中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

语音交互技术
——声纹识别

中国科学院自动化研究所
模式识别国家重点实验室

陶建华
jhtao@nlpr.ia.ac.cn

本节课提纲

- 简介
  - 概念及应用
  - 评价指标
  - 发展历程
- 传统模型
  - 说话人表征
  - 分类器
- 端到端模型
- ASVspoof
  - 什么是ASVspoof
  - 欺骗与对策

本节课提纲

- 简介
  - 概念及应用
  - 评价指标
  - 发展历程
- 传统模型
  - 说话人表征
  - 分类器
- 端到端模型
- ASVspoof
  - 什么是ASVspoof
  - 欺骗与对策

概念

什么是“声纹识别”

- 声纹识别（又称说话人识别），就是从某段语音中识别出说话人身份的过程。
- 与指纹类似，每个人说话过程中蕴含的语音特征和发音习惯等是唯一的。

与“语音识别”的不同

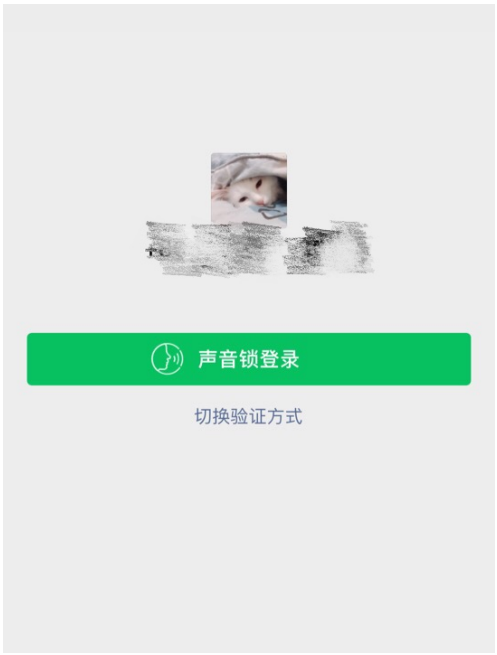
- “语音识别”是共性识别，判定所说的内容（说的什么）。
- “声纹识别”是个性识别，判断说话人身份（是谁说的）。

声纹识别的独特优势

- 语音采集装置造价低廉，只需电话、手机或者麦克风即可，无需特殊的设备。
- 与指纹、人脸相比，声纹更适合远程身份认证。
- 声纹口令可以动态变化。

应用

智能音箱



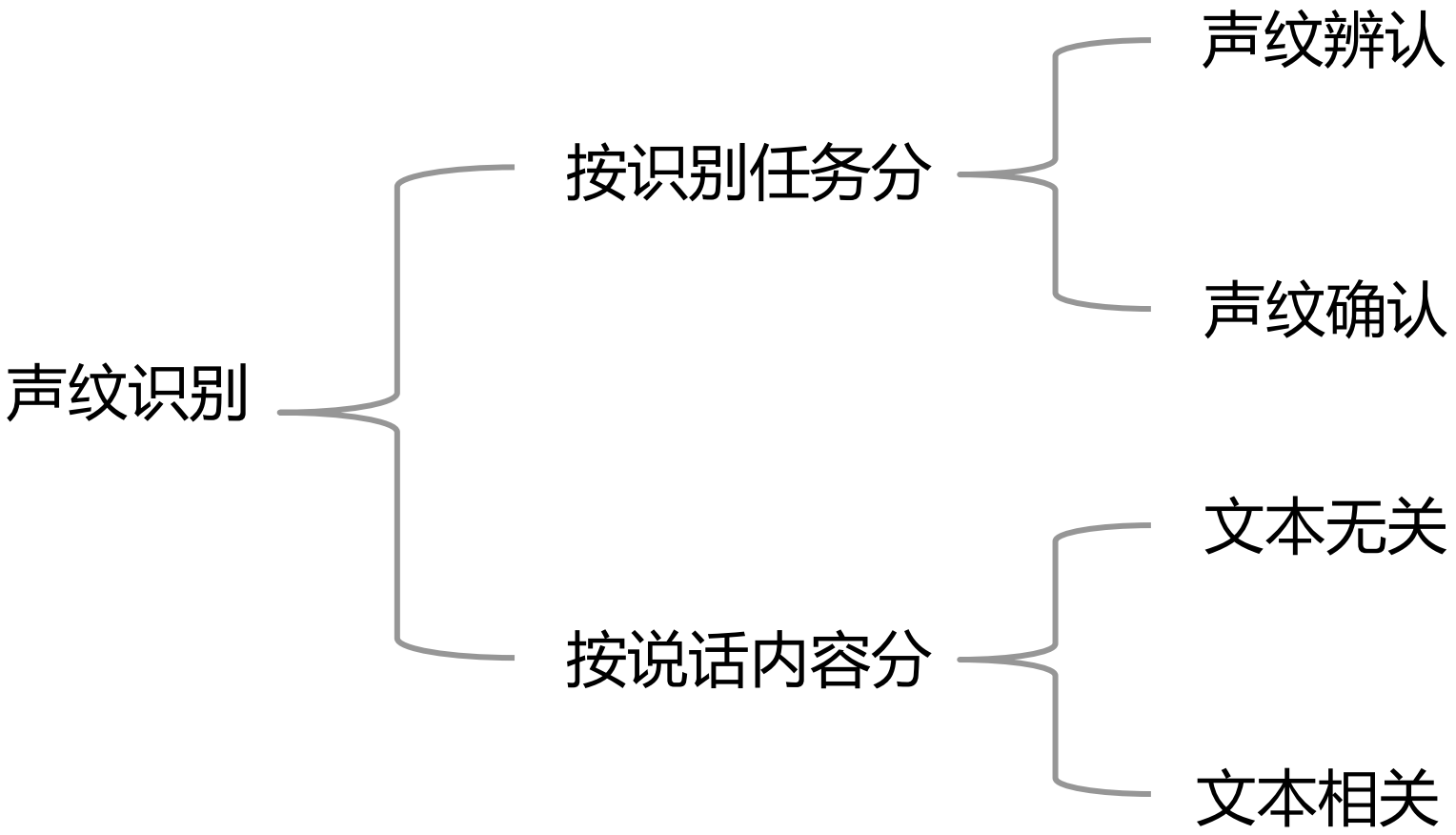
微信声纹登陆

声纹信用卡

类别	采集便利性	准确率	采集成本	采集是否接触	远程识别	造假难度	用户接受度
声纹识别	高	高	低	非接触式	是	高	高
人脸识别	高	高	高	非接触式	是	中	高
指掌纹系统	高	高	高	接触式	否	低	高
虹膜识别	低	高	中	半接触式	否	中	中
DNA识别	中	极高	高	接触式	否	极高	低

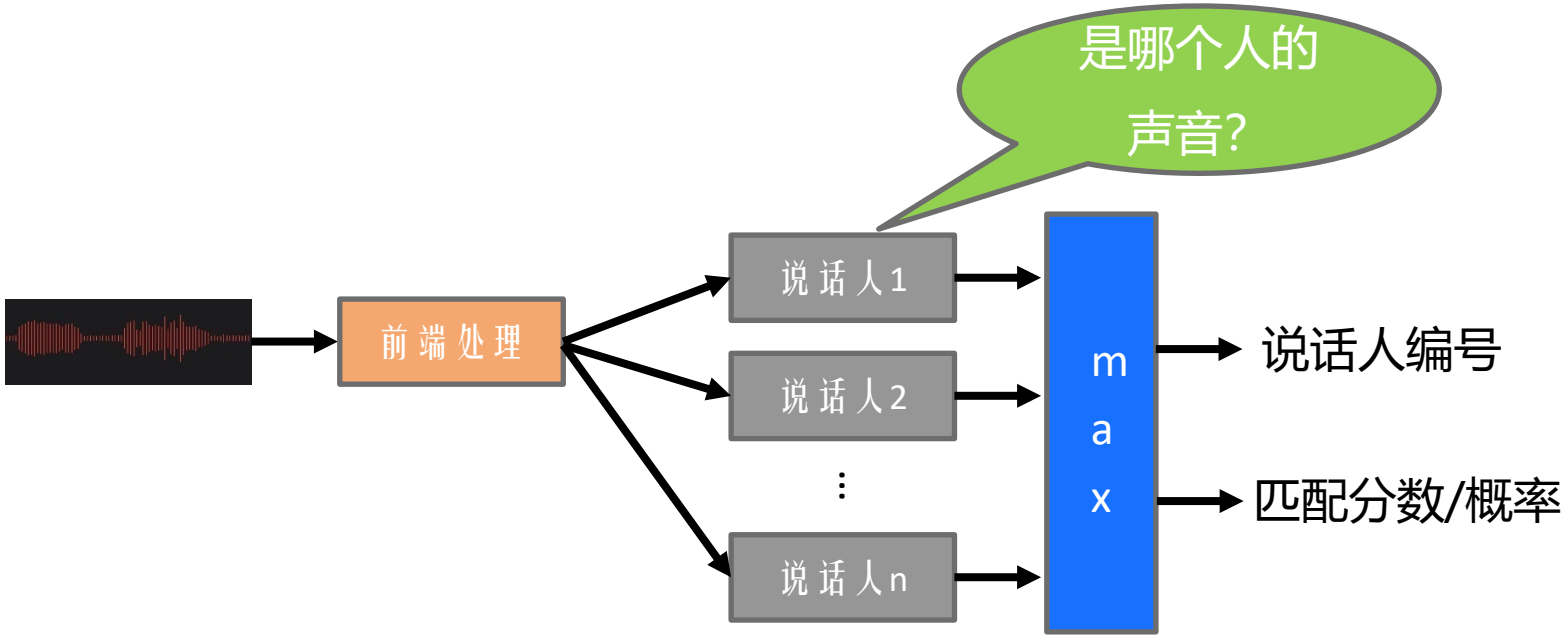
公共安全领域中的识别技术

分类方式



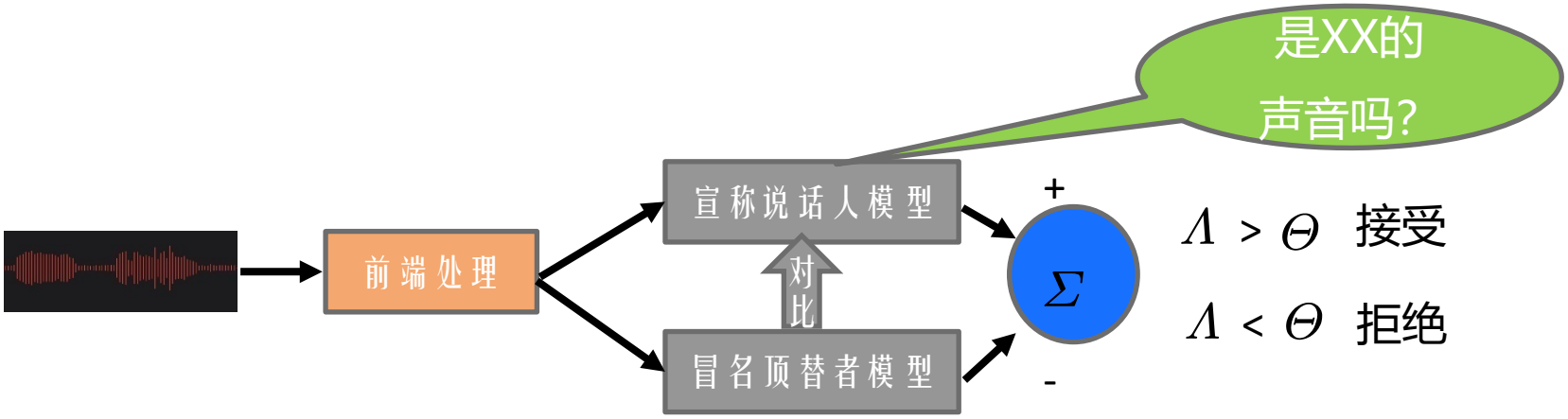
分类方式一

- 按识别任务分：
  - 声纹辨认（identification）



分类方式一

- 声纹确认（verification）



分类方式二

- 按说话内容分类
  - 文本无关（Text-independent）
    不限定说什么文本
    又可分为语种无关、语种相关
  - 文本相关（ Text-dependent ）
    要求说特定的文本
    必定是语种相关

评价指标

- 对于声纹辨认系统，其性能的评价标准主要是正确识别率。
- 对于声纹确认系统，其重要的两个指标是错误拒绝率（FRR）与错误接受率（FAR），前者是拒绝真实的说话人，又称“拒真率”，后者是接受冒认者而造成的错误，又称为“认假率”，两者均与阈值的设定相关。

评价指标

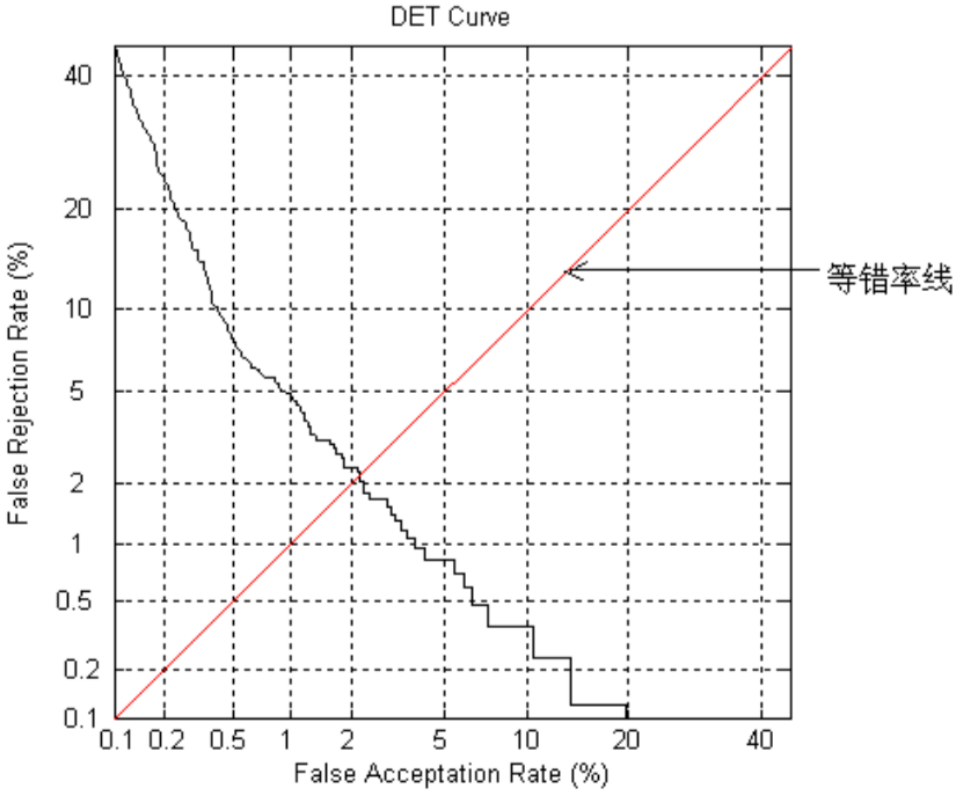
$$FRR=\frac{a}{a+b},\ FAR=\frac{c}{c+d}$$

其中a是指将目标说话人识别为非目标说话人；b是指将非目标说话人识别为非目标说话人；c是指将非目标说话人识别为目标说话人；d是指将将目标说话人识别为目标说话人。

等错误率（EER）：FRR=FAR

$$EER=\frac{a}{a+b}=\frac{c}{c+d}$$

DET曲线图

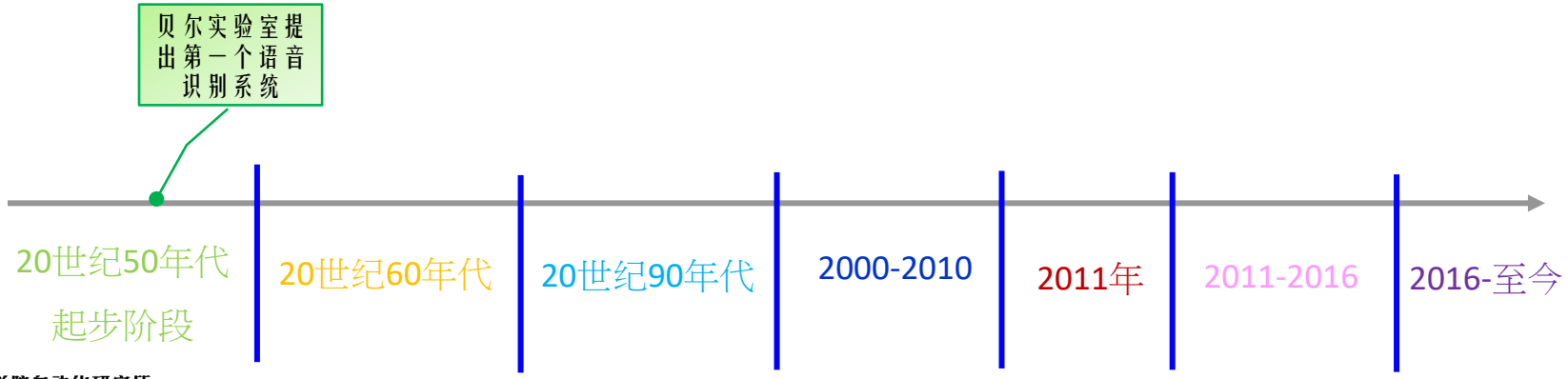


DET(Detection Error Tradeoff)曲线是对二元分类系统误码率的曲线图，绘制出错误拒绝率FRR（False Reject Rate）与错误接受率（False Accept Rate）之间随着判断阈值的变化而变化的曲线图。

发展历程

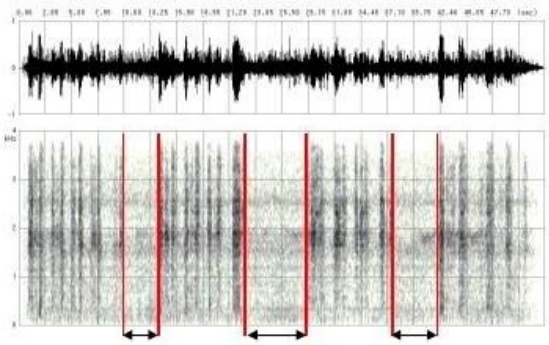
语音识别的研究工作可以追溯到20世纪50年代AT&T贝尔实验室的Audrey系统，它是第一个可以识别十个英文数字的语音识别系统。

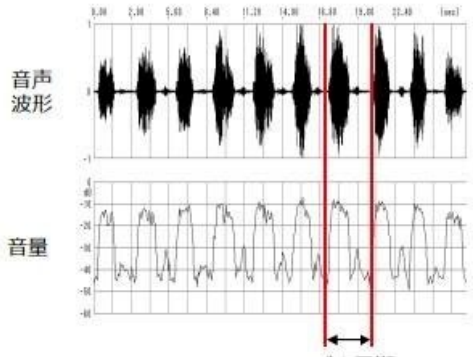
贝尔实验室

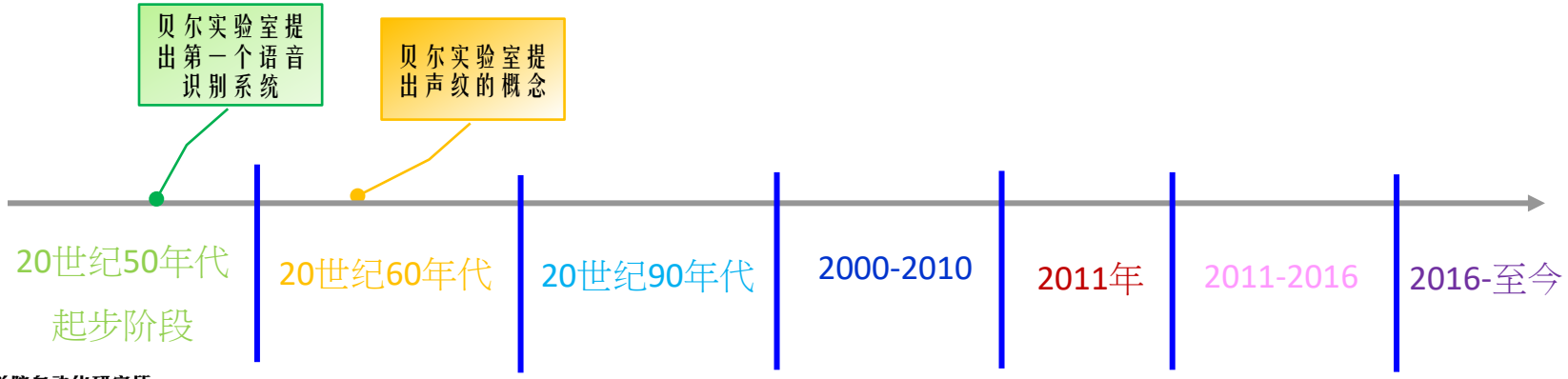


发展历程

贝尔实验室在语音语谱的基础上，提出了声纹 这个概念。

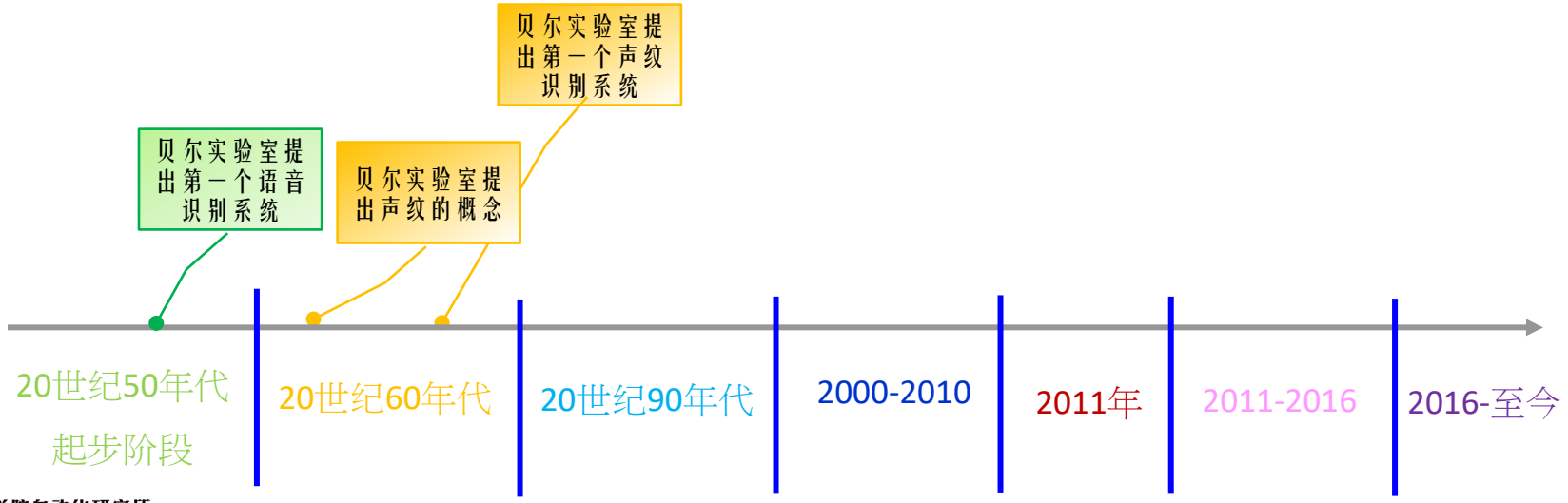




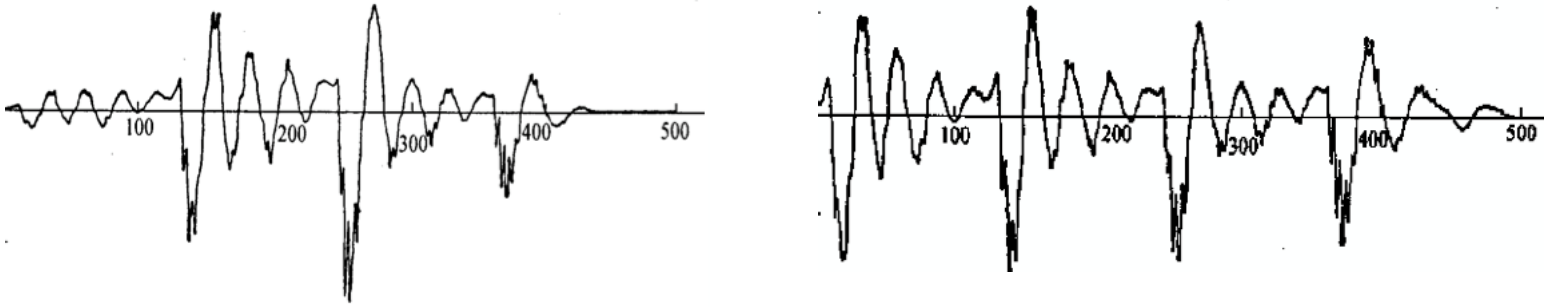


发展历程

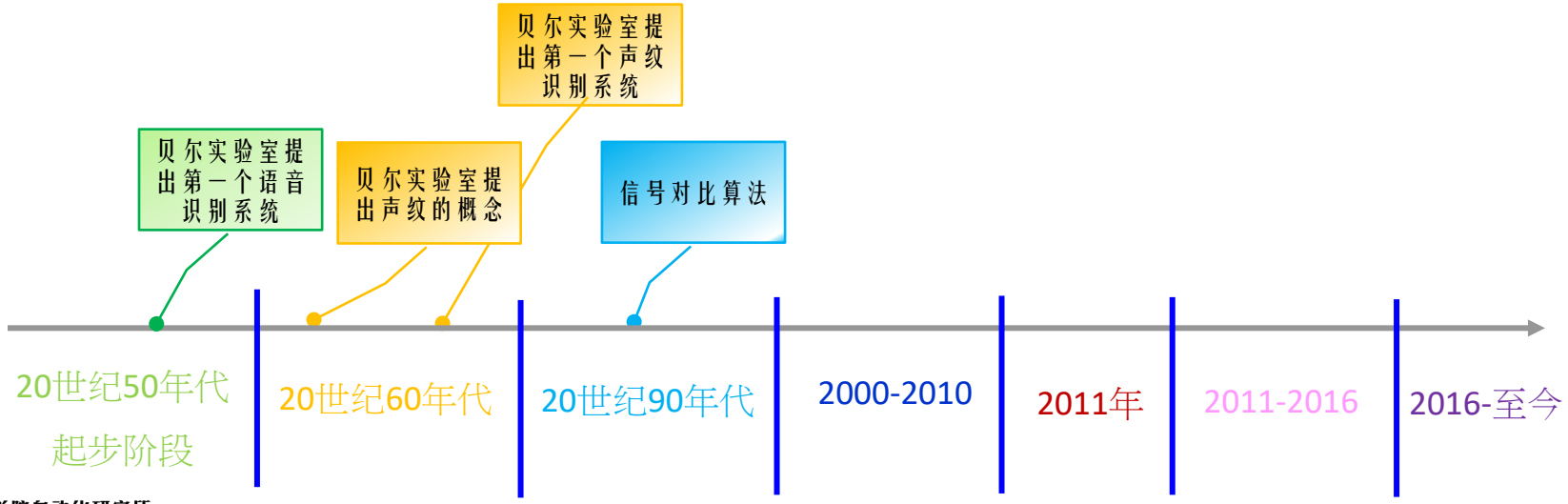
在声纹这个概念基础上，贝尔实验室提出了基于模式匹配和概率统计方差分析的说话人识别方法，此后声纹识别技术得到快速发展，从单模板模型发展到多模板模型，从模板模型发展到矢量量化模型、高斯混合模型、隐马尔可夫模型，再到人工神经网络……



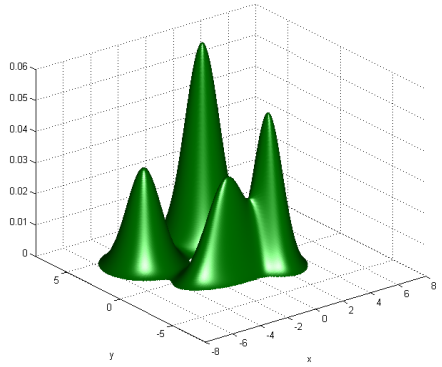
发展历程



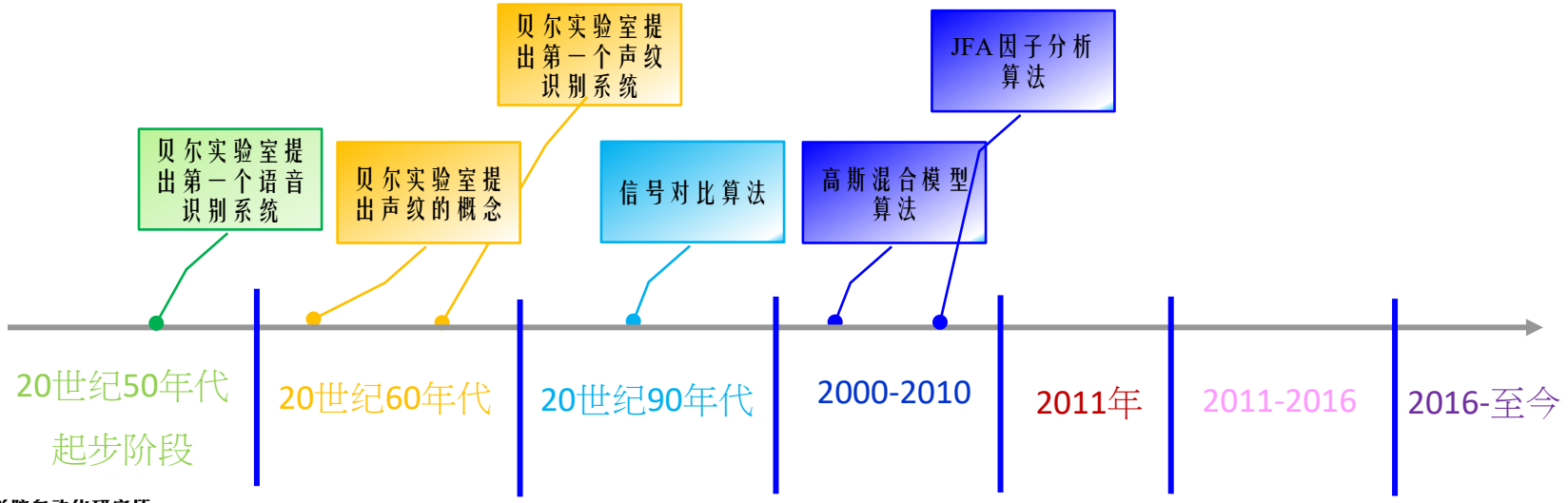
信号对比算法：说同样的内容，对比双方的信号相似度才能验证是否为同一说话人。



发展历程

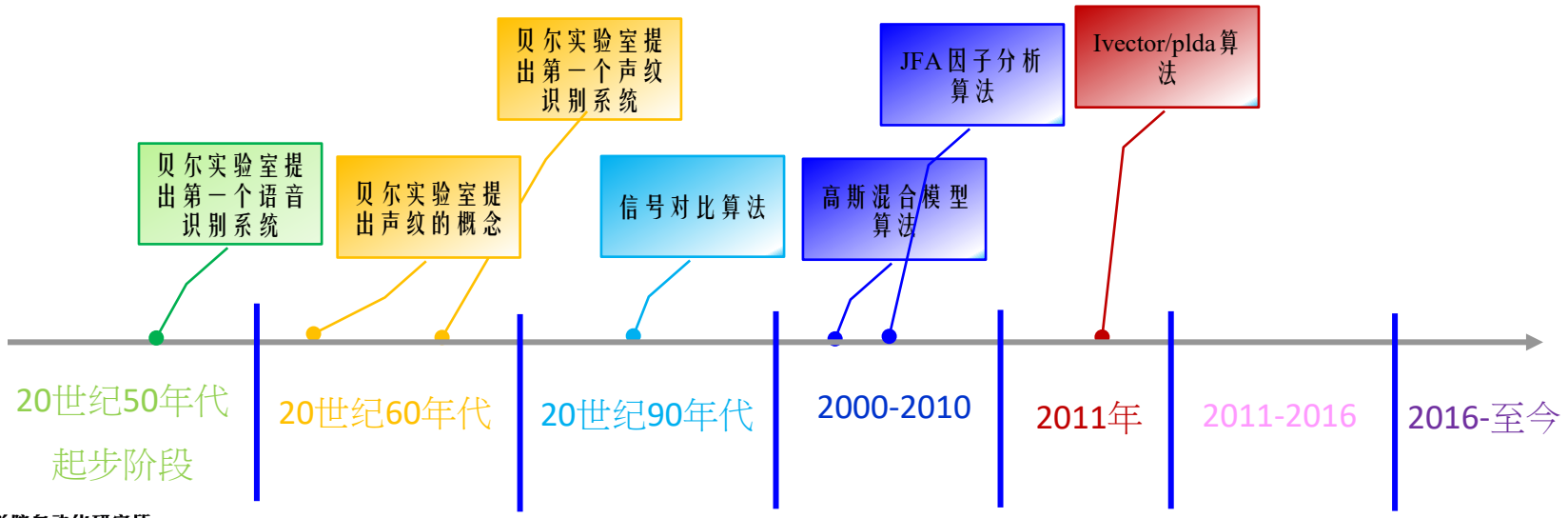


- 说话人需要建立自己的模型时，就可以通过最大后验概率 (MAP)自适应UBM（Universal Background Model）来得到个性特征，即修正后的参数，从而得到自己的GMM（高斯混合模型）。

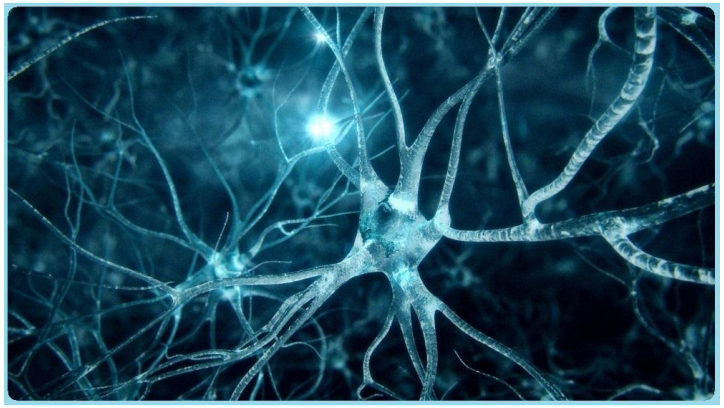


发展历程

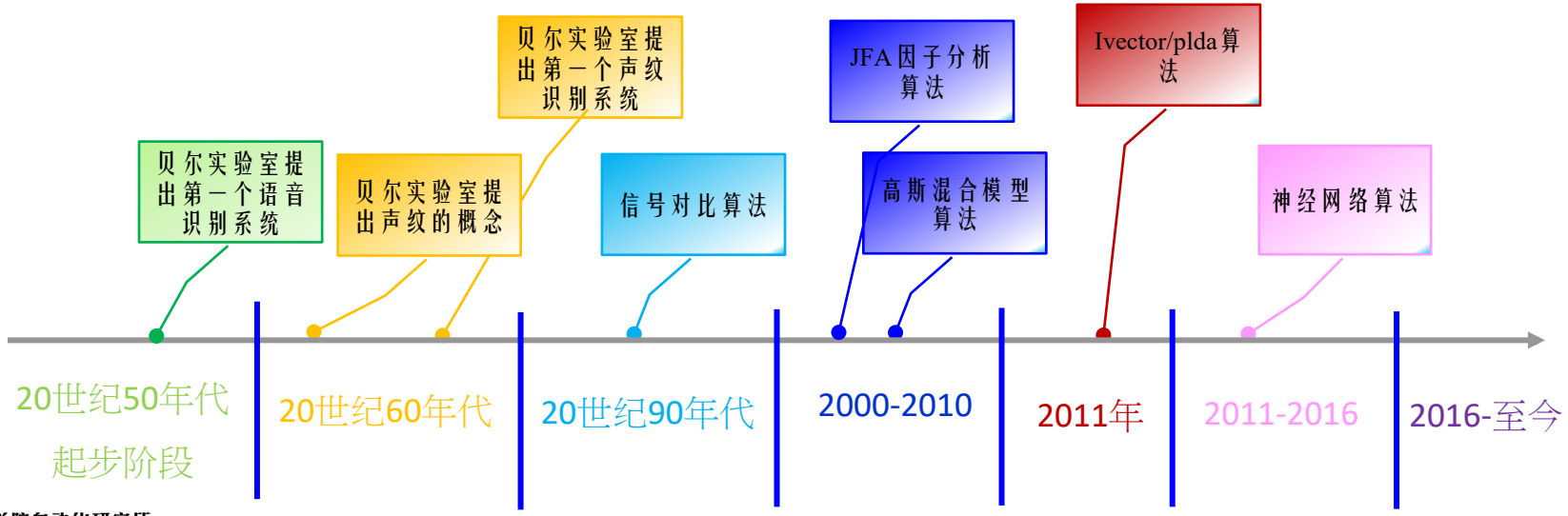
高斯混合模型的缺点是注册语音的时间过长，通常注册一条语音至少需要3-5分钟的时间。2011年，i-vector/plda系统的提出是具有颠覆意义的。Ivector的声纹特征embedding将语音从高维向量转化为低维向量。Plda算法具有信道补偿的作用。



发展历程

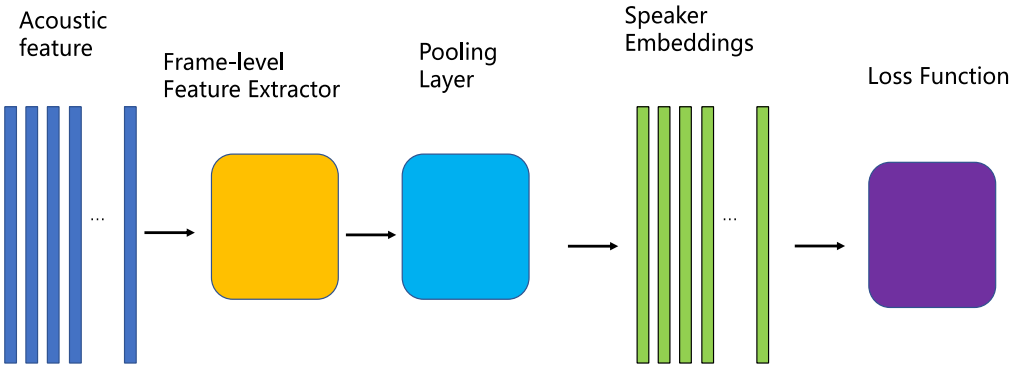


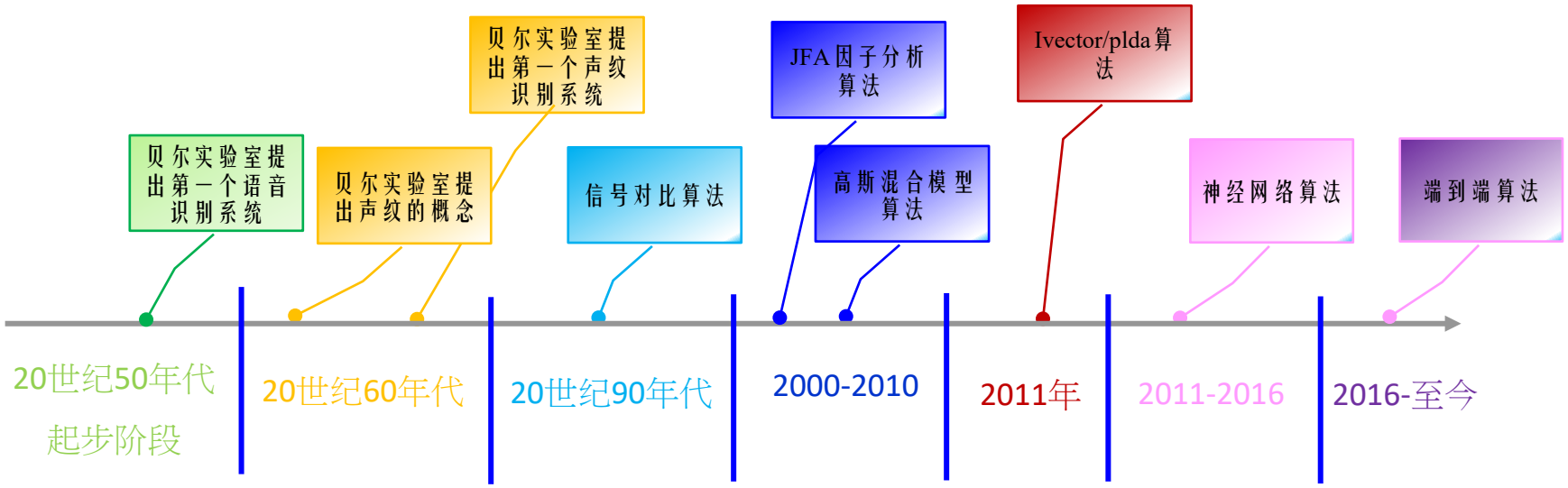
随着神经网络的快速发展，声纹识别中也将神经网络代替传统的高斯混合模型。神经网络的引入大大提高了识别正确率并且具有较好的鲁棒性。



发展历程

- 端到端声纹识别

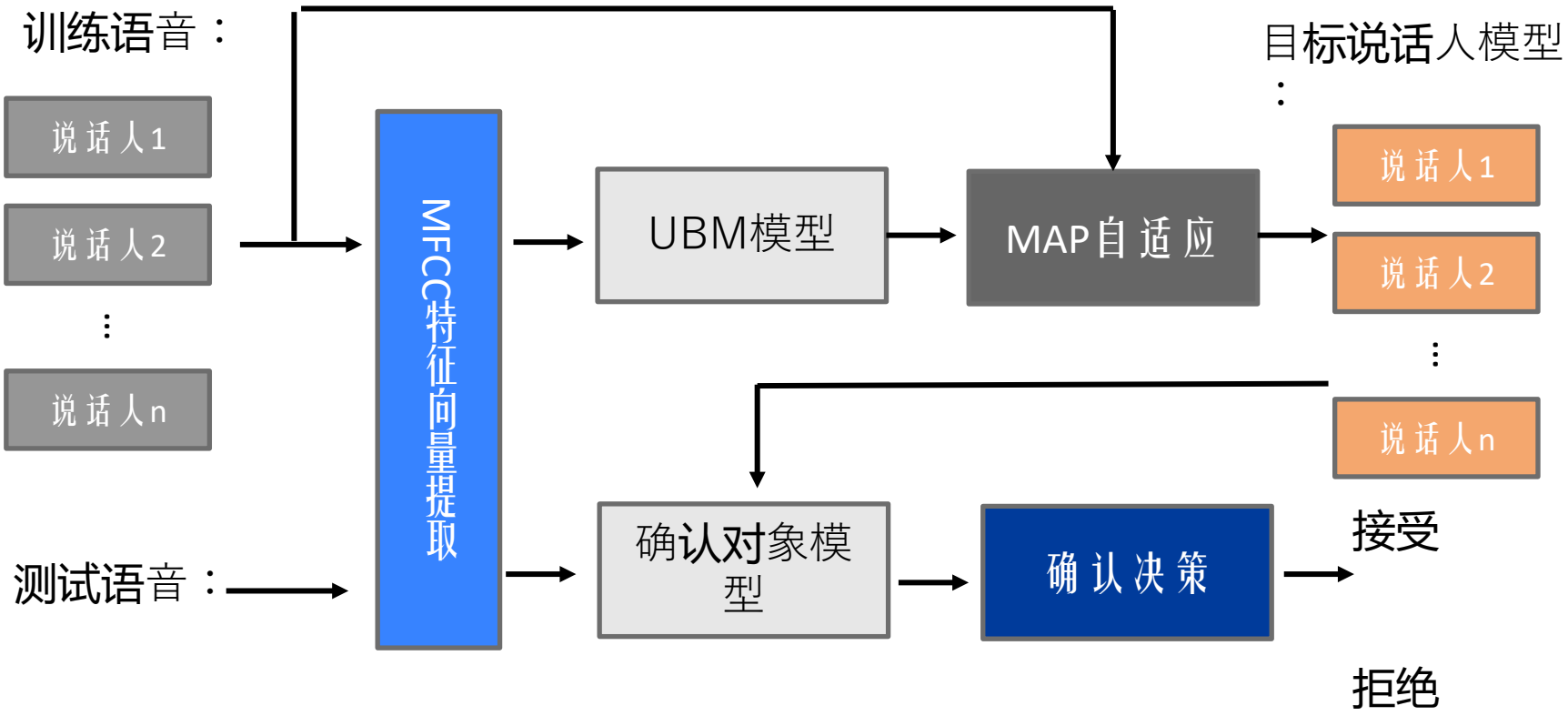




本节课提纲

- 简介
  - 概念及应用
  - 评价指标
  - 发展历程
- 传统模型
  - 说话人表征
  - 分类器
- 端到端模型
- ASVspoof
  - 什么是ASVspoof
  - 欺骗与对策

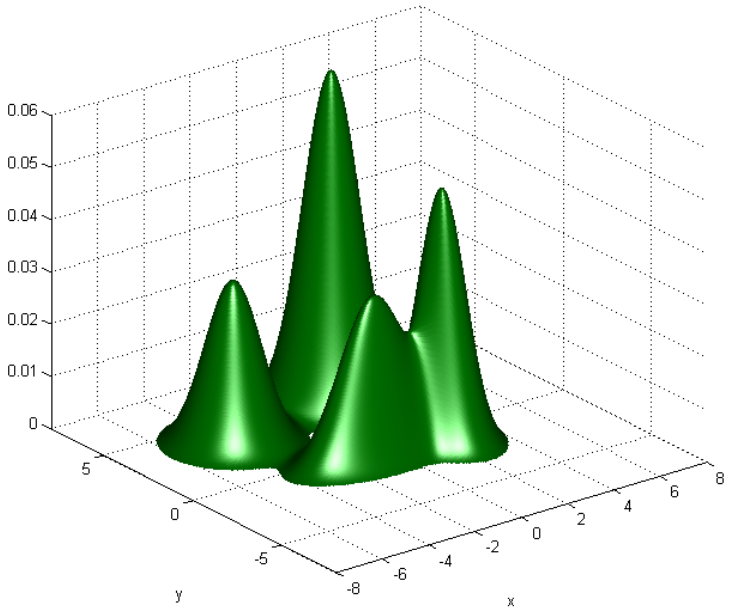
传统框架（GMM-UBM）



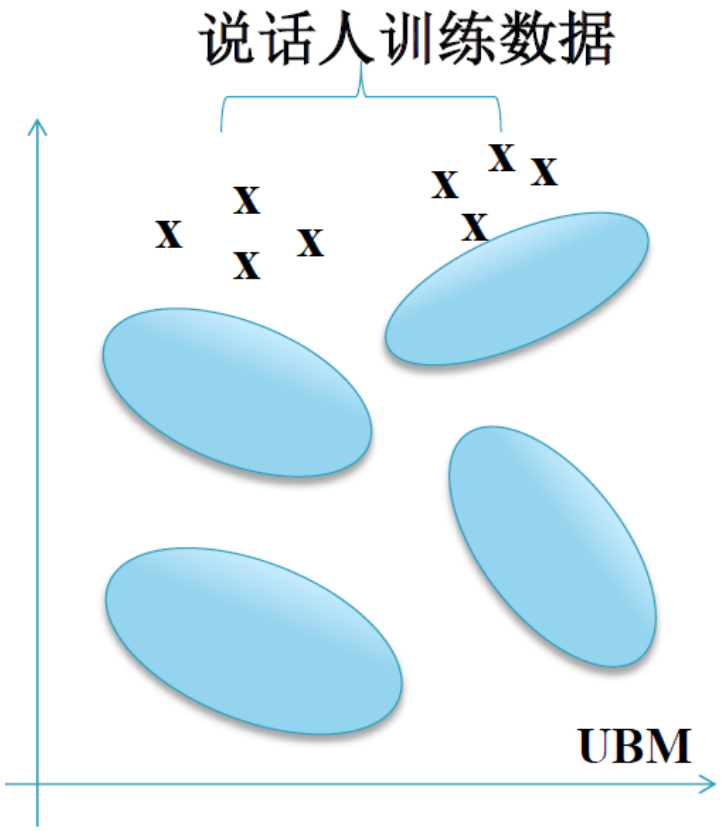
- 说话人需要建立自己的模型时，就可以通过最大后验概率 (MAP)自适应UBM来得到个性特征，即修正后的参数，从而得到自己的GMM。

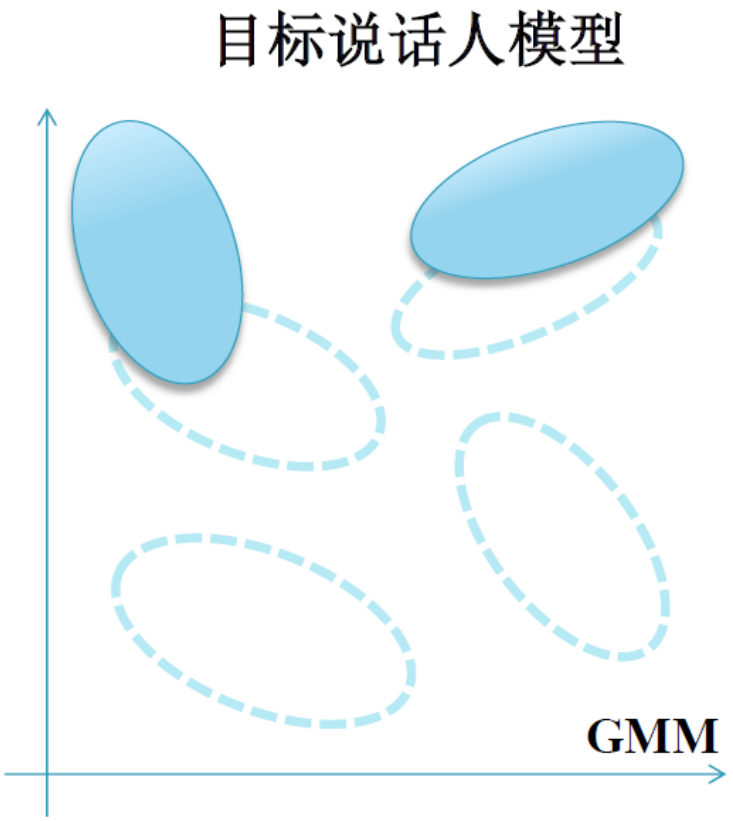
UBM-通用背景模型

- UBM （Universal Background Model）也是一个GMM，只是这个GMM需要用大量不同的说话人的语音数据经过训练来表示说话人无关的特征分布，这种特征是大多数说话人的共性特征。



UBM ⇨ GMM





GMM的优缺点

- GMM优点：

- 概率统计模型，可较好地刻画目标说话人不同情况下的特点，具有一定的鲁棒性。
- 同信道效果很好，已可实用。

- GMM缺点：

- 因为声纹识别的训练数据量是少量的，有限的数据不一定能充分代表说话人地真实分布；只考虑某一类地模型参数和本类训练数据之间的相似程度，而没有考虑与其他类别的区分性。
- 跨信道性能急剧下降！

JFA联合因子分析算法

- 联合因子分析(Joint Factor Analysis, JFA)的方法认为说话人信息由三个空间组成：本征说话人信息空间、本征信道空间和残差空间。

- 即假设每个说话人可以用一个与说话人和信道相关的GMM均值超矢量M来表示，并且可以分解为说话人超矢量s和信道超矢量c和的形式:

$$M = s + c$$

其中，s和c各自独立且服从高斯分布。s表示的是说话人之间的差异，c表示的是信道之间的差异。

JFA联合因子分析算法

- 说话人超矢量s与信道超矢量c分别可以由隐含变量表示:

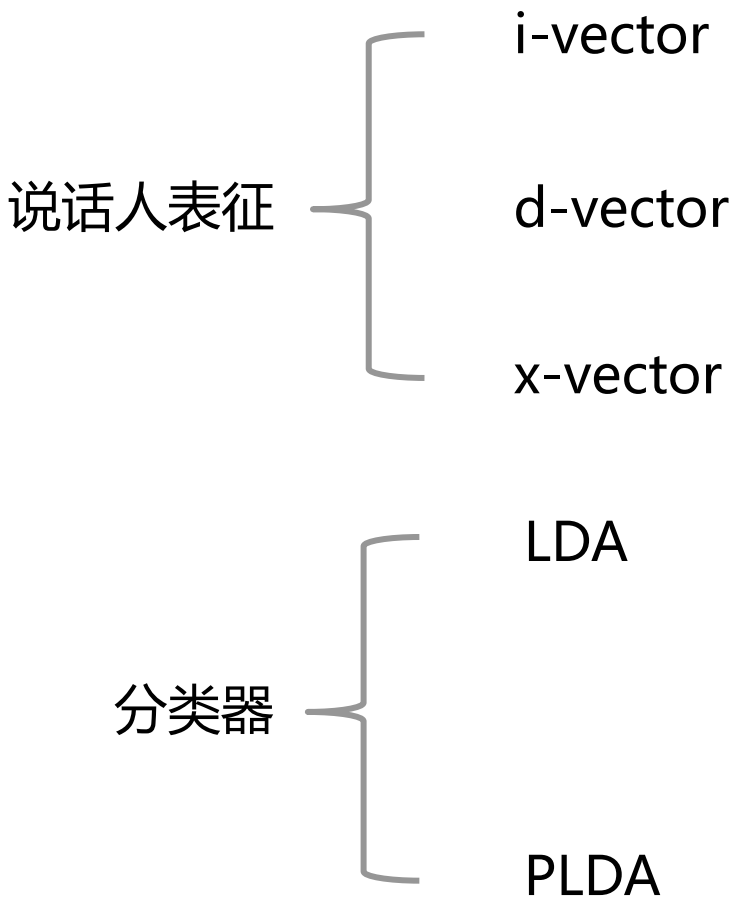
$$S = m + Vy + Dz$$

$$c = Ux$$

其中，m表示说话人及信道无关的均值超向量，V和y表示说话人空间及其对应的说话人相关的因子， D和z表示残差空间及其对应的残差因子，U和x表示信道空间和信道相关的因子。

- 由此得到说话人的GMM均值超矢量作为说话人模型，最后通过对数似然比的方法对说话人进行判决。

传统模型



说话人表征

- i-vector
- d-vector
- x-vector

i-vector

- i-vector[1]是基于单一空间的跨信道算法，该空间既包含了说话人空间的信息也包含了信道空间信息。

  对于给定的语音，高斯超向量表示如下:

$$M = m + Tw$$

- 其中， m是话者无关且信道无关的超向量，通常由 UBM的均值向量拼接而成;T是一个低秩的矩阵; 而w则是服从标准正态分布的随机向量，简称i-vector。

[1]N. Dehak et al. “Front-end factor analysis for speaker verification”. In: IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2011), pp. 788–798.

零阶、一阶统计量

- 已知一个UBM，有C个高斯，每个高斯表示如下 $\lambda_c = W_c$，μ_c，σ_c^2，$c = 1, 2, \cdots, C$

- 其中W_c是权重，μ_c是均值，σ_c^2是方差。
  给定一段T帧语音$O = o_1, o_2, \cdots, o_T$,其零阶和一阶 Baum-Welch统计量计算如下:

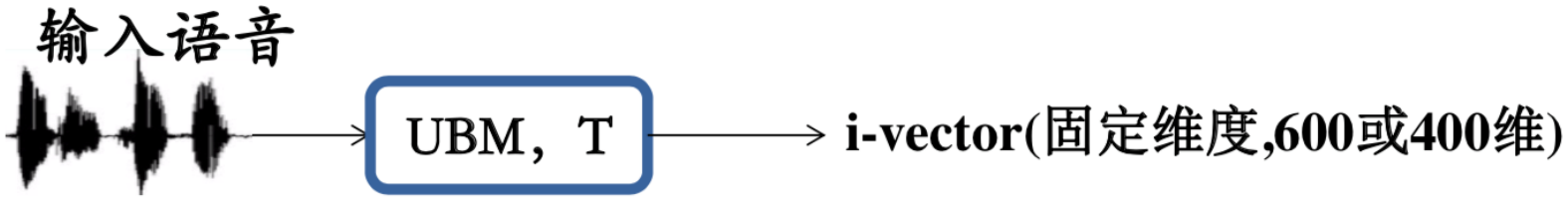
$$N_c = \sum_{t=1}^{T} P(c|o_t, \lambda_c)$$

$$F_c = \frac{1}{N_c} \sum_{t=1}^{T} P(c|o_t, \lambda_c)\,(o_t - \mu_c)$$

提取i-vector

- 在得到总变化矩阵T后，还需计算出总变化因子 w(i-vector), 它的计算过程同样需要UBM作为先验知识。

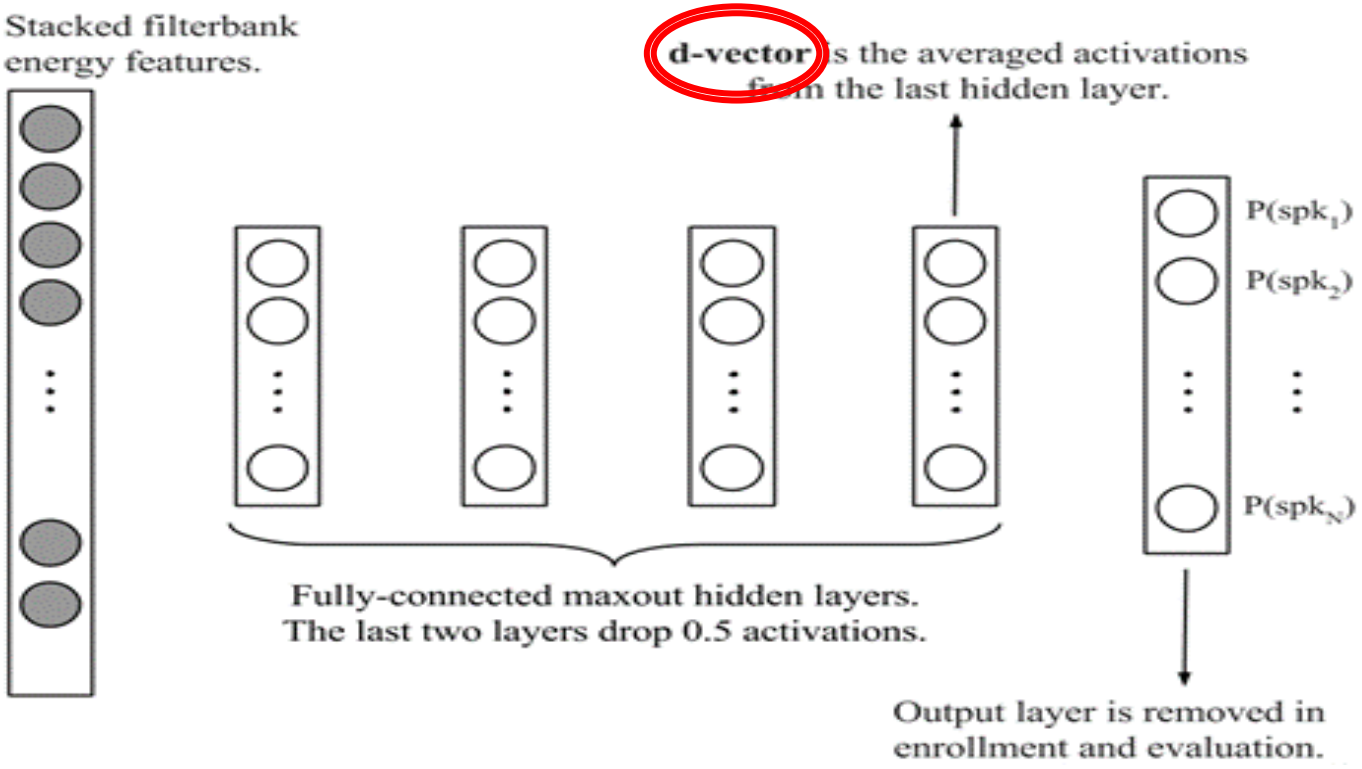
$$w = (I + T^T\Sigma^{-1}N(u)T)^{-1}T^T\Sigma^{-1}F(u)$$



d-vector

- UBM的作用
  - 包含C个高斯，用来做分类器;
  - 给出每一帧对每个高斯的后验概率;
  - UBM分类器缺乏语义信息(音素区分);
- 类似语音识别GMM-HMM=>DNN-HMM，说话人识别的UBM也可替换为DNN。

d-vector



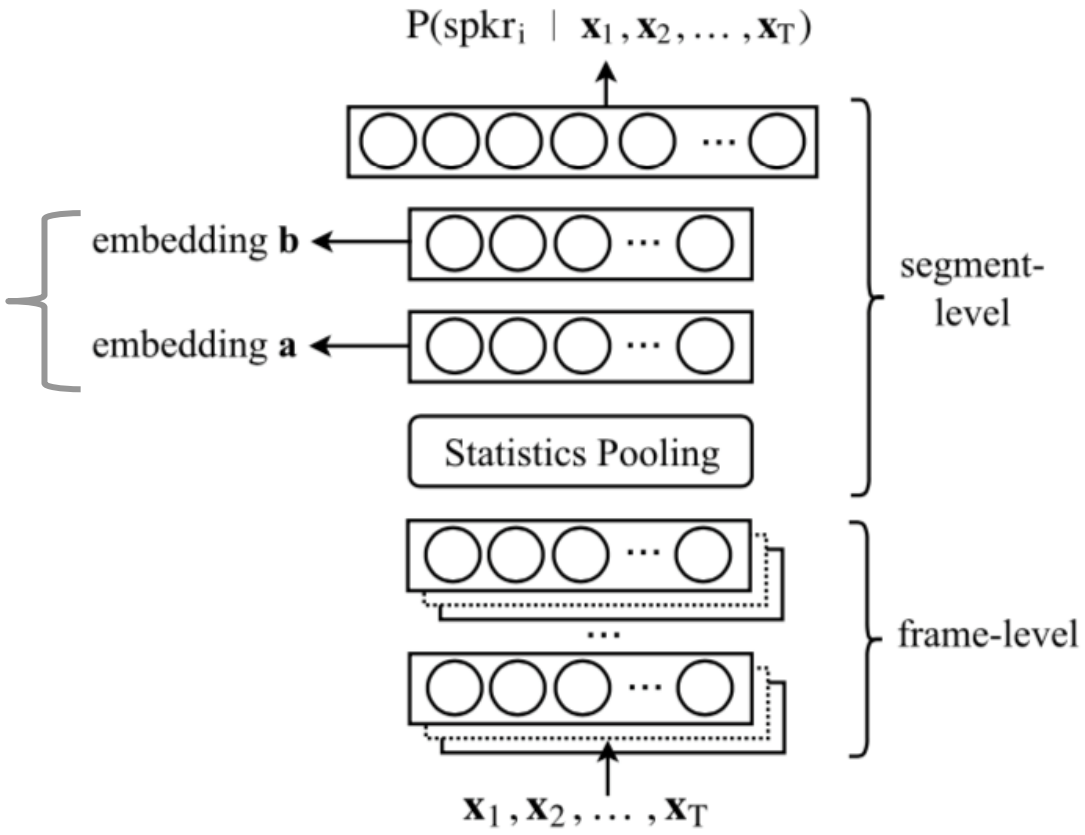
Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, Javier Gonzalez-Dominguez. "DEEP NEURAL NETWORKS FOR SMALL FOOTPRINT TEXT-DEPENDENT SPEAKER VERIFICATION", In: Proc. of ICASSP. IEEE. 2014

x-vector

- 采用大数据，模拟噪声和远场环境
- 输出标签与说话人ID对应
- 基于深度模型，提取更有效的说话人特征——Deep Embedding（x-vector）

x-vector

x-vector



Statistics Pooling:

不同帧数的输入特征在这里聚合平均，形成段级别的特征，输出后再用LDA进行降维。

David Snyder, Daniel Garcia-Romero, Daniel Povey, Sanjeev Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification", *Interspeech2017*

x-vector的优点

- X-vector比i-vector更能利用大量的数据。
- X-vector在短时声纹上比i-vector效果好很多。
- X-vector比i-vector更具鲁棒性。

Cosine距离

- 在计算得到说话人表征之后，直接利用两向量间的余弦(Cosine)距离，它也被证明在提高训练与测试效率的同时仍不降低识别性能。

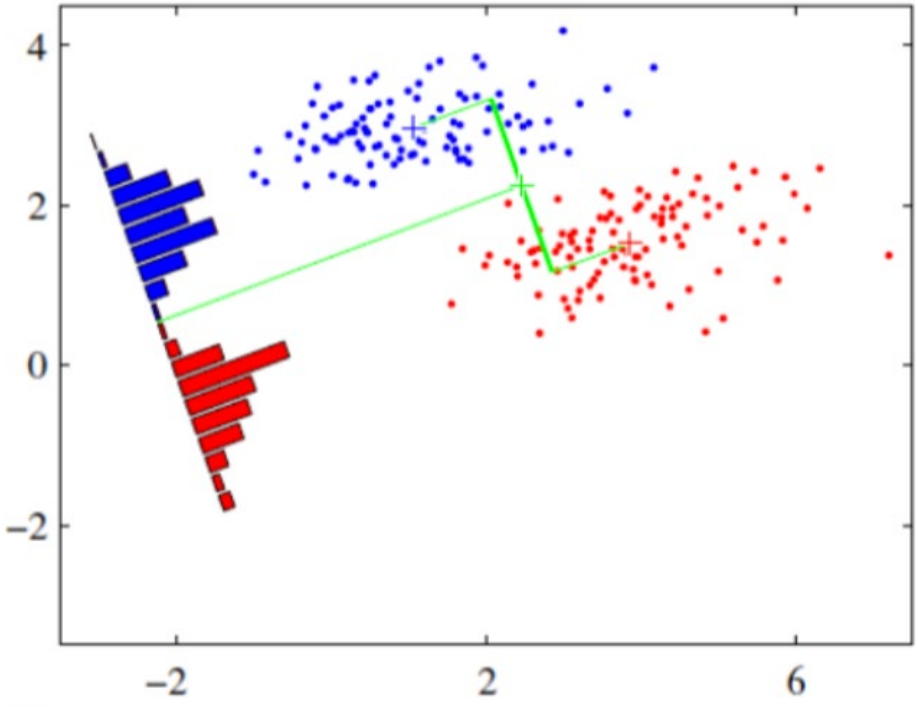
$$k(w_1, w_2) = \frac{< w_1, w_2 >}{\|w_1\| \bullet \|w_2\|}$$

分类器

- LDA
- PLDA

线性判别分析（LDA）

- 模式分类中，在数据处理中的降维步骤经常会用到线性判别分析(Linear Discriminant Analysis, LDA)方法。
- LDA可以在不破坏良好的类别区分度的前提下，将数据集投影到更低维空间。



LDA原理

- 定义类内散布矩阵S_W

$$S_W = \sum_{c=1}^{C} \frac{1}{N_c} \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^T$$

- 定义类间散布矩阵S_B

$$S_B = \sum_{c=1}^{C} (\mu_c - \mu)(\mu_c - \mu)^T$$

- 设投影矩阵为w， 求解

$$w^* = arg \max_w \left\{ \frac{w^T S_B w}{w^T S_W w} \right\}$$

LDA降维

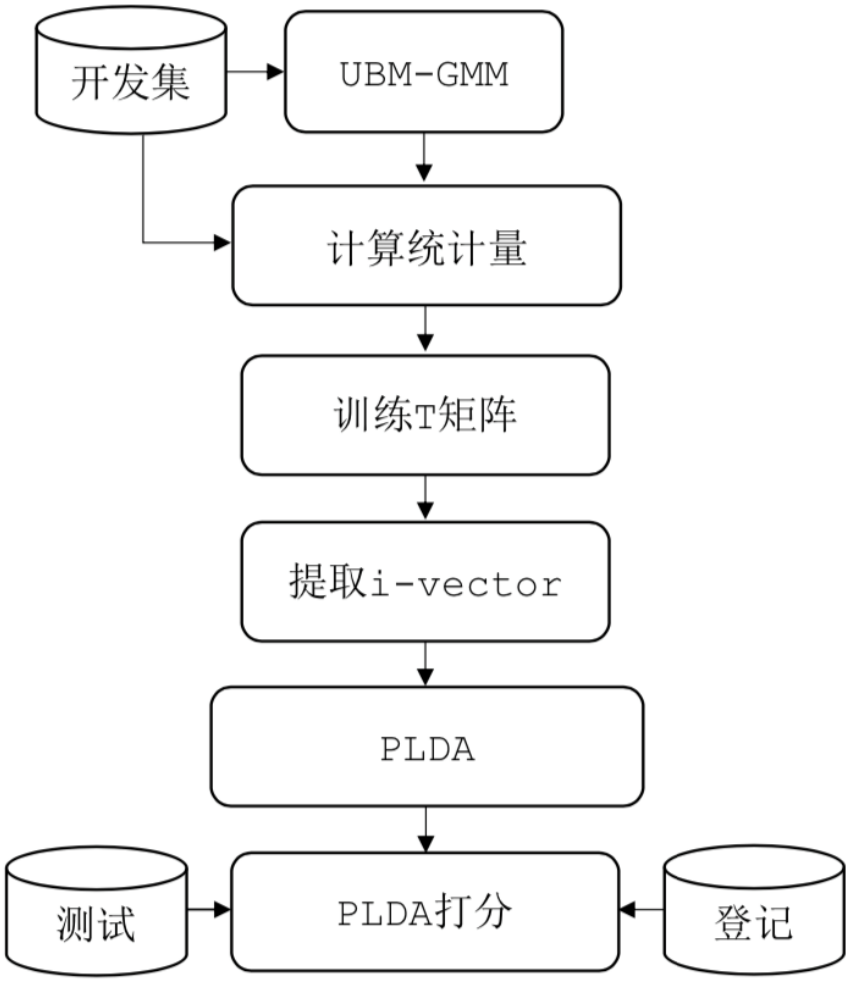
- 求解矩阵$S_W^{-1}S_B$的广义本征值，得到线性判别器。线性判别器最多有C个，C是实验对象的类别总数。
- LDA的形变程度可以由本征值和本征向量共同描述
  - 本征值表示形变的幅度;
  - 本征向量表示形变的方向。
- 在LDA降维中，根据本征值的大小，由高到低排序对本征向量进行排序，然后选择使用前k个本征向量，即抛弃末尾的本征向量。
- 设输入样本x有d维，LDA矩阵为$d \times k$矩阵w，则降维后的k维输出样本为:

$$y=xw$$

PLDA

- PLDA(Probabilistic Linear Discriminant Analysis)是一种信道补偿算法，号称概率形式的LDA算法。
- PLDA同样通常是基于i-vector特征的，因为i-vector特征即包含说话人信息又包含信道信息，而我们只关心说话人信息，所以才需要信道补偿。
- PLDA算法的信道补偿能力比LDA更好，已经成为目前最 好的信道补偿算法。

i-vector/PLDA系统



传统模型的缺点

- 训练过程复杂：需要先提取说话人表征再送到分类器打分。
- 鲁棒性不足，跨信道表现较差
- 训练时间还是太长

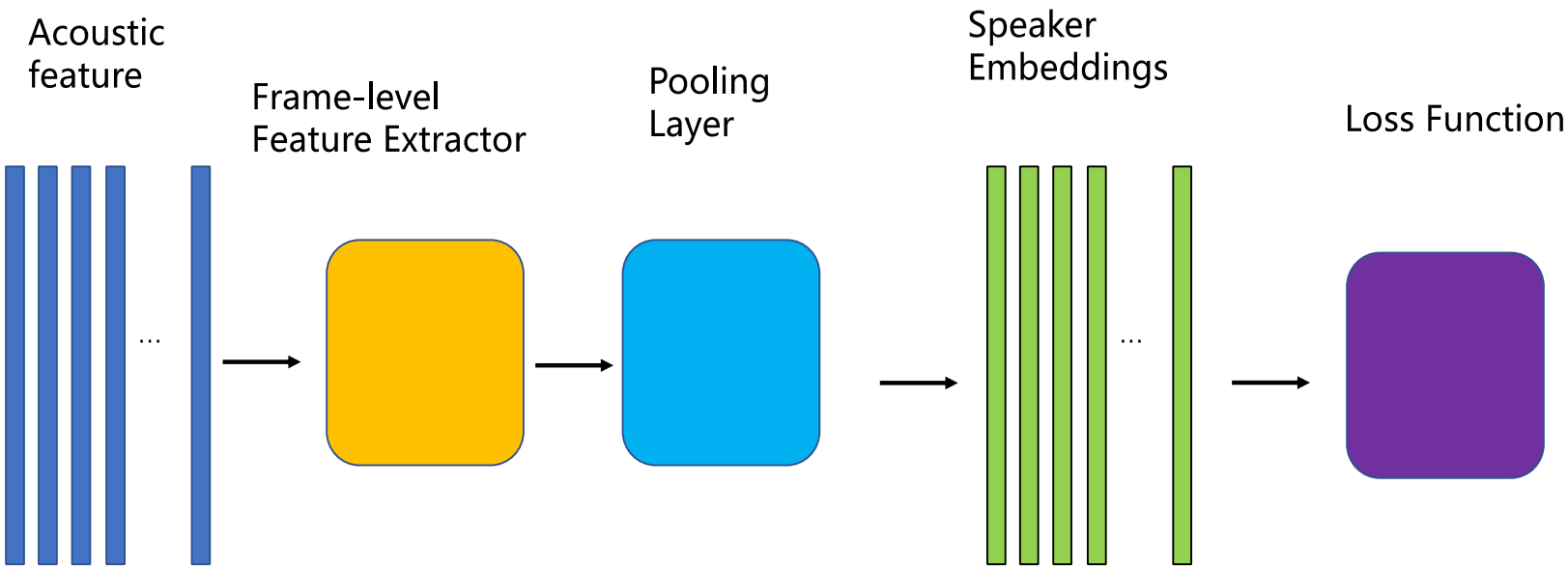
本节课提纲

- 简介
  - 概念及应用
  - 评价指标
  - 发展历程
- 传统模型
  - 说话人表征
  - 分类器
- 端到端模型
- ASVspoof
  - 什么是ASVspoof
  - 欺骗与对策

端到端模型的优点

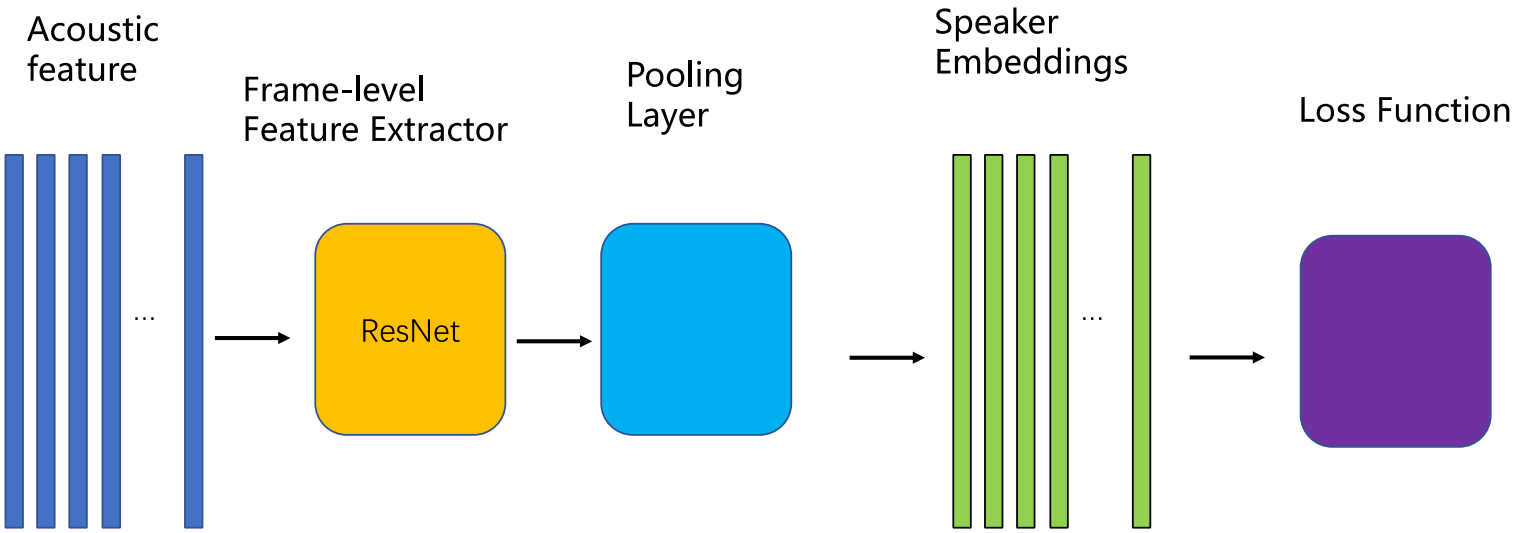
- 有一个整体的优化目标（loss function）
- 结构简单明确
- 训练速度更快
- 系统更具鲁棒性

端到端声纹识别框架



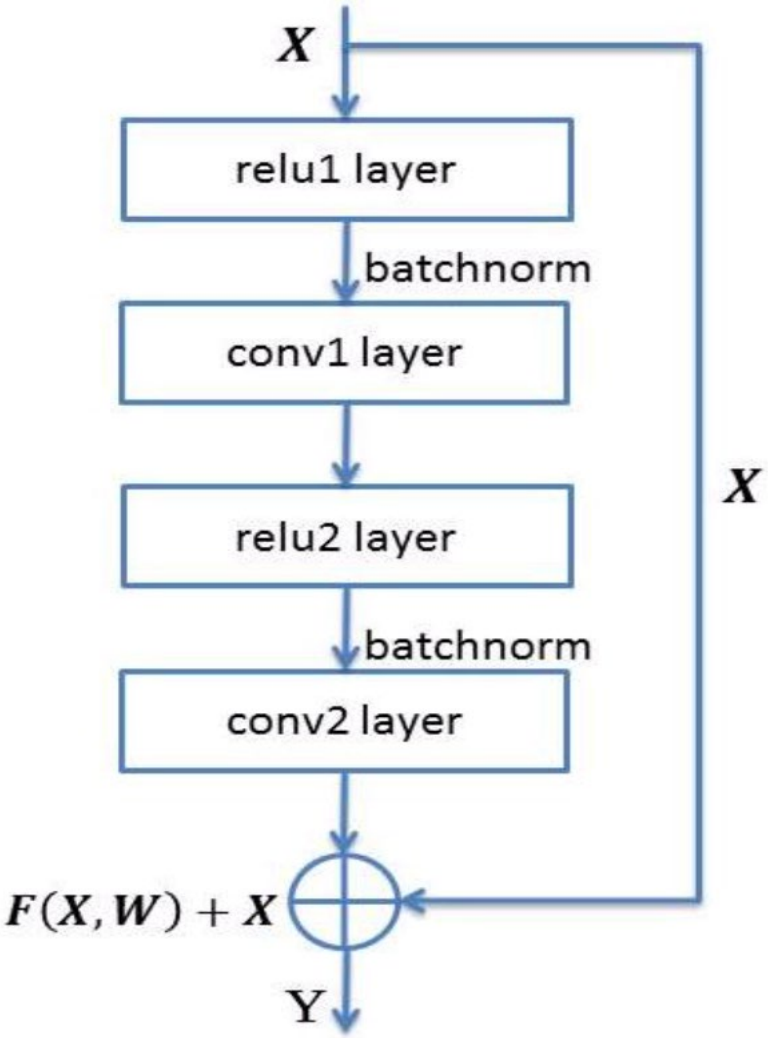
- Acoustic feature：输入帧级别的语音
- Frame-level Feature Extractor: 提取帧级别的语音特征，常用的有CNN、RNN、ResNet等
- Pooling Layer：池化层的目的是将帧级特征整合为段级特征，常用的有最大值池化、平均值池化和注意力池化

基于ResNet的端到端声纹识别



整个系统流程如下：首先对语音数据进行预处理、声学特征提取（如如40维滤波器组特征，Fbank）。其次基于ResNet 的方法构建深度神经网络，提取说话人的声纹特征。ResNet作为帧级特征提取器，不仅能够提取语音声学特征中表征声纹的深层局部信息特征，还可以在建模上减小频域变化。接着将ResNet的最后一层输入平均池化层，其目的是将帧级层面转化到段级层面。最后，输出向量通过softmax损失函数映射到(0，1)区间，计算声纹类别之间的概率。

残差网络ResNet


X
relu1 layer
batchnorm
conv1 layer
relu2 layer
batchnorm
conv2 layer
X
F(X, W) + X
Y

基于注意力机制的端到端声纹模型

- 注意力机制的引入：

- 传统平均池化层将其权重进行平均分配，并获得最终编码表示。该方法的问题主要在于，假设序列的所有元素必须在获得话语水平表示上做出同等的贡献。自注意力机制是一种方法，它通过可训练的层能够为序列的每个表示分配权重。因此，在给定这些权重的情况下，通过这些表示的相应加权平均值获得段级特征表示。

自注意力机制

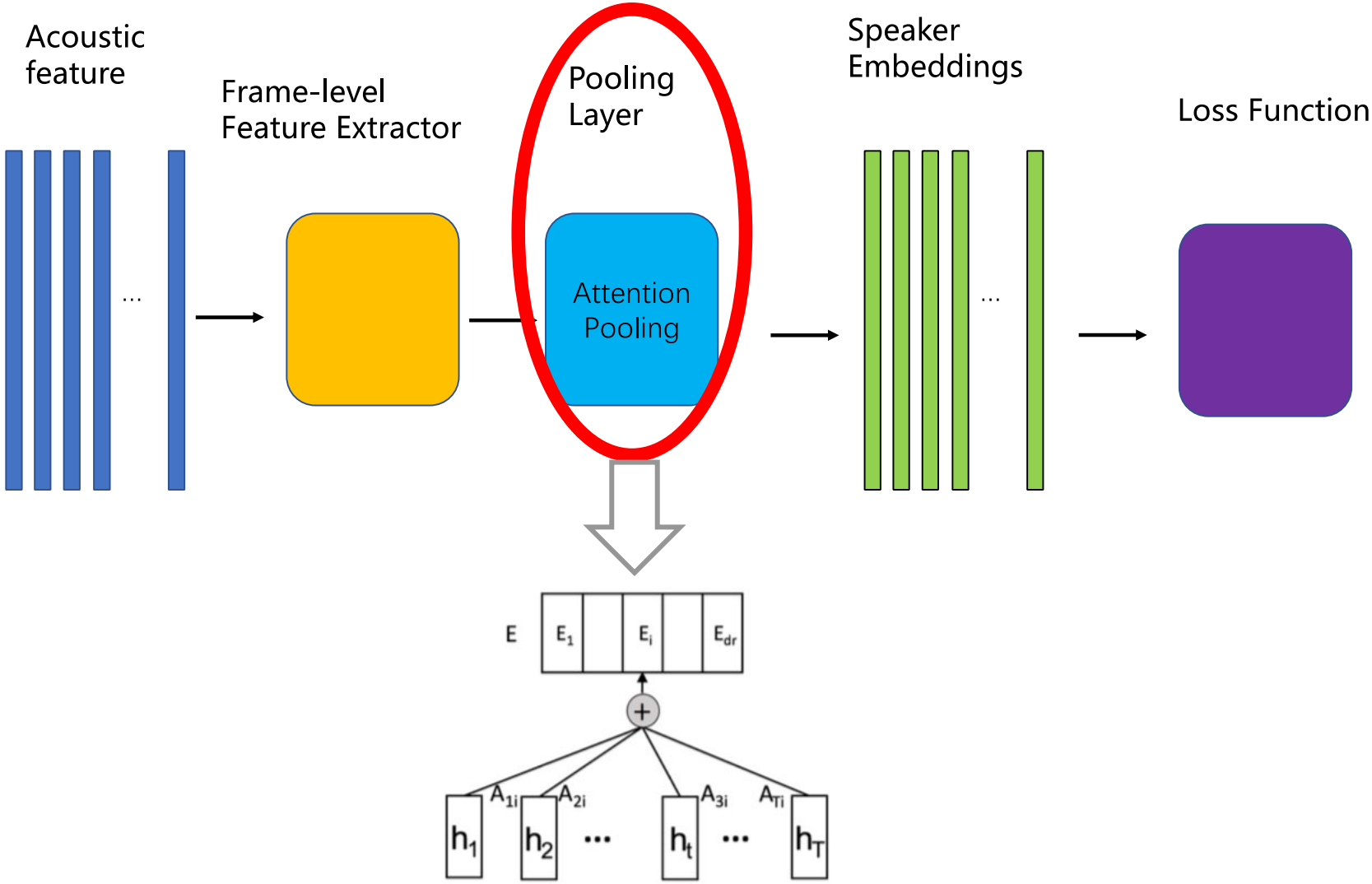
- 考虑一个序列的状态为h=$\{h_1, h_2, \ldots, h_n\}$的序列，u为可训练的系数，可以通过softmax层为序列的每个元素定义相关标量权重：

$$w_t = \frac{\exp(h_t^T u)}{\sum_{l=1}^{N} \exp(h_t^T u)}$$

- 给定所有序列元素的权重集，我们就可以得到合并的表示形式，作为序列的加权平均值：

$$c = \sum_{t=1}^{N} h_t^T \ w_t$$

基于注意力机制的端到端声纹识别的框架图



本节课提纲

- 简介
  - 概念及应用
  - 评价指标
  - 发展历程
- 传统模型
  - 说话人表征
  - 分类器
- 端到端模型
- ASVspoof
  - 什么是ASVspoof
  - 欺骗与对策

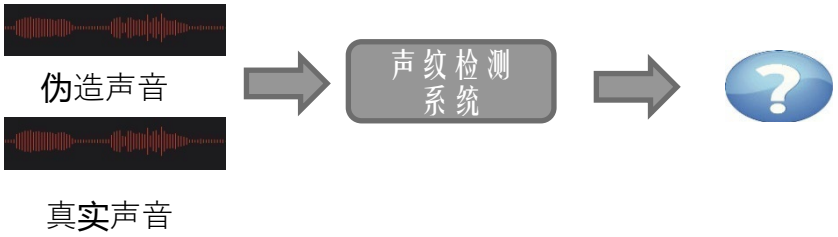
ASVspoof

什么是ASVspoof?

ASVspoof（Auto Speaker Verification spoof）是随着声纹识别技术的发展，攻击者会提出各种欺骗策略以冒充真正的用户。这就称为自动说话人识别欺骗攻击技术。

背景意义

伪造声音攻击声纹检测系统



安全性

- 生活方面：窃取用户声纹信息
- 政治方面：模仿政府官员的声纹

伪造数据库

- VCTK base corpus
- 训练集：20人（8男，12女）
- 开发集：10人（4男，6女）
- 测试集：48人（21男，27女）
- 来自VCTK数据库的17种不同TTS和VC系统生成的数据，其中6种被指定为已知攻击。

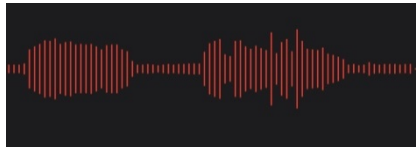
伪造声音的四种形式

- 伪造声音攻击声纹验证系统的四种形式

声音模仿

录音重放

我的密码是：



语音合成

语音转换

欺骗与对策

欺骗：声音模仿

- 语音模仿是最明显的欺骗手段之一，指的是使用人类改变的声音进行攻击，在这里，攻击者试图模仿目标说话人的声音音色和韵律，而不使用计算机辅助技术。

对策：

- 伪装检测器

欺骗与对策

欺骗：录音重放

- 录音回放是一种欺骗形式，使用收集的预先录制的语音样本，该样本来自真正的目标发言人。由于高质量和低成本的记录设备例如智能手机特别常见，录音回放欺骗攻击可以说是最容易获得的 。

对策：

- 信道噪声检测法

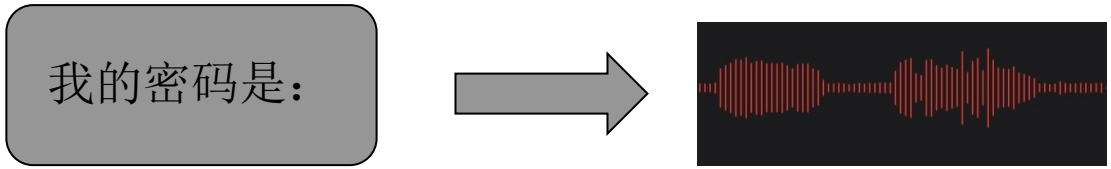
欺骗与对策

欺骗：语音合成

- 语音合成是一种欺骗形式，使用合成的方法伪造说话人的语音。目前，随着语音合成技术的快速发展，语音合成已经成为了自动说话人认证系统的主要威胁。

对策：

- 相位检测法

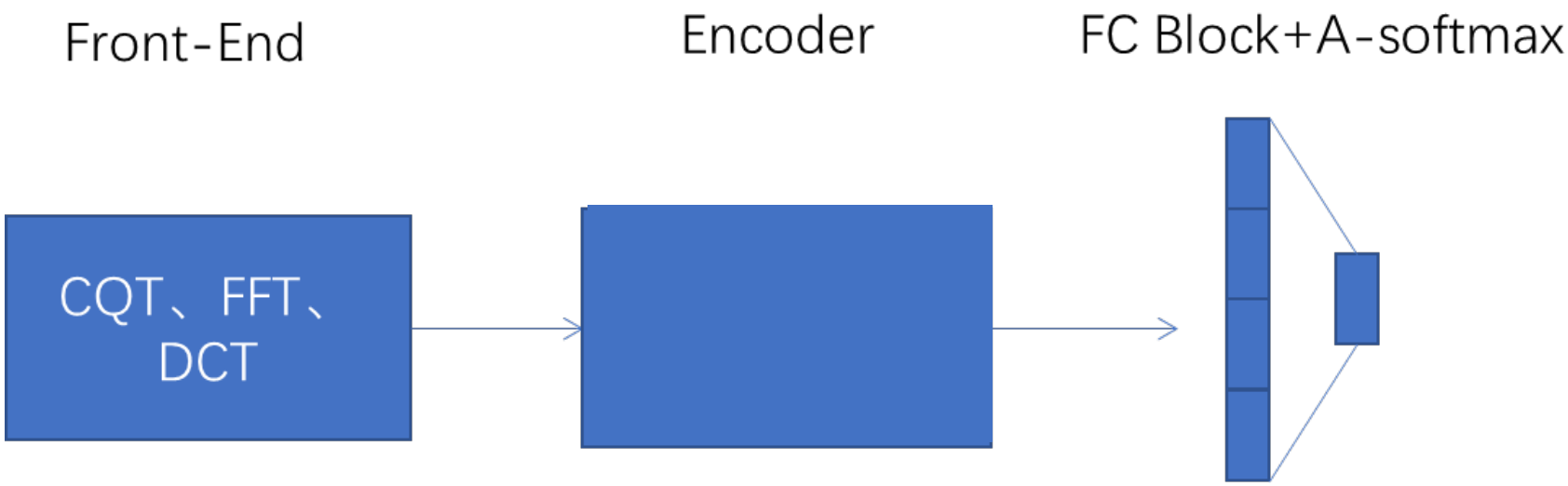


欺骗与对策

欺骗：语音转换

- 语音转换的目的是通过分析语音信号中发音人的个性信息，通过声学手段改变语音信号中发音人的属性而保持语音内容以及背景信息不变，使得伪造语音听起来像是目标说话人发出的。
- 对策：
- 相位检测法

防语音伪造模型框架



其中，Encoder可以是任意的模型，比如传统的GMM模型或者是神经网络模型。

特征与模型

- 特征层面：MFCC、CQCC、IMFCC、log Spectrum

- 模型层面：GMM模型、CNN模型、Resnet模型

存在挑战

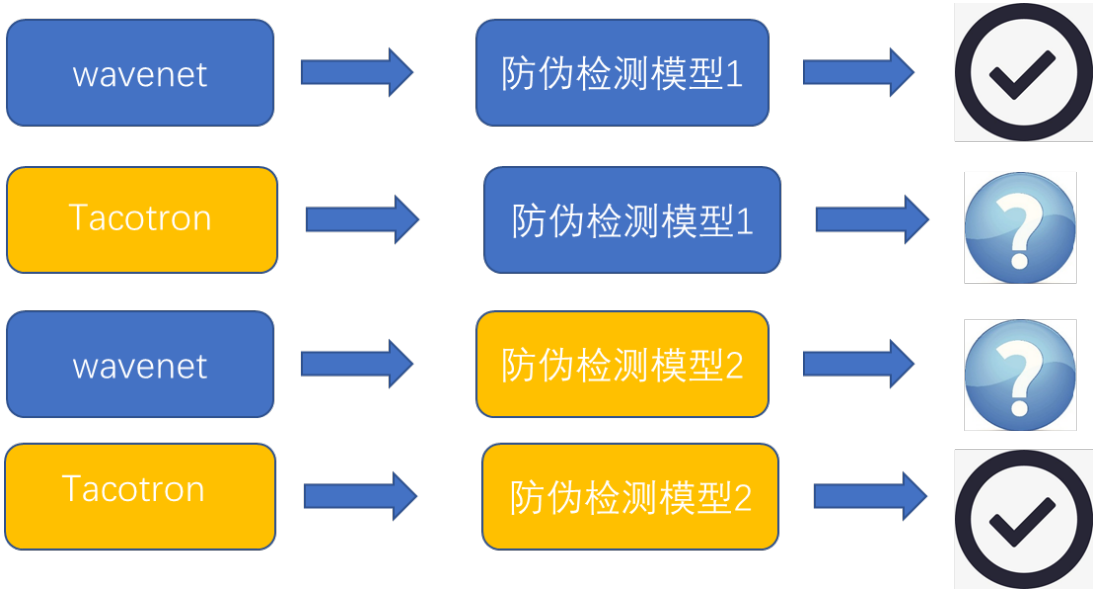
可解释性不足

- 目前的神经网络模型，本质上类似一个“黑盒子”，虽然效果较传统方法有所提高，但是溯源性不足。

存在挑战

伪造检测模型的普适性不足

- 声音的防伪检测要求系统具有鲁棒性，即能够检测出来自于多种不同的伪造系统的伪造声音。由于缺乏大规模数据集，使得目前伪造的声音不具有多样性，基于这样数据集训练出来的模型虽然能够鉴别出部分伪造声音，但是普适性不足。



本节课总结

- 声纹识别
  - 声纹、评价指标EER等基本概念
- 模型与应用
  - 经典模型与端到端模型
- ASVspoof
  - 欺骗与对策

谢谢！