P2.2 GMM系统框架

- 判决模块：

通常比值很小，因此取对数，得到对数似然比（Log likelihood ratio, LLR）,判决函数表示为：

$$\Lambda(X) = \log P(X|H_0) - \log P(X|H_1)$$

由于全局背景模型描述了多语种特征参数分布，非目标语种模型可以采用全局背景模型。

$$\Lambda(X) = \frac{1}{T}\sum_{t=1}^{T}[\log p(x_t \mid \lambda_{target}) - \log(p(x_t \mid \lambda_{ubm})]$$

P2.2 GMM系统框架

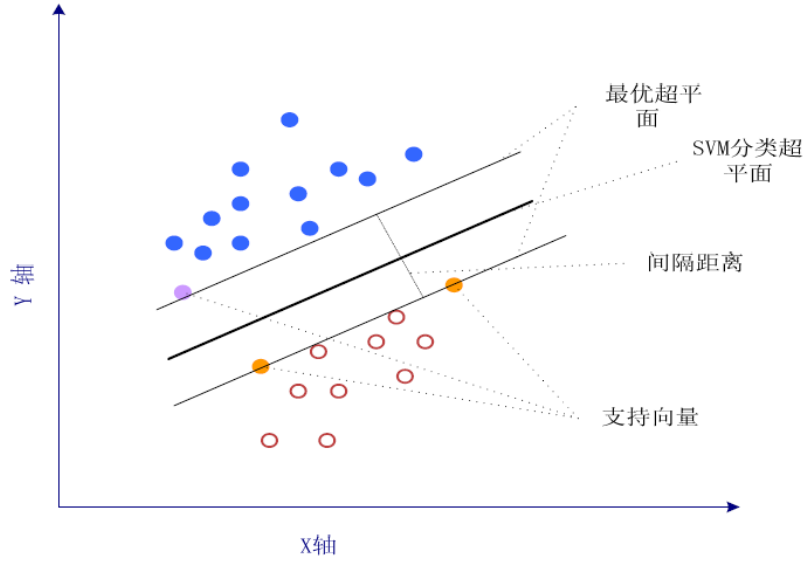
- **语种识别面临的问题：**
  - 说话人差异：发音情绪、讲话方式等
  - 信道差异：麦克风、固定电话、移动电话GSM和CDMA、小灵通等。
  - 噪声问题：信道噪声、环境噪声、对话抢入、垃圾语音等。
  - 不流畅、多语现象、发音变体等。

- **目标**
  - 寻找更具有语种区分度的特征
  - 增强系统的鲁棒性

P2.3 基于SVM的语种识别系统

- **支持向量机：Supported Vector Machine, SVM**

  是一种最大间隔分类器，通过最大间隔超平面将在特种空间中的线性可分的训练数据分开。



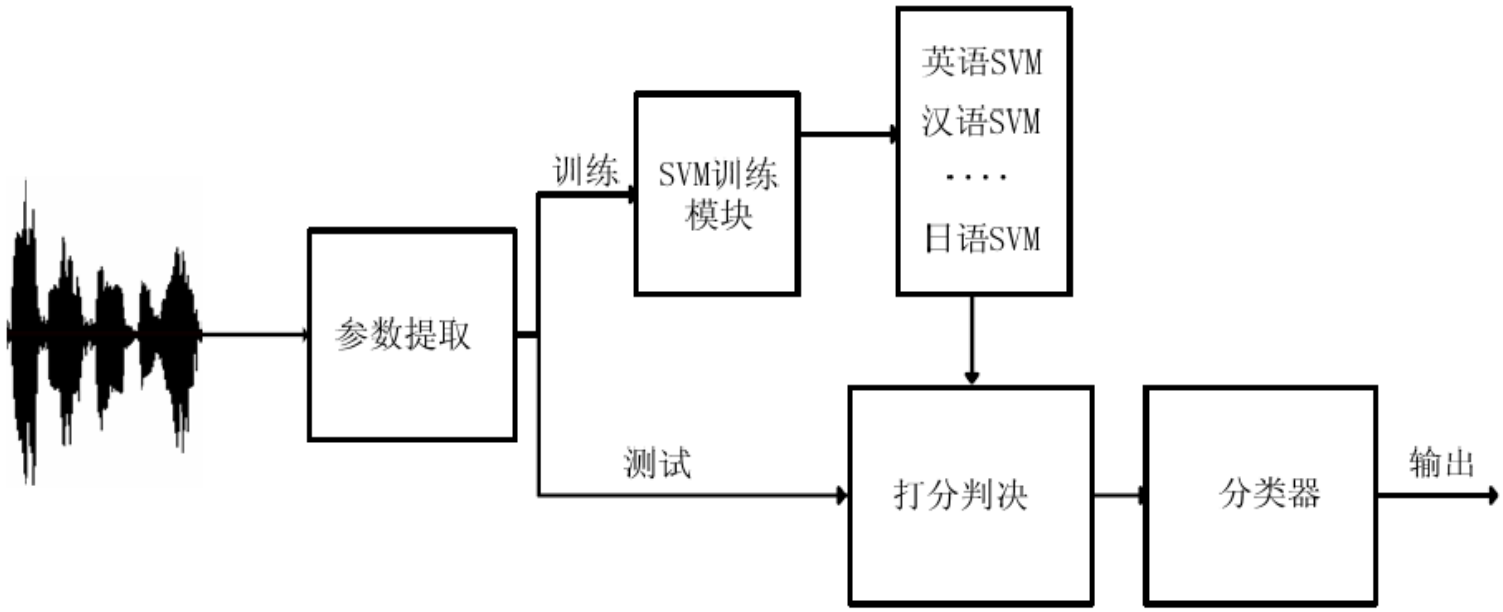
P2.3 基于SVM的语种识别系统

- **支持向量机常用工具包：**
  - LIBSVM：台湾大学，支持多类分类
  - SVM-LIGHT：美国康奈尔大学，二类分类
  - SVMTorch: Support Vector Machines for Large-Scale Regression Problems
  - LS-SVMLab: Matlab

    http://www.esat.kuleuven.be/sista/lssvmlab/

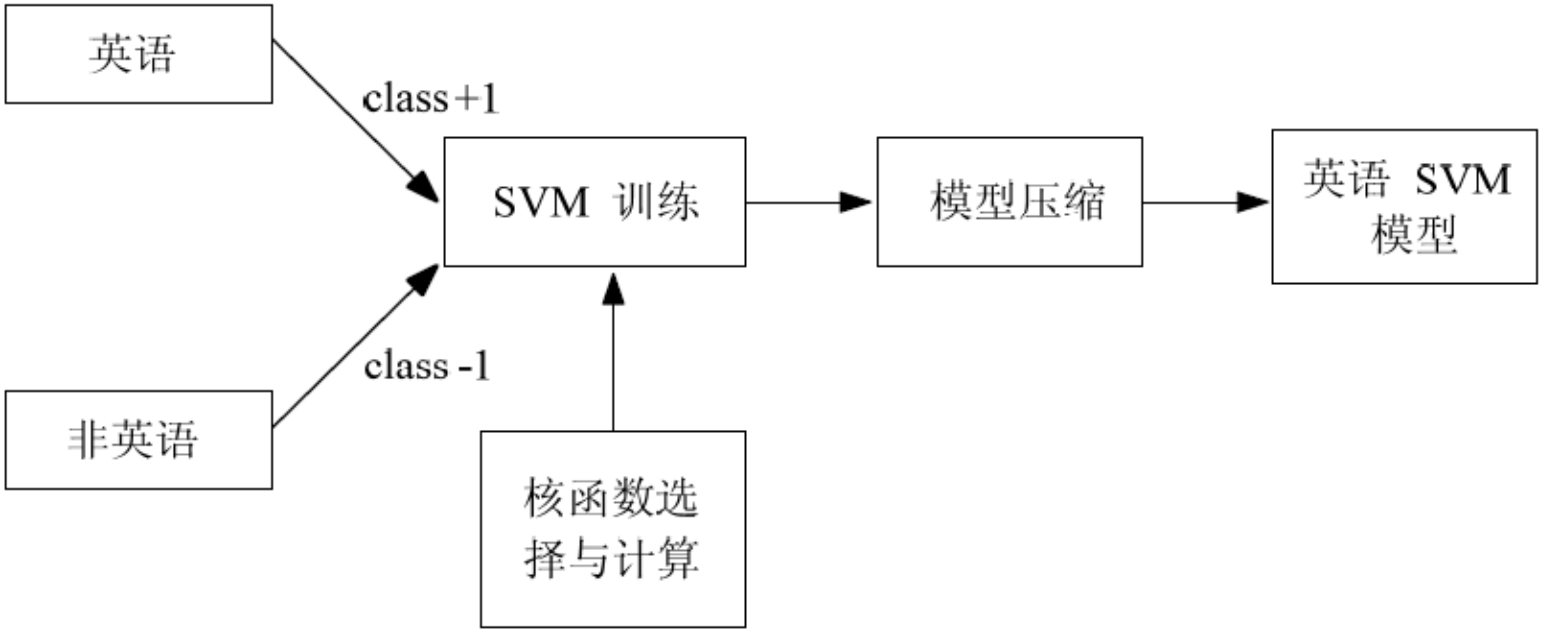
P2.3 基于SVM的语种识别系统

- SVM语种识别系统框图



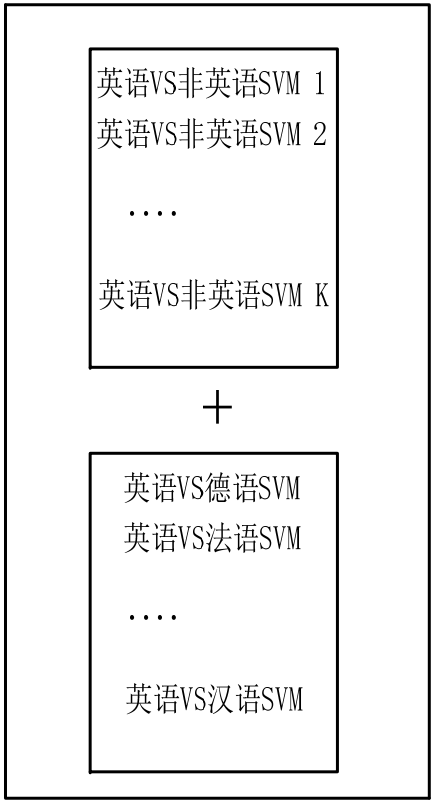
P2.3 基于SVM的语种识别系统

- 目标语种模型建模



P2.3 基于SVM的语种识别系统

- 几种改进的建模思路
  - 扩展一对多的模型
    - 例如：英语和非英语的区分，把英语作为正样本，把非英语作为负样本。假设目标语种的数目为N，负样本中因为包含了很多语种，所以其数据量应该是正样本的N－1倍。为了解决样本中数据平衡问题，通过聚类的方法，将负样本拆分成了K份。
  - 增加一对一模型
    - 例如：我们认为如果目标语种是法语，那么不仅仅能在英语对非英语SVM分类器的距离度量上偏向非英语，同时在英语和法语的SVM分类器上，更加明显的偏向法语。这样语种间的区分性，在高层的得分矢量上更加清晰的反映了出来，以方便后端的分类。



P2.3基于SVM的语种识别系统

- 基于GMM 的语种识别方法是对每个目标语种的发声特征进行概率分布建模，能够很好地描述不同语种的发声差异。

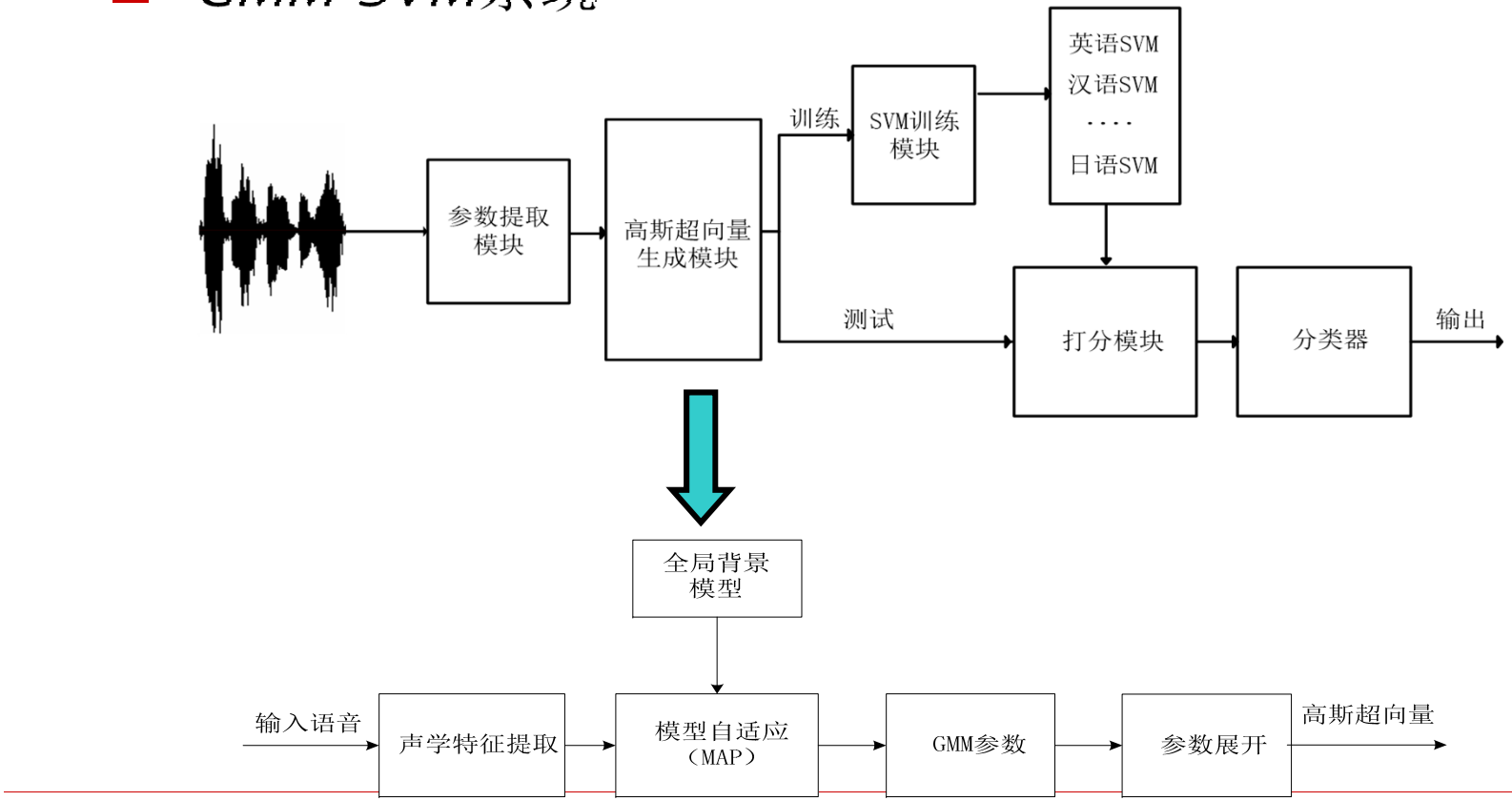
2006年MIT Campbell提出

GMM-SVM系统，简称GSV系统

- 基于SVM 的语种识别方法则是利用超平面在最大边界条件下将目标语种与非目标语种分开，具有很强的分类能力

P2.3基于SVM的语种识别系统

- GMM-SVM系统



P2.3 基于SVM的语种识别系统

- 高斯超矢量生成过程（Supervector）

对于输入语音 U_a，我们经过建模可以用一个GMM来描述：

$$g_a(x) = \sum_{i=1}^{N} \lambda_i N(x; \bar{m}_i^a, \Sigma_i)$$

则输入语音的高斯超矢量可以表示为：

$$\bar{m}_a = ((\sqrt{\lambda_1}\,\Sigma_1^{-1/2}\,\bar{m}_1^a)^t, (\sqrt{\lambda_2}\,\Sigma_2^{-1/2}\,\bar{m}_2^a)^t, \ldots (\sqrt{\lambda_N}\,\Sigma_N^{-1/2}\,\bar{m}_N^a)^t$$

P2.3 基于SVM的语种识别系统

- 结果比对

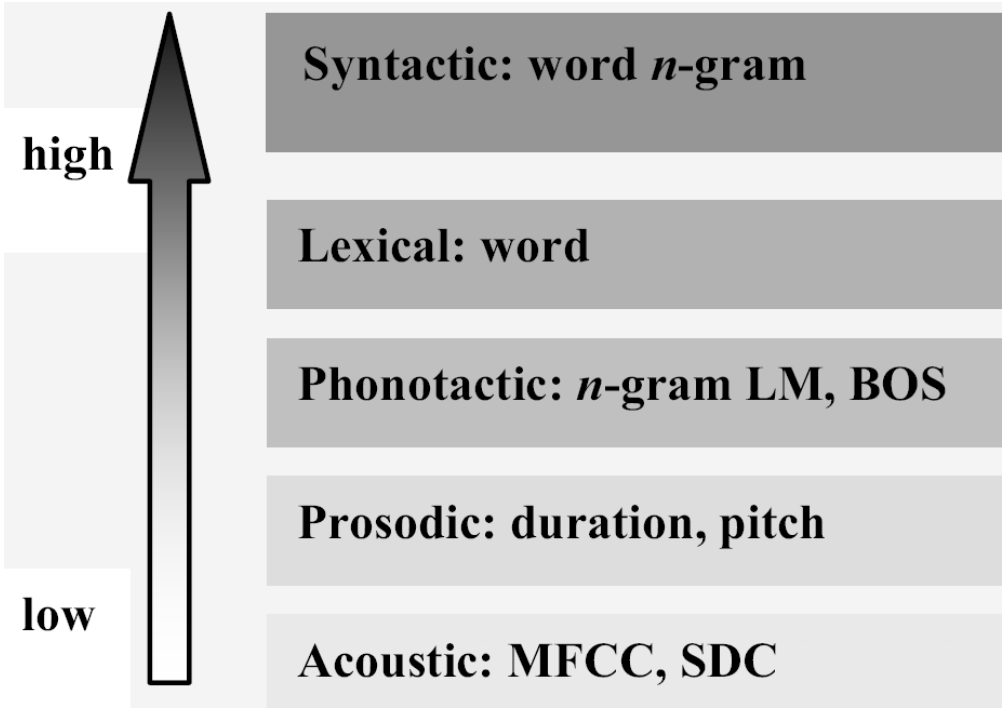
EER（%） / 系统	LRE03	LRE05
GMM 系统	4.8	8.3
SVM 系统	4.0	7.1
GSV 系统	2.9	4.0

P2.4 基于语法建模的语种识别系统

- 语音的表示单位
  - 从大到小：篇章、段落、句群、句子、短语、词、语素、音节、音素
  - 语音学研究对象：音节和音素
  - 音子：（phoneme，音素，音位）是人类语言中能够区别意义的最小声音单位
  - 每种语言都有自己的一组音子/音位。

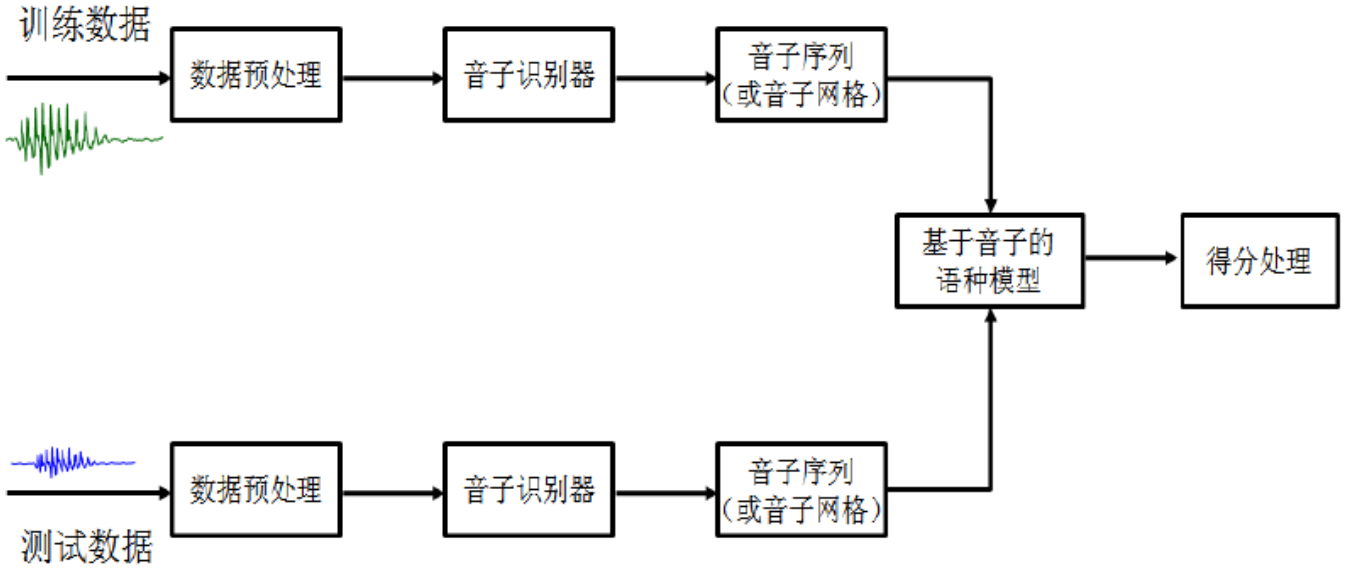
P2.4 基于语法建模的语种识别系统

- 不同层次的语种鉴别性信息



P2.4 基于语法建模的语种识别系统

- 语法建模，亦称音子建模，利用语音识别解码器，将语音信号表征成一串音子序列，对音子序列的建模采用了语言模型概念。



P2.4 基于语法建模的语种识别系统

- 音子识别器：

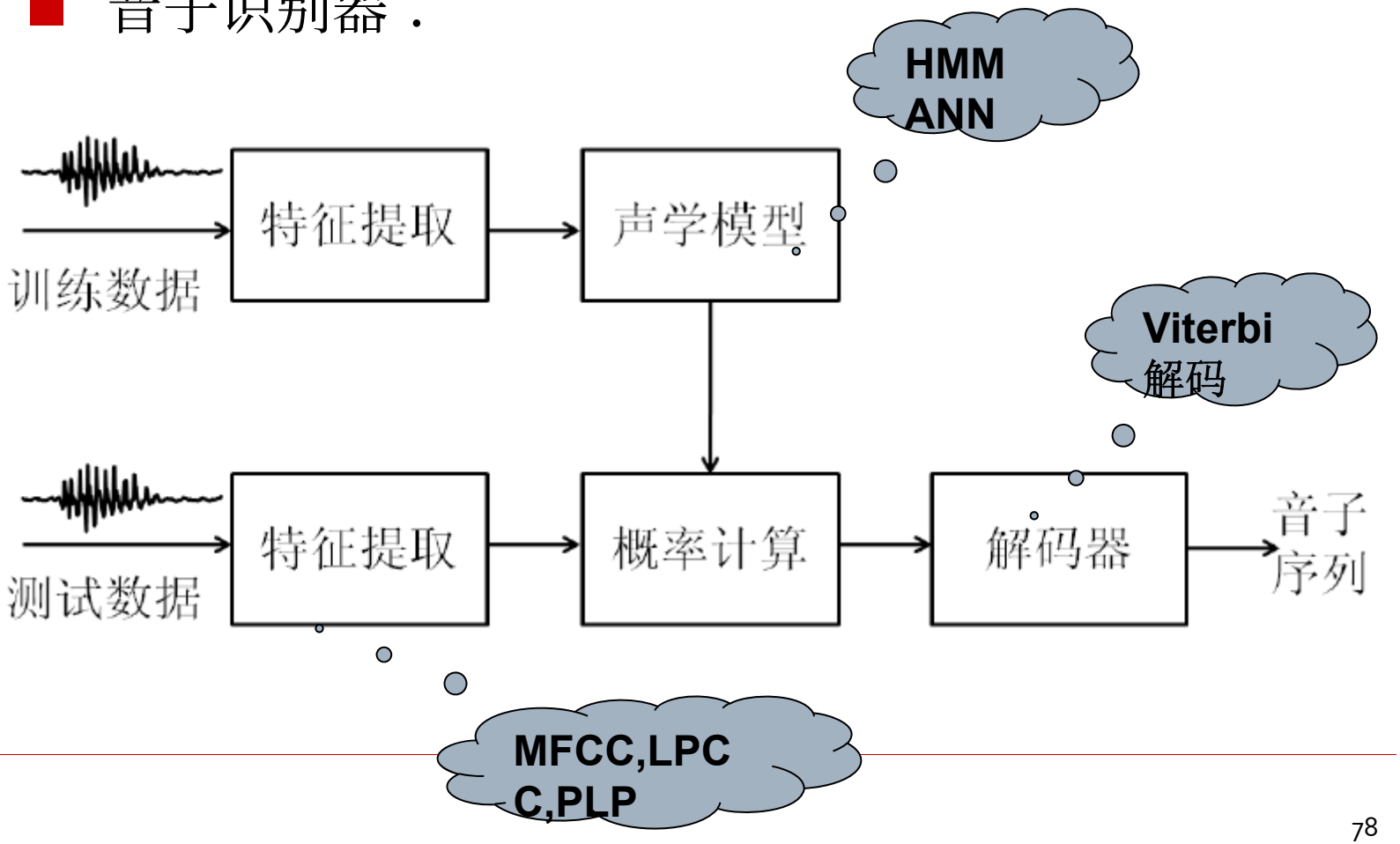
音子识别技术可以被视为音子层次的语音识别技术。

与大词表连续语音识别技术不同，音子识别技术并不关注语义信息或者语法信息的应用，而重点关注特征提取和声学模型建模，即音子识别技术重点考虑语音信号从声学层到音子符号之间映射的正确性与稳定性。

在语种识别领域，正是利用不同语种的音子配列结构不同进行语种区分。因此语种识别中采用的音子识别技术又可以被视为无语法约束或者弱语法约束的语音识别技术。

P2.4 基于语法建模的语种识别系统

- 音子识别器：



P2.4 基于语法建模的语种识别系统

- BUT音子识别器：
  - Brno University of Technology: http://speech.fit.vutbr.cz/
  - 开源软件包括：STK toolkit,
    PHNREC (phone recognizer)
    SNet/Tnet (neural net training software)
  - 包含四种语言：CZ, RU, HU, EN

```
phnrec -c PHN_EN_TIMIT_LCRC_N500 -l list -m out.mlf
#!MLF!#
"*/faem0.rec"
000000 1300000 pau
1300000 2000000 ah
2000000 3500000 s
3500000 4500000 ih
```

P2.4 基于语法建模的语种识别系统

- 音子语法模型建模—Ngram

给定一段音子序列 $W = \{W_1, W_2, ..., W_m\}$ 其统计概率表示为：

$$P(W) = \prod_{i=1}^{m} P(W_i | W_1, ..., W_{i-1})$$

Ngram语法模型采用N阶马尔可夫近似，假设每一个当前音子只与他前面的N个因子有关，概率表示近似为：

$$P(W) = \prod_{i=1}^{m} P(W_i | W_{i-N-1}, ..., W_{i-1})$$

$$P(W_i | W_{i-N-1}, ..., W_{i-1}) = \frac{C(W_{i-N-1}, ..., W_{i-1}, W_i)}{C(W_{i-N-1}, ..., W_{i-1})}$$

P2.4 基于语法建模的语种识别系统

- 音子语法模型建模—Ngram

常用CMU/MIT EN列表包含39个因子，建模所需参数量：

N = 1,2,3,4,5….

模型	参数量
unigram	38
bigram	39×38=1,482
trigram	39×39×38=57,798
fourgram	39×39×39×38=2,254,122
fivegram	39×39×39×39×38=87,910,758

P2.4 基于语法建模的语种识别系统

- Ngram音子模型平滑方法：
  - 插值平滑
  - 回退平滑
  - 加1平滑
  - Good-Turning方法
  - Witten-Bell方法
  - ……

P2.4 基于语法建模的语种识别系统

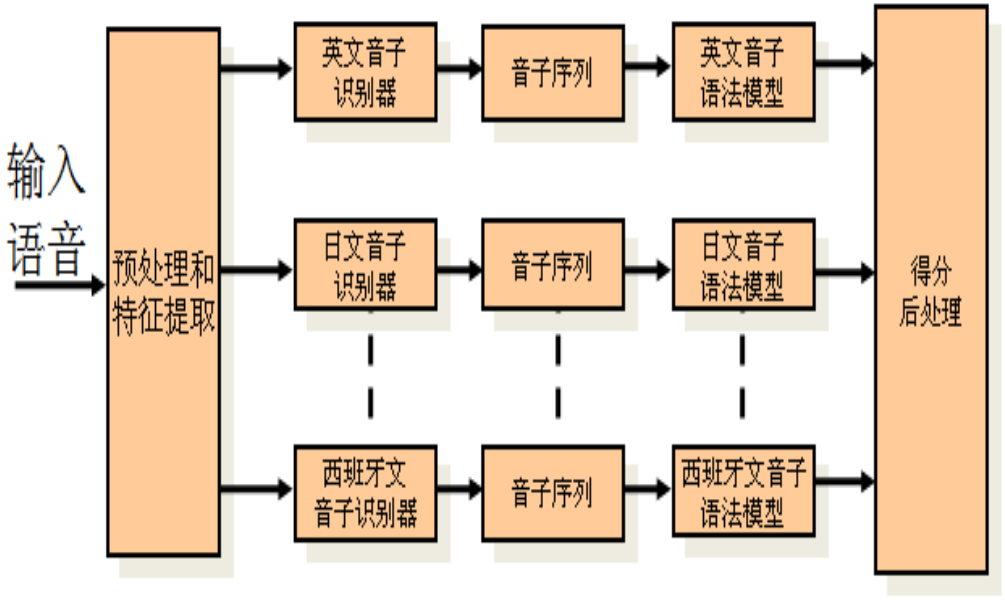
- 打分判决：模型复杂度（PP值，Perplexity）

$$PP = 2^{H(W)} = \left[P(W_1) \prod_{i=2}^{N} P(W_i | W_1, ..., W_{i-1}) \right]^{-\frac{1}{N}}$$

其中$H(W)$是音子语法模型与音子序列W之间的信息熵。直观上，如果 PP 值越小，该语法模型描述的概率分布越与音子序列W相吻合。因此在实验中，我们将负 PP 值作为一个语法模型与一段音子序列之间的“打分”，PP 值越小，该语法模型与音子序列“打分”越高。

P2.4 基于语法建模的语种识别系统

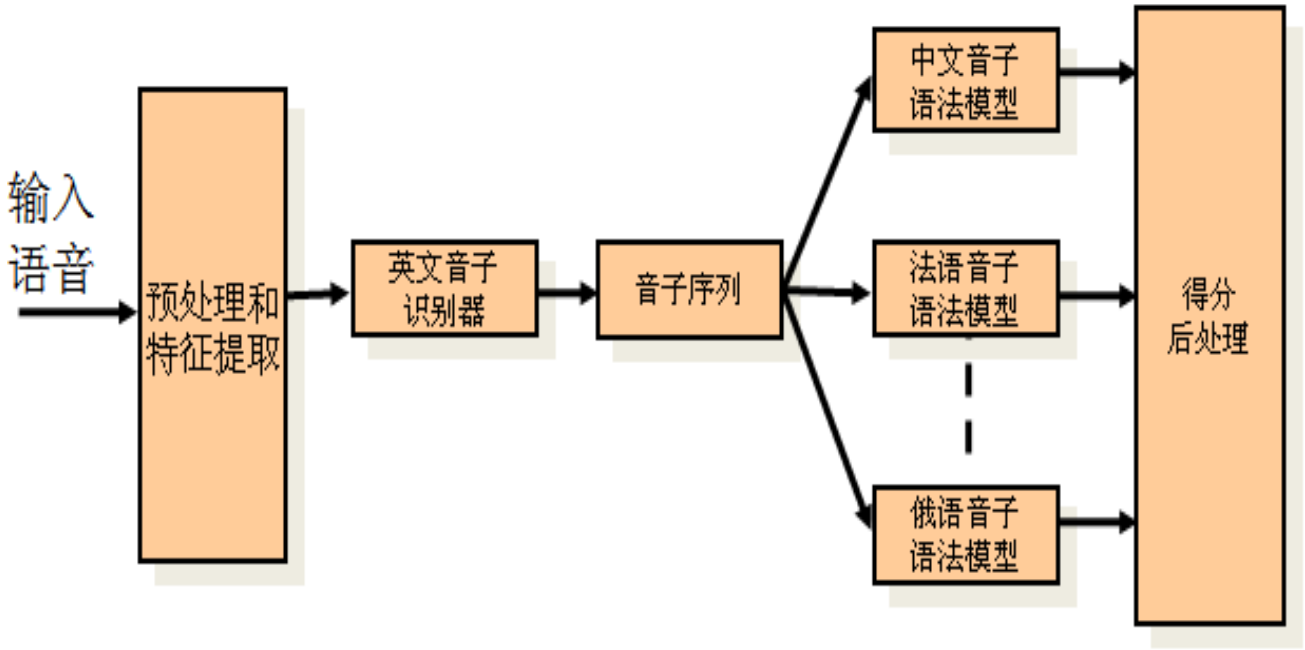
- PPR(Parallel Phone recognition)系统框架



- PPR(Parallel Phone recognition)系统缺点：
  - 为每个语种建立音子识别器和音子语法模型，数据标注和数据量问题；

P2.4 基于语法建模的语种识别系统

- PRLM系统
  - Phone Recognition followed by Language Modeling



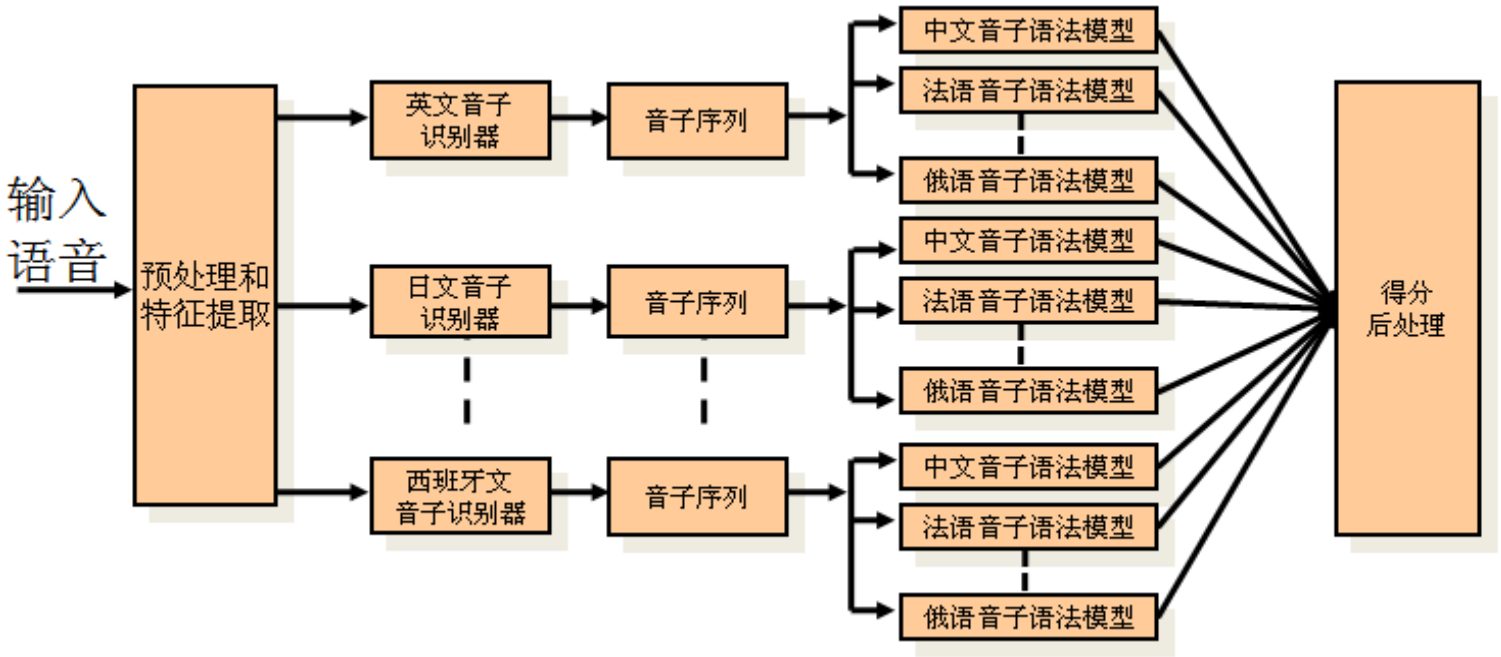
P2.4 基于语法建模的语种识别系统

- PRLM系统优点：
  - 单一音子识别器，解决对标注数据的依赖
  - 扩展性好，增加新语种只需要未标注数据

- 存在问题
  - 一是用一种语种的音子描述其他语种的语音，这种描述比较粗糙，不够精细。例如，在汉语中常见的一个音子或某种音调的变化，在英语中并不存在，那么用英语识别器来识别这个音子就只能找一个近似的音子来替代，甚至某种汉语的音调变化可能在英语中根本不存在。这种情况势必会降低系统的性能
  - 二是音子识别器本身不可避免的会带来识别错误也就是说用音子识别器的最优识别序列并不能准确的反映出原始语音数据的信息。

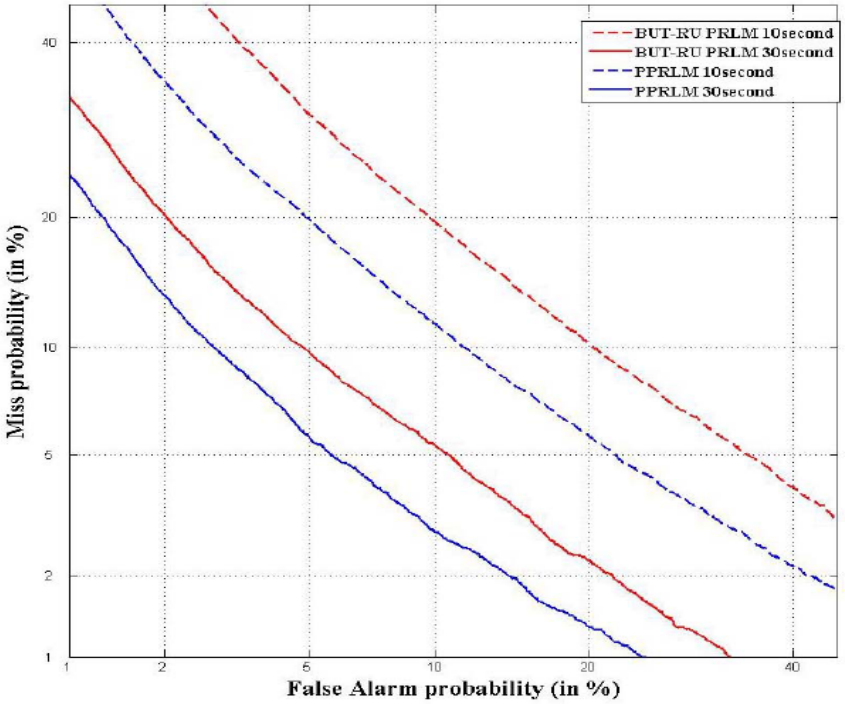
P2.4 基于语法建模的语种识别系统

- PPRLM系统框架：
  - Parallel Phone Recognition followed by Language Modeling



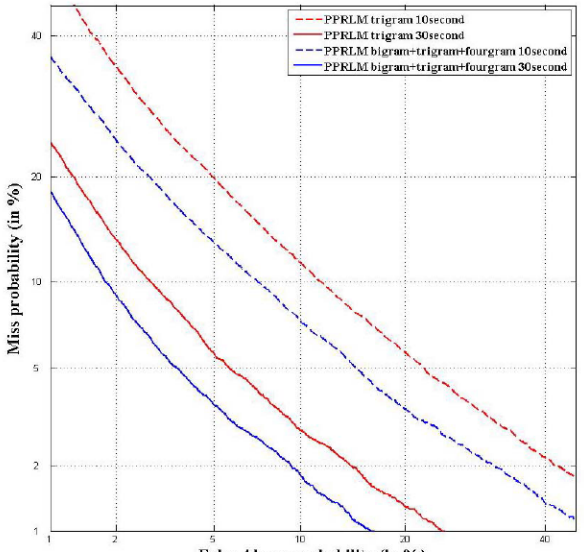
P2.4 基于语法建模的语种识别系统

- PRLM 与 PPRLM性能比较示例：



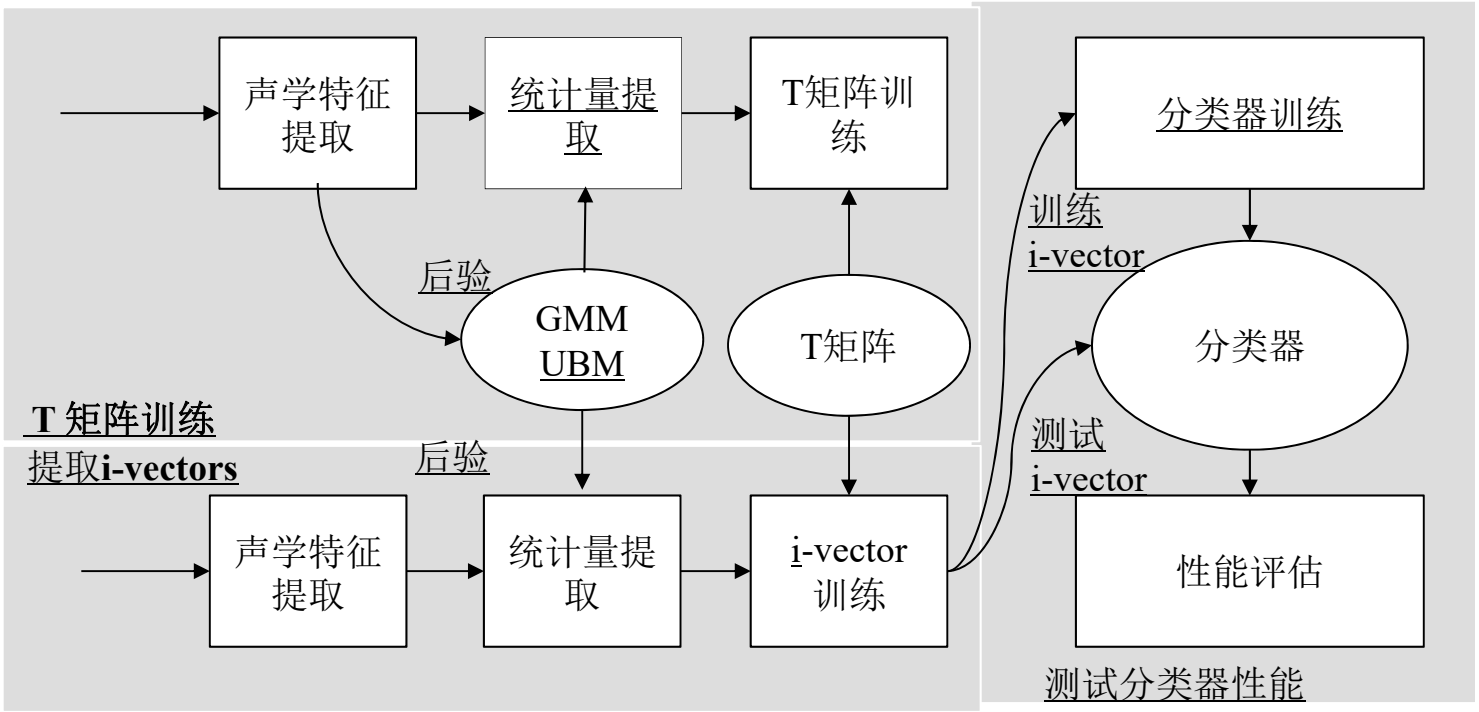
P2.4 基于语法建模的语种识别系统

- 语法系统需要考虑的问题：
  - 后端融合策略：
    - Log似然打分
    - SVM
    - LDA + 单高斯
    - ....
  - 语法模型的选取问题
    - Trigram
    - fourgram



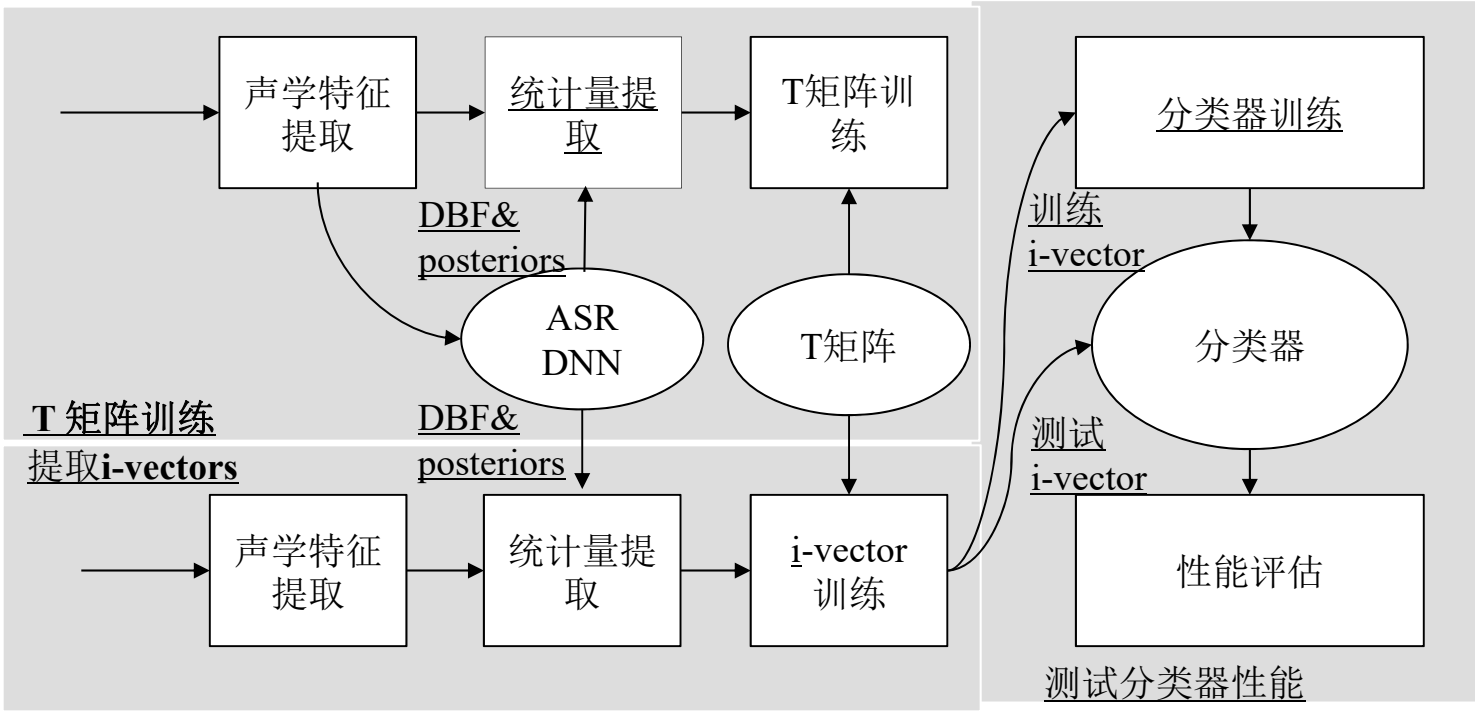
最新进展

传统 i-vector 说话人/语种识别系统

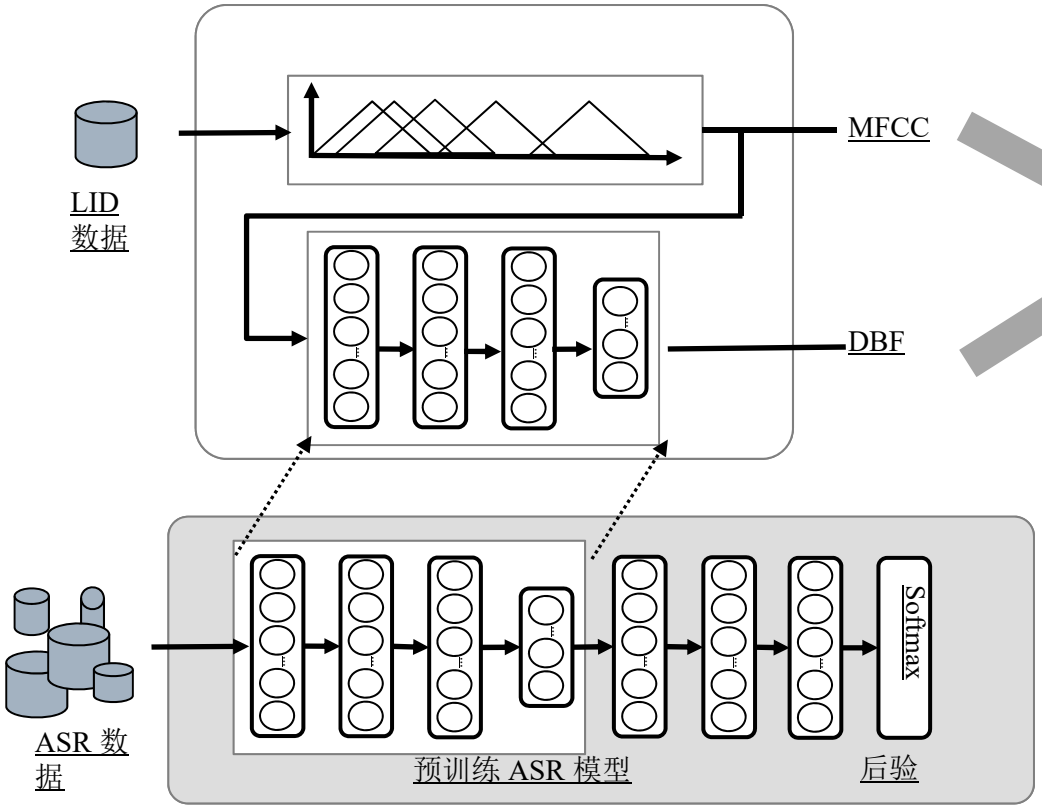


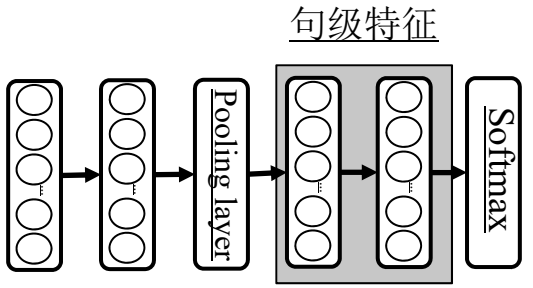
基于深度学习的 i-vector 说话人/语种识别系统

1) DNN i-vector　2) DBF i-vector　3) DNN DBF i-vector



P3 深度学习——端到端系统



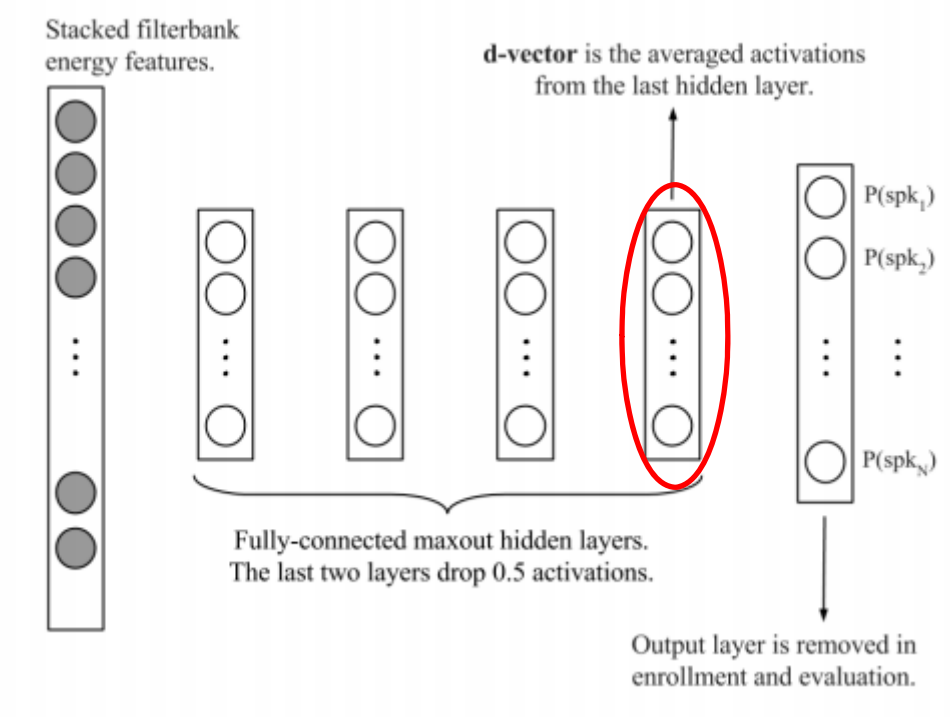


- 通过一个声学模型作为特征提取的预训练模型
- 多种声学特征送入句级特征提取器
- 语种分类

P3 深度学习——dvector

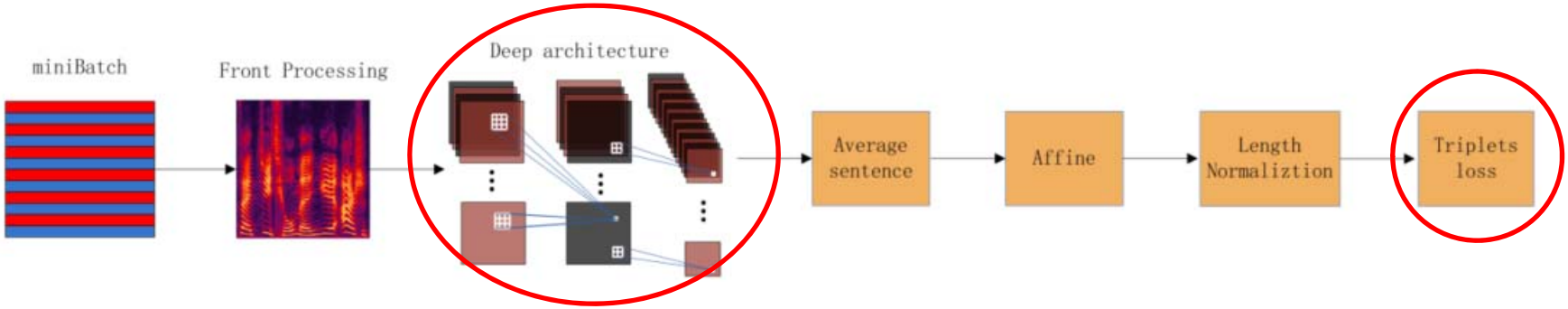
- Deep Speaker
  - Fbank特征输入
  - 学习目标为目标类别
  - 隐层输出作为特征
  - 多帧长取平均

  - 模型小，计算速度快
  - 嵌入式迁移方便
  - 随着数据量的增加性能有明显提升



P3 深度学习——ResNet

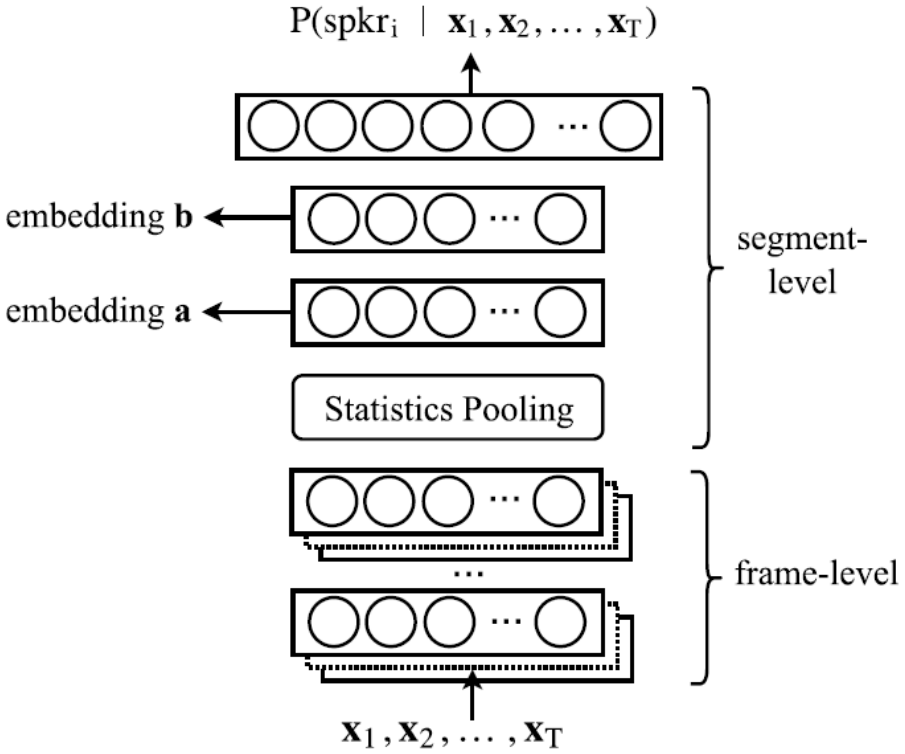
- Deep Speaker



- Fbank或者语谱图作为输入
- 多层卷积增加模型深度，提升建模能力
- 多帧平均提取整句特征
- 可扩展的鉴别性优化函数

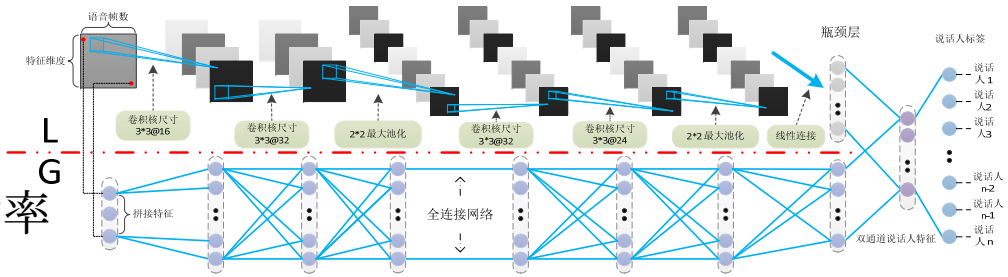
P3 深度学习——xvector

- Deep Speaker
  - 增加方差统计层
  - 数据扩增提升数据量，提高模型对数据变化的学习能力
  - Kaldi开源，模型简单，易于实现



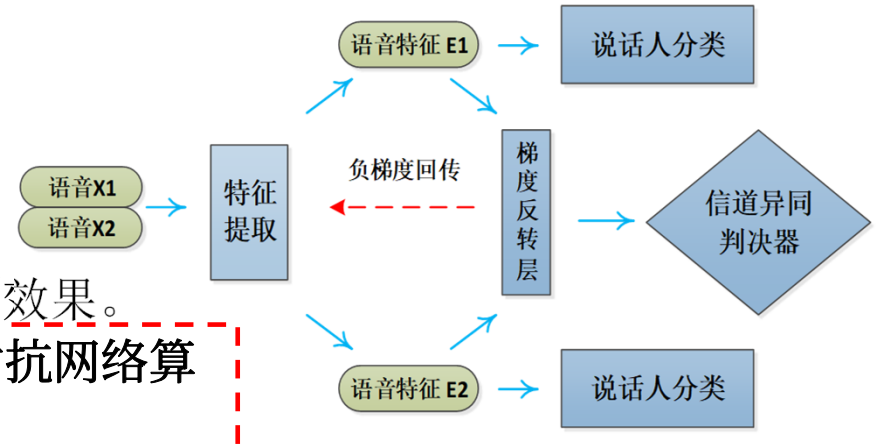
P3 深度学习——前沿成果

- 双流建模
  - 多尺度学习
  - 有效提高短时语音识别正确率



首次将类脑研究与深度学习相结合，使声纹识别语音时长最短限制减少5秒。

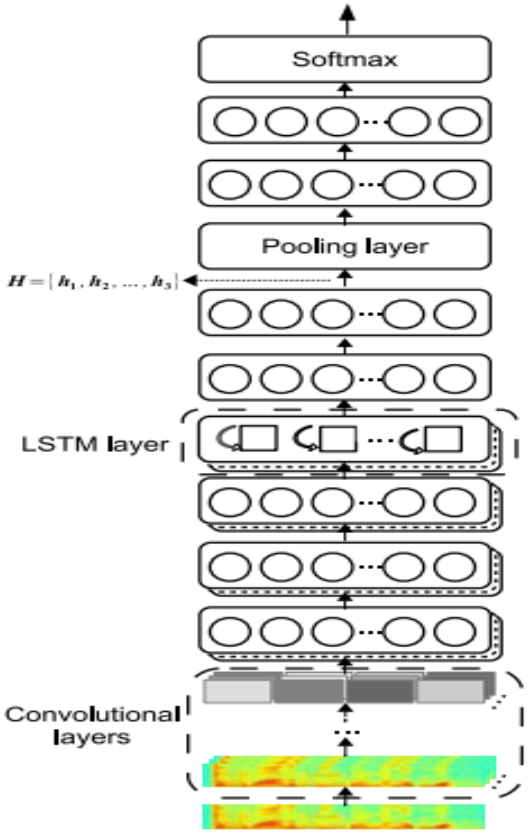
- 跨信道补偿
  - 孪生网络成对学习
  - 信道异同判决器
    - 克服训练数据不平衡问题
    - 提升在未知信道场景下的识别效果。



业界目前最好的跨信道补偿算法，相比一般对抗网络算法性能相对提升20%。

P3 深度学习——前沿成果

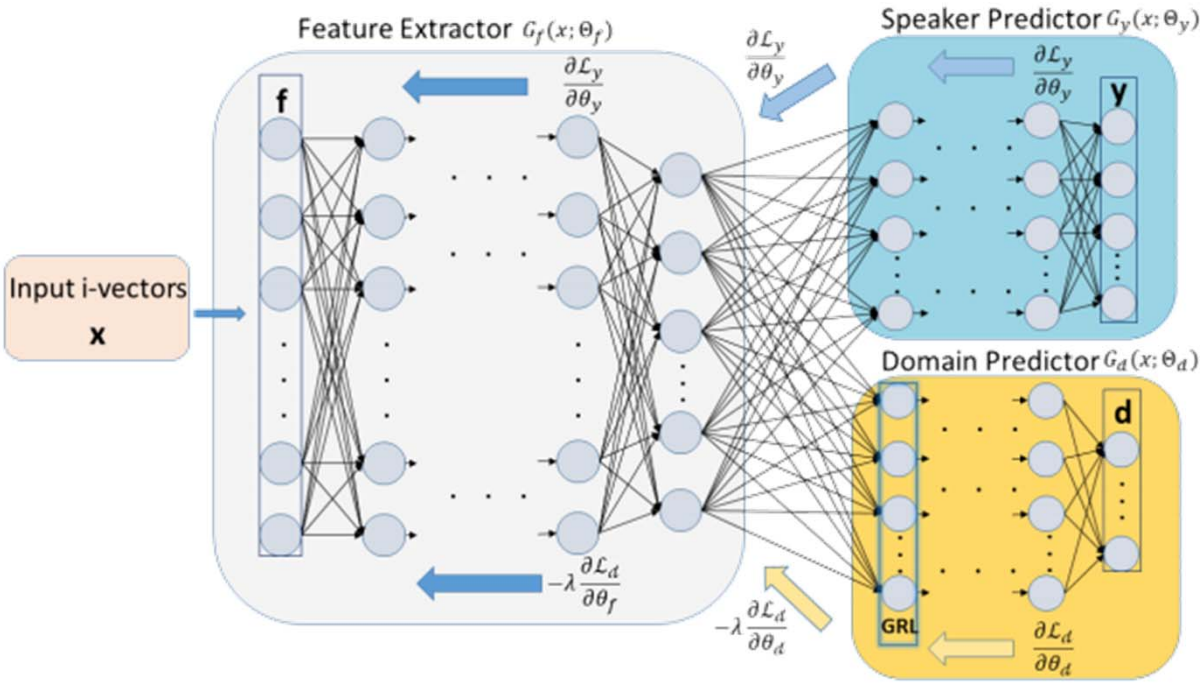
- 时序累积建模
  - 基于注意力机制的时频域信息融合
  - 时序累积、鉴别建模分布学习
  - 提高模型鲁棒性和长时信息学习能力



弥补TDNN上下文时序累积较短的问题

域自适应

- Domain adversarial training (DAT)[13]



Thank you