



## 第六章 声纹识别

# 第六章 声纹识别

- 6.1 声纹识别概述
- 6.2 传统声纹识别算法（GMM-UBM）
- 6.3 基于深度学习的声纹识别算法
- 6.4 声纹识别技术的展望

# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念（基本概念）
- 6.1.2 声纹识别方法的回顾（了解、基本概念）
- 6.1.3 声纹识别的典型应用（了解）

## ■ 6.2 传统声纹识别算法（GMM-UBM）

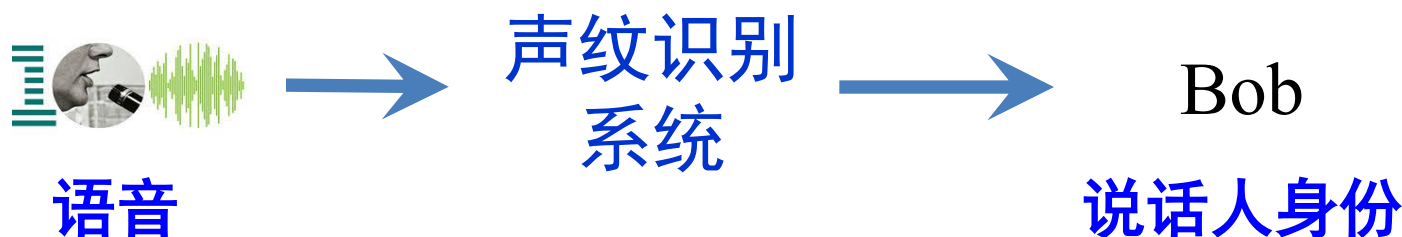
## ■ 6.3 基于深度学习的声纹识别算法

## ■ 6.4 声纹识别技术的展望

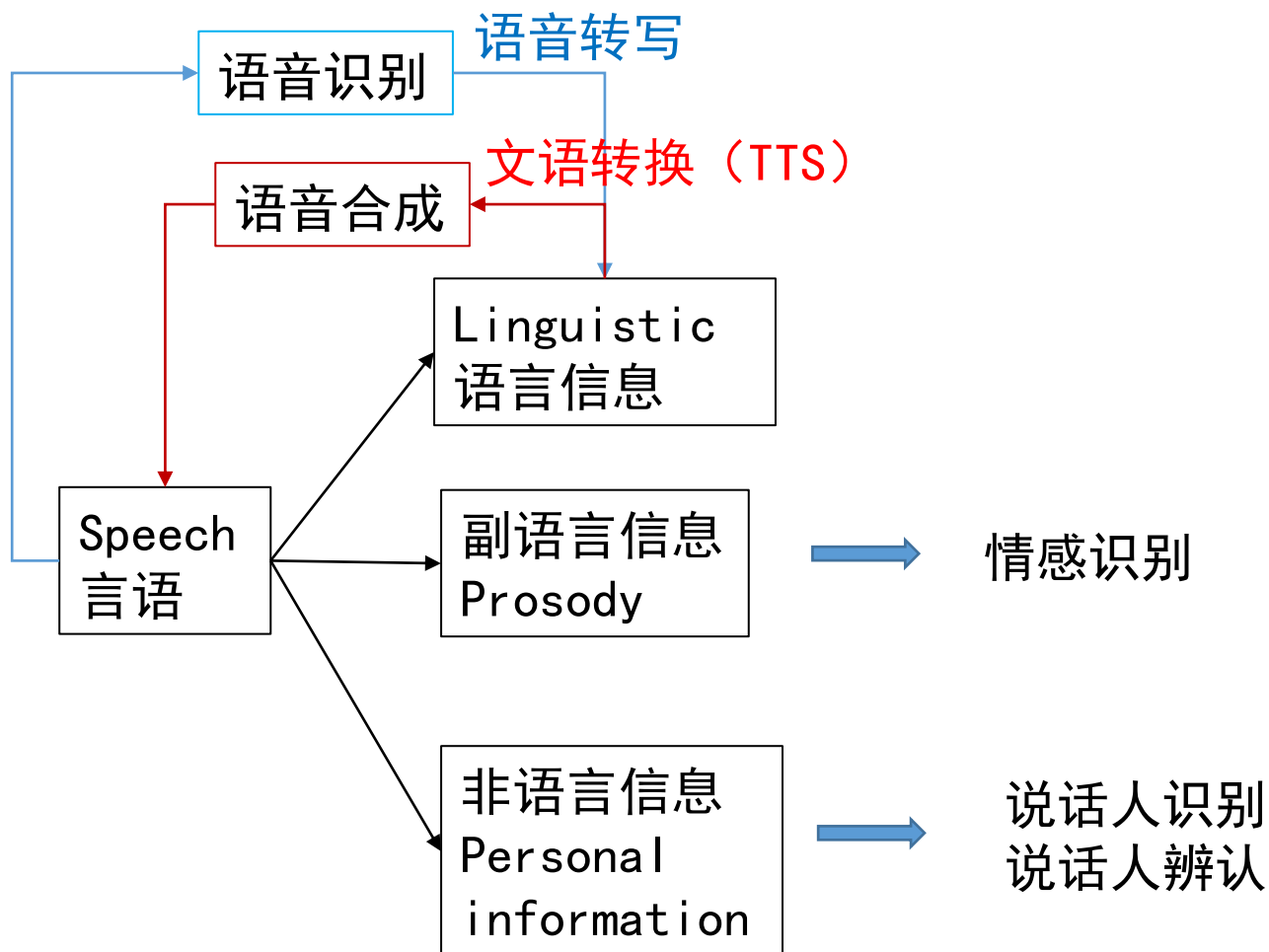
# 什么是声纹识别？

## ■ 定义：

- 声纹识别（说话人识别）作为生物识别（指纹识别、掌纹识别、人脸识别和虹膜识别等）的一种，是根据说话人的声音特性进行身份辨识的任务。
- 输入：语音
- 输出：说话人身份

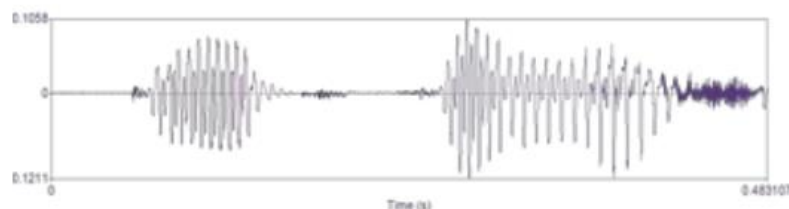


# 语音中语言、副语言和非语言信息的感知（回顾）



# 与其他语音处理方法的区别

- 语音识别：输出语音携带的文本信息；
- 情感识别：输出语音携带的情感信息；
- 语种识别：输出语音所属的语种；
- 语音合成：由文本输出语音；
- **声纹识别**：输出语音携带的说话人身份信息。



“I am Bob”

Identity: Bob

Message: “I am Bob”

English, Male, Happy,  
Middle age, etc.

# 说话人身份信息相关的语音属性

- 低级属性：与发声器官的生理方面相关
  - 声道频谱；
  - 共振峰轨迹。
- 高级属性：与说话方式的行为差异相关
  - 韵律；
  - 口音；
  - 话语中词的选择。

# 声纹识别的分类

## ■ 按照输入的模式

### □ 文本相关的说话人识别（Text-dependent）

- 注册和测试话语共享（至少部分）相同的内容；
- 目的是分析给定语音语境下的说话人特征；
- 更高的性能。

### □ 文本无关的说话人识别（Text-independent）

- 注册和测试话语可以使用任意文本（甚至不同的语言）；
- 不使用关于口语内容的先验信息；
- 更高的便利性。



# 声纹识别的分类（续）

## ■ 按照输出的模式

### □ 说话人辨认（Speaker Identification）

- 判定待识别说话人来自注册集中的哪个说话人；
- 1：N 语音的对比。

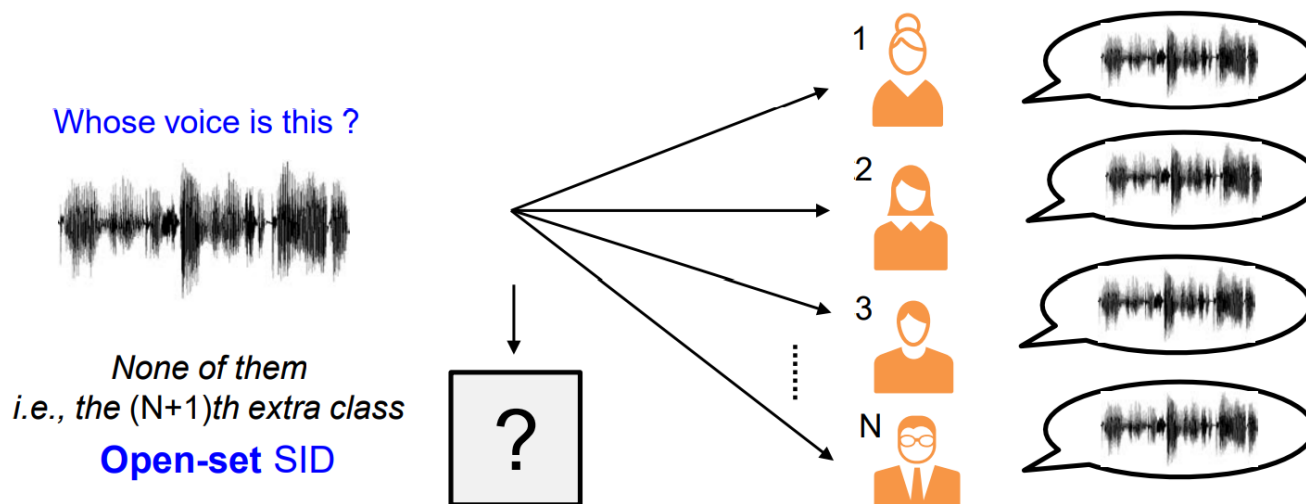
### □ 说话人确认（Speaker Verification）

- 判定待识别语音是否来自注册集中的某个说话人；
- 成对或1：1语音的对比。

# 声纹识别的分类（续）

## ■ 说话人辨认（Speaker Identification）

- 判定待识别说话人来自注册集中的哪个说话人；
- 来自未知说话者的语音样本与一组标记样本进行比较，未知说话人被识别为语音与输入语音样本最匹配的说话人。



确定未知说话人是否是一组人中的特定人

# 声纹识别的分类（续）

## ■ 说话人确认 (Speaker Verification)

- 判断说话人身份是否与其声明的身份相符；
- 由未知说话人提出身份声明。来自说话人的语音样本与声称身份的注册样本进行比较。如果匹配足够好（即通过给定的阈值），则系统接受身份声明。
- 辨认任务可以分解为N对说话人的比较。验证任务更为基础。

Is this Bob's voice ?



Same or  
different?



# 声纹识别评价指标

## ■ 等错误率 (Equal Error Rate)

- 调整阈值, 使得误拒绝率(False Rejection Rate, FRR)等于误接受率 (False Acceptance Rate, FAR), 此时的FAR与FRR的值称为等错误率。

- 误拒绝率(False Rejection Rate, FRR):

- 被错误拒绝的有效身份声明的比例。

$$FRR = \frac{\text{本该匹配成功却被判匹配失败的次数}}{\text{总的匹配成功次数}}$$

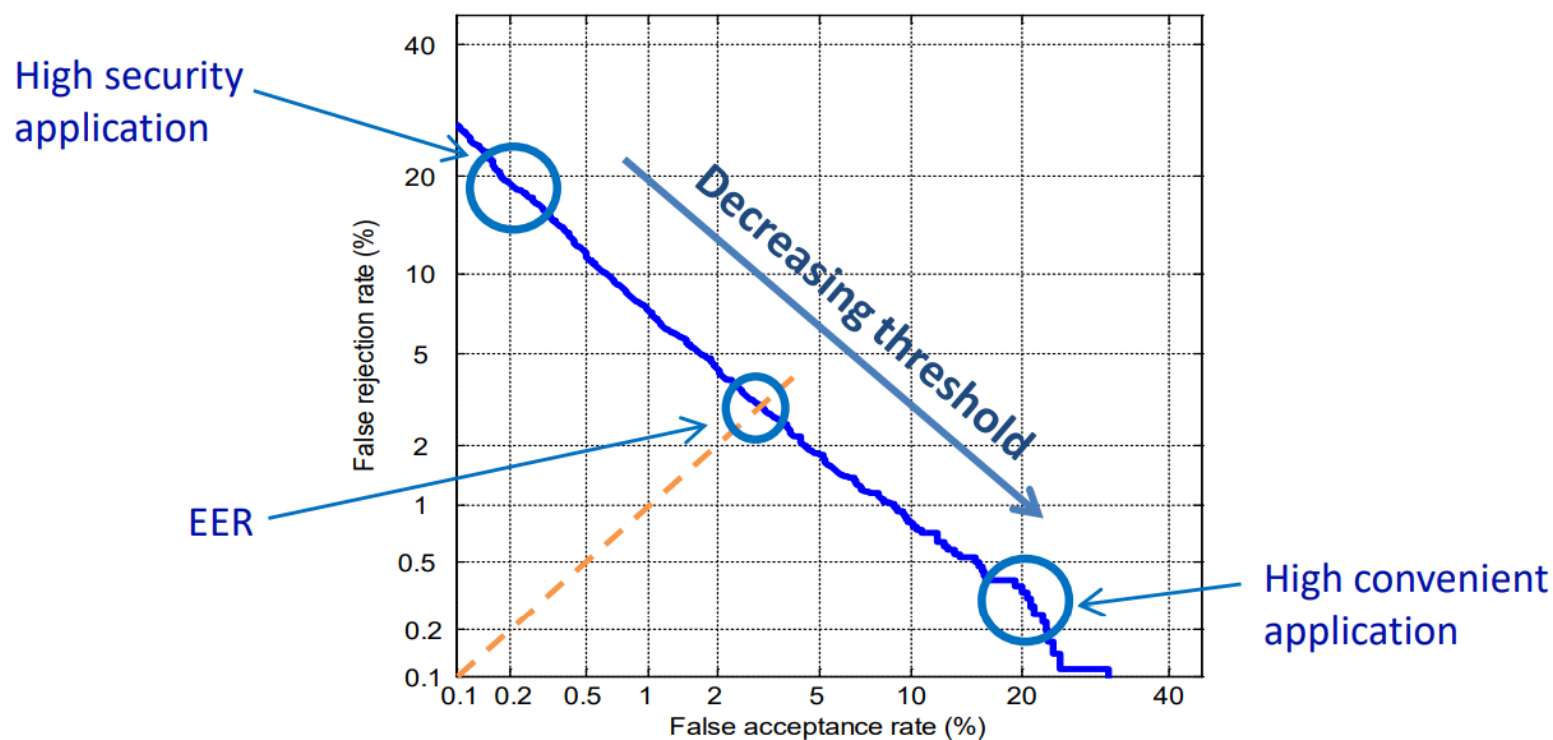
- 误接受率 (False Acceptance Rate, FAR):

- 错误接受冒名顶替者身份声明的比例。

$$FAR = \frac{\text{本该匹配失败却被判匹配成功的次数}}{\text{总的匹配失败次数}}$$

# 声纹识别评价指标（续）

- 等错误率（Equal Error Rate）
- 检测-错误-均衡曲线（Detection Error Tradeoff: DET）



# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念

- 6.1.2 声纹识别方法的回顾

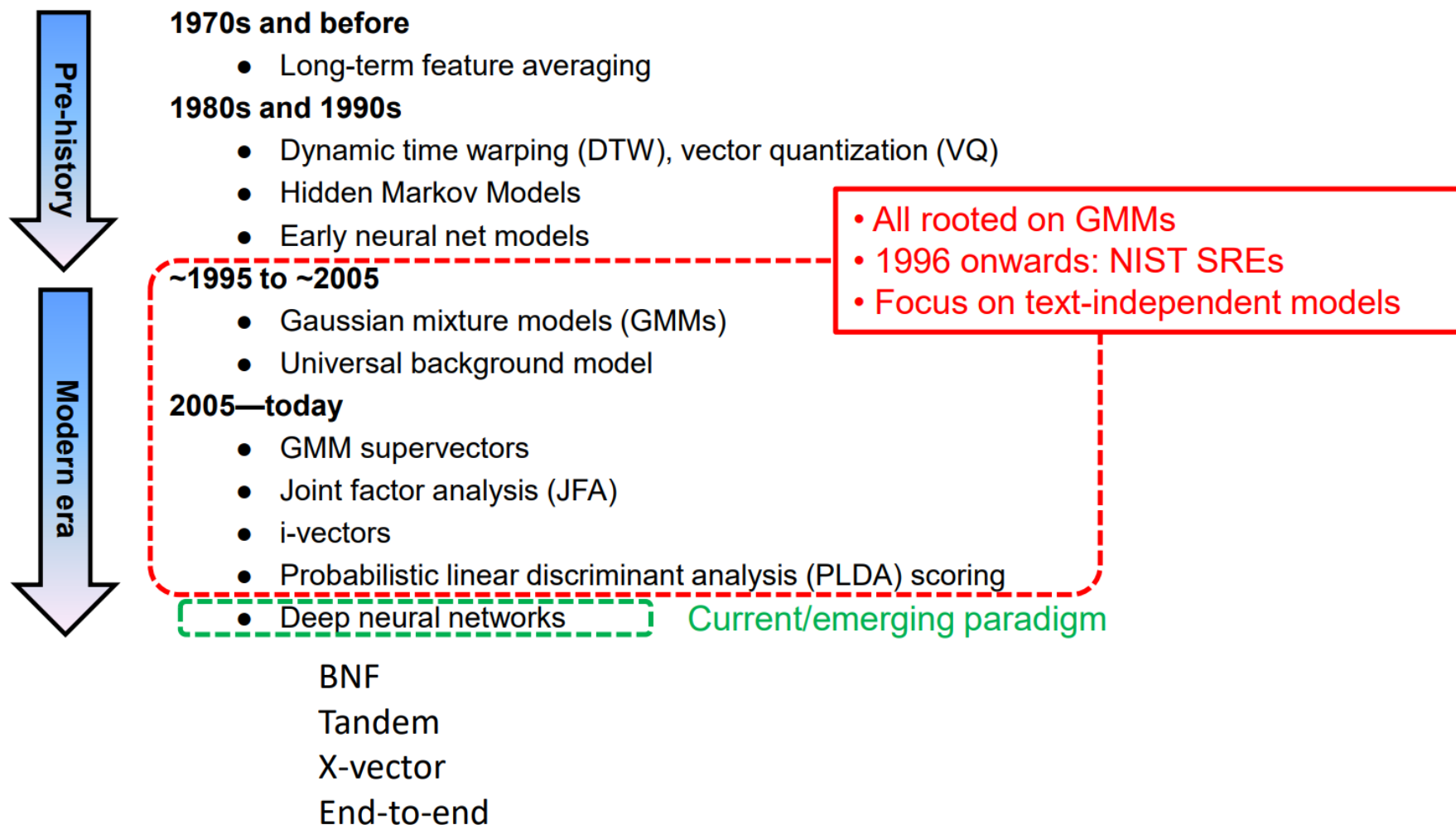
- 6.1.3 声纹识别的典型应用

## ■ 6.2 传统声纹识别算法（GMM-UBM）

## ■ 6.3 基于深度学习的声纹识别算法

## ■ 6.4 声纹识别技术的展望

# 声纹识别技术的演化（了解）



# 声纹识别系统基本框架（掌握）

## ■ 特征提取

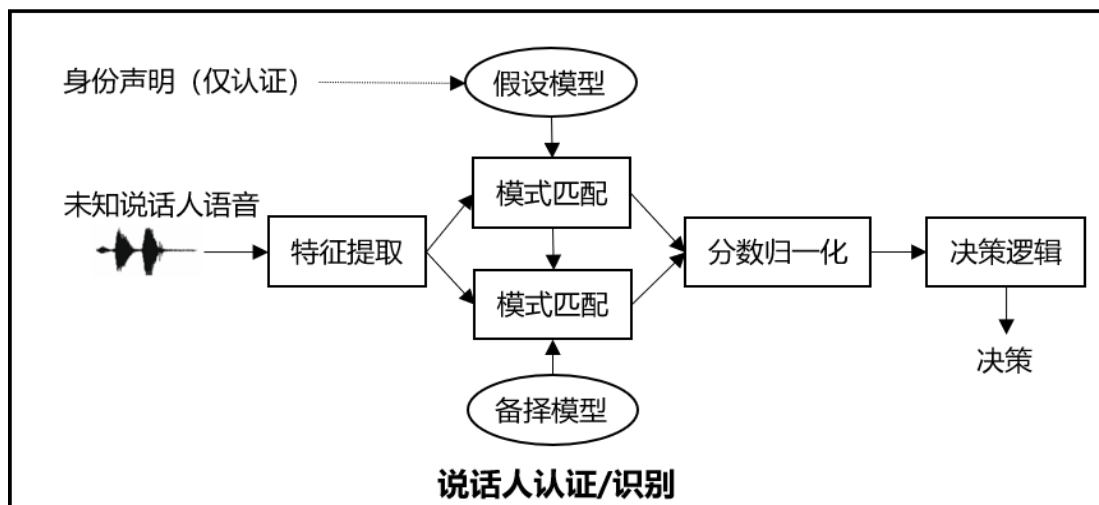
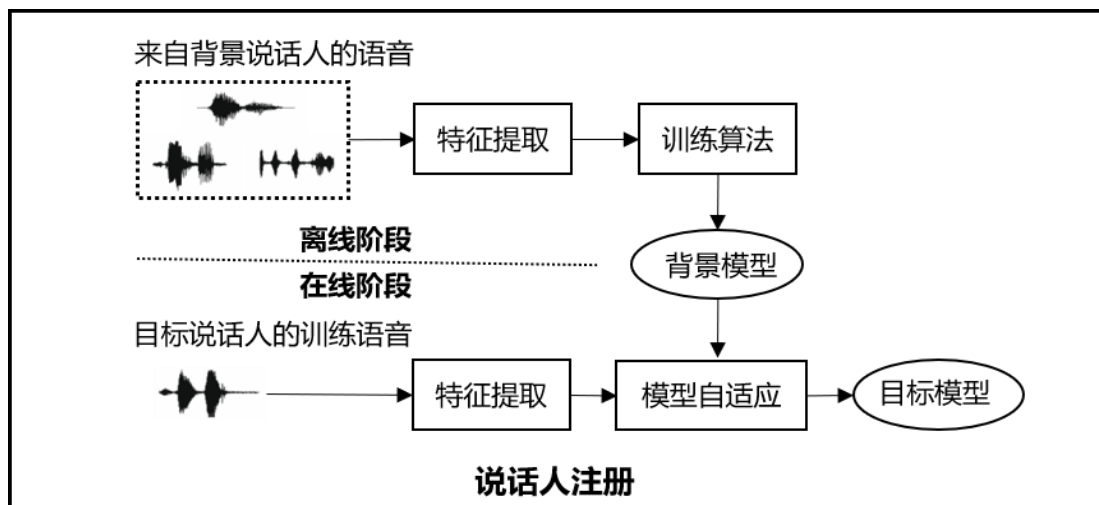
- 将原始语音信号转换为包含说话人信息为主的特征向量序列。

## ■ 注册

- 用以与测试段对照，判断访问者是否与其声明身份相符。

## ■ 测试

- 从待测语音中提取特征向量，并与每个注册说话者的存储模型进行比较。
- 根据相似度（或似然度）值做出识别决策。





# GMM-UBM（基本概念）

## ■ Gaussian mixture model (GMM)

- 为目标说话人的声纹特征建模；
- GMM将空间分布的概率密度用多个高斯概率密度函数的加权来拟合，可以平滑地逼近任意形状的概率密度函数；
- 是一个易于处理的参数模型，对实际数据具备较强的表征能力。

$$P(\mathbf{x}|\theta) = \sum_{m=1}^M c_m N(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$
$$N(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{D/2}(\boldsymbol{\Sigma}_m)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{x} - \boldsymbol{\mu}_m)\right\}$$

# GMM-UBM (续) (基本概念)

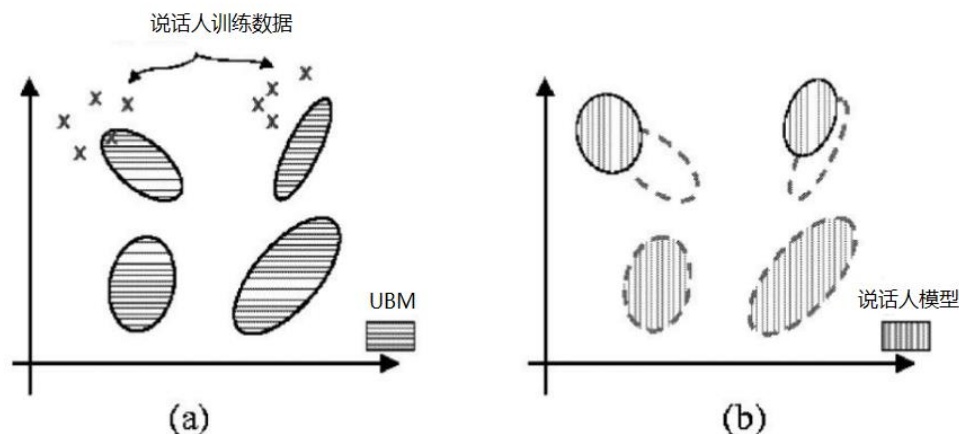
## ■ Universal Background Model (UBM)

- 实际场景中每一个说话人的语音数据很少，导致无法训练出高效的GMM模型；
- 在复杂的声学环境和信道场景下，训练GMM模型的语音与测试语音存在失配的情况；
- 为解决上述问题：
  - 使用EM算法对说话人无关的领域模型或通用背景模型(UBM)进行训练；
  - 基于UBM通过自适应算法（如最大后验概率：MAP）来得到目标说话人模型。

# GMM-UBM (续) (基本概念)

## ■ 使用MAP的原因

- UBM是初始模型，可以减少注册新说话人所需的语音发生的持续时间/数量（导出依赖于说话人的GMM参数）；
- 由于说话人相关的GMM和说话人无关的UBM具有相同的模型复杂度，因此可以计算更好的对数似然分数；
- 使用高斯选择可以更快。



MAP算法自适应

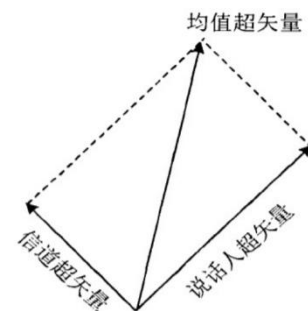
# 联合因子分析（Joint factor analysis: JFA）（了解）

## ■ 目的

- 在GMM超向量空间中进行信道补偿。

## ■ 实现

- GMM 超向量 $\mathbf{m}_r$ 被建模为说话人和信道分量的总和。



均值超矢量分解示意图

$$\mathbf{m}_r = \mathbf{m}_o + \mathbf{U}\mathbf{x}_r + \mathbf{V}\mathbf{y}_r + \mathbf{D}\mathbf{z}_r = \mathbf{m}_o + \underbrace{\mathbf{V}\mathbf{y}_r + \mathbf{D}\mathbf{z}}_{\text{Speaker supervector}} + \underbrace{\mathbf{U}\mathbf{x}_r}_{\text{Residual variability}}$$

UBM supervector

Channel supervector

# 联合因子分析（续）（了解）

## ■ 注册

- 联合估计说话人因子 $\mathbf{y}_r$ 和信道因子 $\mathbf{x}_r$ 以及残差 $\mathbf{z}_r$ 。
- 信道补偿通过丢弃信道分量 $\mathbf{U}\mathbf{x}_r$ 来实现。

## ■ 测试

- 从测试段估计信道因子 $\mathbf{z}_{test}$
- 计算说话人验证分数

$$s = (\mathbf{V}\mathbf{y}_r + \mathbf{D}\mathbf{z}_r)^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_{test} - \mathbf{N}_{test}\mathbf{m}_o - \mathbf{N}_{test} \overbrace{\mathbf{U}\mathbf{x}_{test}}^{\text{Channel supervector discarded}})$$

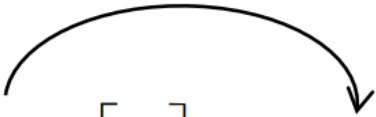
# Identity vector (i-vector) （了解）

## ■ 目的

- 为了解决JFA估算出来的说话人子空间与信道子空间存在互相掩盖的问题，不严格区分说话人空间以及信道空间

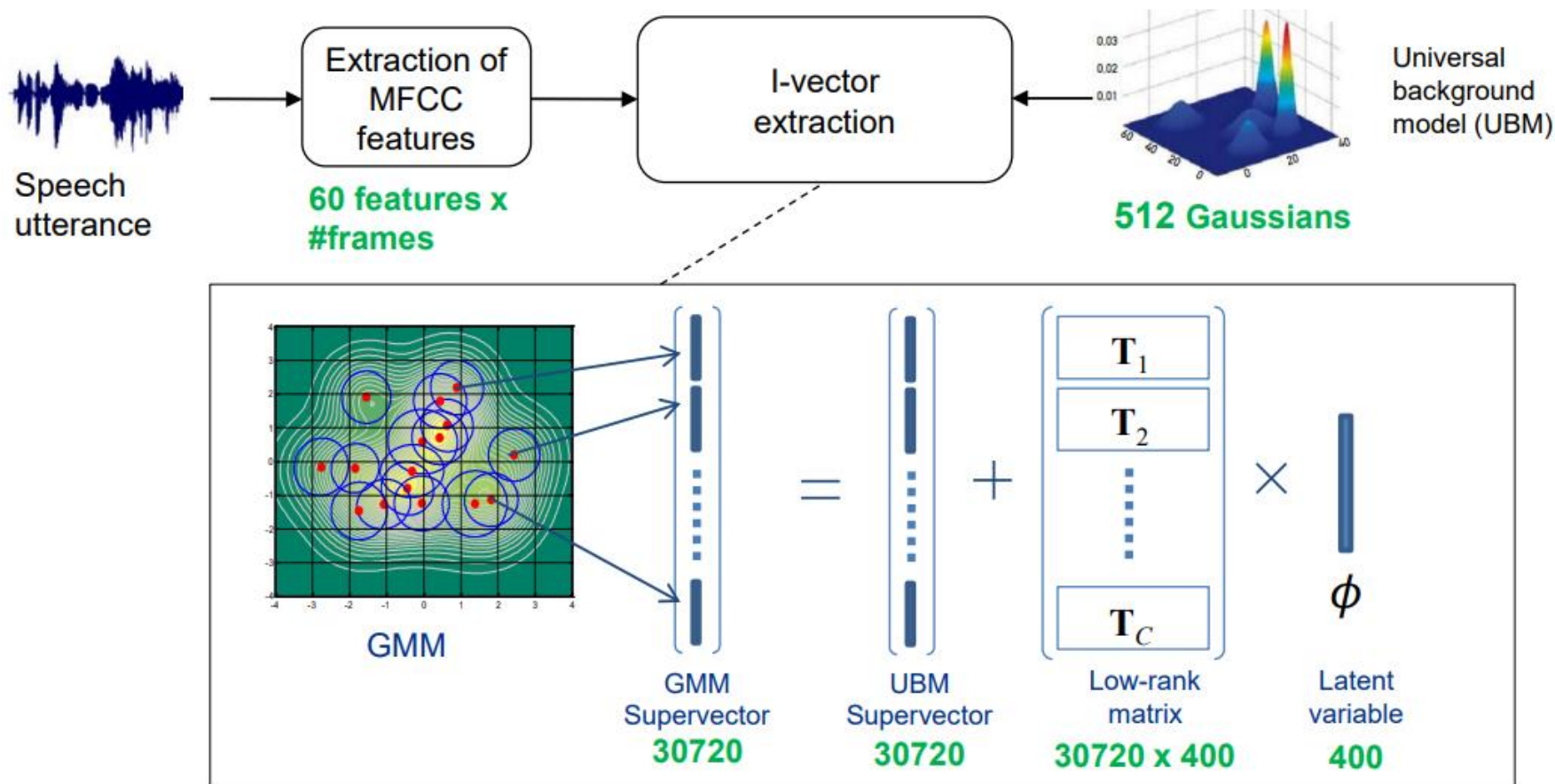
## ■ 简化版本的JFA

“Channel” + “Speaker” = “Total”


$$\mathbf{m}_r = \mathbf{m}_o + \mathbf{U}\mathbf{x}_r + \mathbf{V}\mathbf{y}_r + \cancel{\mathbf{D}\mathbf{z}_r} = \mathbf{m}_o + [\mathbf{U}, \mathbf{V}] \times \begin{bmatrix} \mathbf{x}_r \\ \mathbf{y}_r \end{bmatrix} = \mathbf{m}_o + \mathbf{T}\mathbf{h}_r$$

- 信道和说话人子空间形成一个总的可变空间
- l-vector包含人的语音特征（来自说话人子空间）和信道因子（来自信道子空间）

# Identity vector (i-vector) (了解)



1. 通过EM算法迭代计算全局差异空间矩阵 $T$ ;
2. 利用 $T$ 计算 $\mathbf{h}_r$ 的后验分布的均值 (i-vector)。

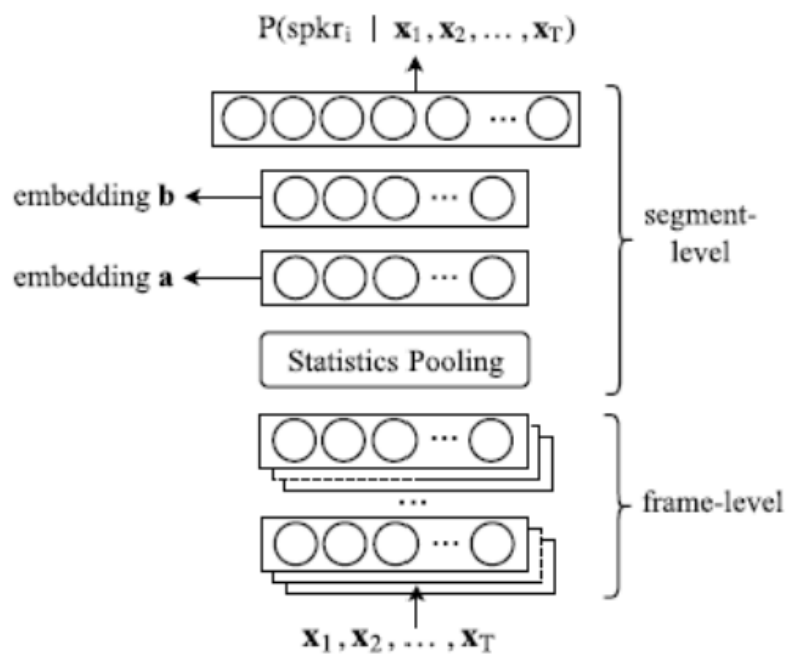
# X-vector (基本概念)

## ■ 目的

- 使用监督方式训练的深度神经网络导出句子级别的嵌入表示。

## ■ X-vector提取器

- TDNN：经过训练以在输出层区分说话人；
- 统计池化层：计算均值和标准差；
- 段的统计信息被传递到另一个隐层以产生嵌入；
- 数据增广是必要的。



[Source: Snyder et al, 2017]



# Probabilistic LDA (PLDA)简介（了解）

## ■ PLDA定义

- 嵌入表示（例如i-vector和x-vector）中既包含了说话人信息，也包含了信道信息。
- PLDA根据说话人身份和信道效应解释观察到的嵌入表示，以实现对话说话人因子的“提纯”。
- 可以看作是 LDA 的概率形式。

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\mu} + \boldsymbol{V}\mathbf{y}_i + \boldsymbol{U}\mathbf{x}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

$\boldsymbol{\eta}_{ij}$ 代表第*i*个说话人的第*j*段语音的嵌入表示； $\boldsymbol{\mu}$ 为所有嵌入表示的全局均值； $\boldsymbol{V}$ 是说话人空间矩阵 (Eigen Voice)，用于描述说话人的特征； $\boldsymbol{U}$ 是信道空间矩阵 (Eigen Channel)，用于描述信道的特征； $\mathbf{y}_i$ 与 $\mathbf{x}_{ij}$ 是其对应子空间内的因子，服从高斯分布； $\boldsymbol{\varepsilon}$ 是残差项，服从协方差矩阵为对角阵的高斯分布。

# Probabilistic LDA (PLDA)简介（了解）

## ■ 训练：

- PLDA的模型参数一共有4个，分别是i-vector/x-vector均值 $\mu$ ，空间特征矩阵 $V$ 和 $U$ ， $\epsilon$ 噪声协方差。由于模型含有隐变量，模型的训练过程采用经典的EM算法迭代求解。

## ■ 测试：

- 对于说话人确认任务，每组试验都需要一个目标说话人和一个测试说话人。分别提取目标说话人和测试说话人的i-vector/x-vector，使用PLDA模型计算它们之间的似然度评分。

# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念

- 6.1.2 声纹识别方法的回顾

- 6.1.3 声纹识别的典型应用

## ■ 6.2 传统声纹识别算法（GMM-UBM）

## ■ 6.3 基于深度学习的声纹识别算法

## ■ 6.4 声纹识别技术的展望

# 访问控制

- 访问房间、建筑物和物理资产
- 身份和凭证认证

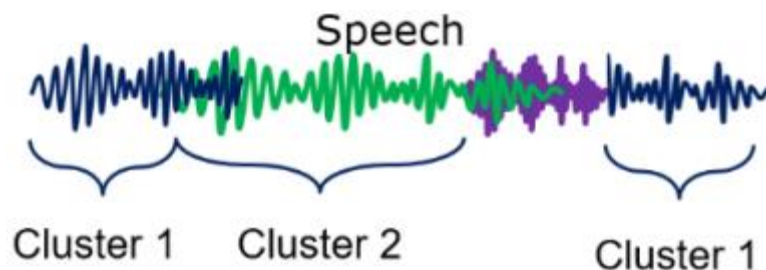


Lenovo A586  
[Lee et al, SLTC  
News Letter,  
2013]

# 说话人日志

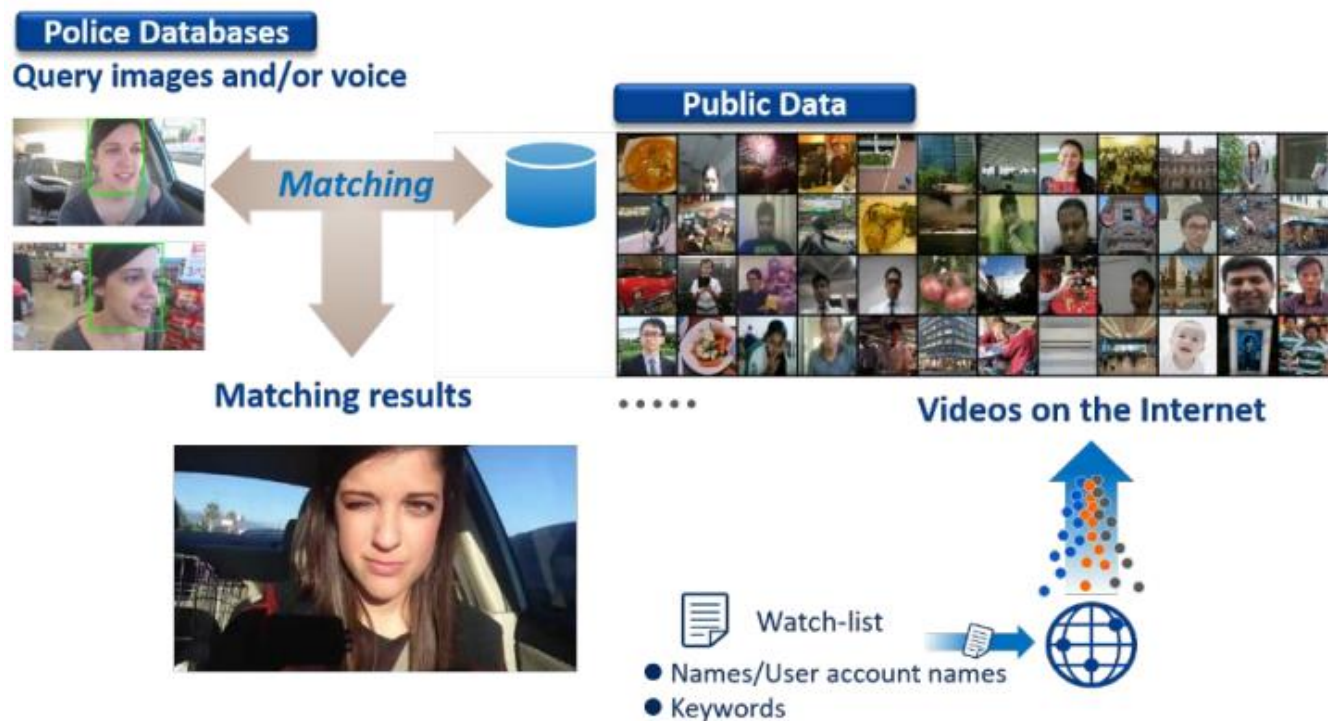
- 谁在什么时候说话
- 会议转录
- 会议分析（不同会议参与者的积极性）

Diarization:  
Who spoke when



# 搜索/索引

- 索引多媒体档案
- 情报、反恐



# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念
- 6.1.2 声纹识别方法的回顾
- 6.1.3 声纹识别的典型应用

## ■ 6.2 传统声纹识别算法（GMM-UBM）（基本概念）

## ■ 6.3 基于深度学习的声纹识别算法

## ■ 6.4 声纹识别技术的展望

# 特征提取

## ■ 目的

- 保留说话人信息；
- 去除噪声和其他干扰信息。

## ■ 声学特征

- 梅尔频率倒谱系数（MFCC）
- 线性预测倒谱系数（LPCC）

## ■ 完整的特征提取前端还包括

- 语音活动检测（丢弃非语音帧）
- 后处理（Cepstral Mean Normalization等）



# 基于GMM的说话人建模

## ■ 注册

- 说话人模型通过来自该说话人的注册语音构建；
- 每个注册语音 $X$ 是一个由特征提取前端生成的特征向量序列 $\{\mathbf{x}_t\}_{t=1}^T$ 。

$$p(\mathbf{x} | \theta) = \sum_{k=1}^K w_k \mathbf{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- 其中， $K$ 表示高斯分量的数量；
- 集合 $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^K$ 表示分布的参数；
- 权重 $w_k$ ，对于所有的 $k$ 值，其和总为1。

# GMM的期望最大化算法 (EM) (了解)

- 给定注册数据 $\{\mathbf{x}_t\}_{t=1}^T$ ， $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^K$ 的最大似然估计可以使用EM算法获得。
- 从一些随机初始化开始迭代更新参数。

# GMM的期望最大化算法 (EM) (了解)

## ■ E-step

- 计算每个特征向量属于第 $k$ 个高斯分布的程度。

$$\lambda_k(\mathbf{x}_t) = \frac{N(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}{\sum_{i=1}^K N(\mathbf{x}_t | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot w_i}$$

$$\sum_{k=1}^K \lambda_k(\mathbf{x}_t) = 1$$

## ■ M-step

- 基于每个帧的 $\lambda_k(\mathbf{x}_k)$ 更新均值 $\boldsymbol{\mu}_k$ 、协方差矩阵 $\boldsymbol{\Sigma}_k$ 和权重 $w_k$ 。

$$w_k = \frac{1}{T} \sum_{t=1}^T \underbrace{\lambda_k(\mathbf{x}_t)}_{n_k} \quad \left| \quad \boldsymbol{\mu}_k = \frac{1}{n_k} \times \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \cdot \mathbf{x}_t \quad \left| \quad \boldsymbol{\Sigma}_k = \frac{1}{n_k} \times \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \cdot (\mathbf{x}_t - \boldsymbol{\mu}_k)(\mathbf{x}_t - \boldsymbol{\mu}_k)^T$$

Total number of frames aligned to the  $k$ -th Gaussian component

# 识别（辨认或确认）

- 测试语音与说话人模型的匹配程度被视为平均对数似然，如下所示：

$$s(Y | \theta) = \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{y}_t | \theta)$$

- 对于辨认（Identification）任务，选择得分最高的说话人模型。
- 对于验证（Verification）任务，决策基于以下形式的对数似然比

$$\Lambda(Y) = s(Y | \theta) - s(Y | \theta_{bg})$$

# 最大后验 (MAP) 自适应

- 调整来自 UBM 的所有参数（权重、均值向量和协方差矩阵）。

- 计算语音帧到混合高斯分量的相似度：

$$\lambda_k(\mathbf{x}_t) = \frac{N(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}{\sum_{i=1}^K N(\mathbf{x}_t | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot w_i}$$

- 计算零阶和一阶统计量：

$$N_k = \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \quad \Bigg| \quad F_k = \sum_{t=1}^T \lambda_k(\mathbf{x}_t) \cdot \mathbf{x}_t$$

- 调整参数（如平均向量  $\boldsymbol{\mu}_k$ ）：

$$\boldsymbol{\mu}'_k = \underbrace{\alpha_k \left( \frac{1}{N_k} F_k \right)}_{\text{New information}} + (1 - \alpha_k) \underbrace{\boldsymbol{\mu}_k}_{\text{UBM mean vector}}$$

# 最大后验（MAP）自适应（续）（了解）

## ■ 适应系数

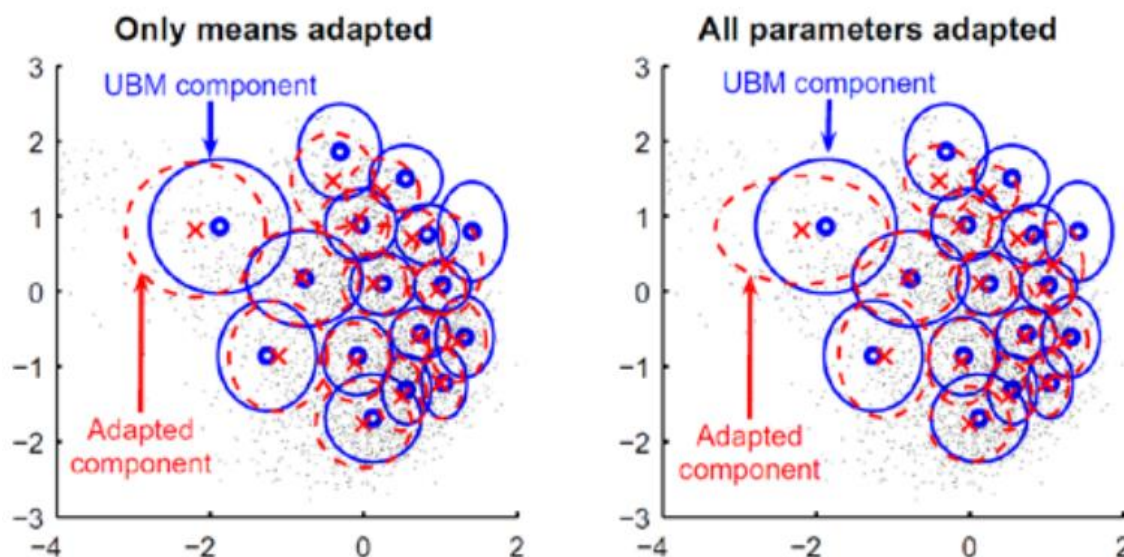
$$\alpha_k = \frac{N_k}{N_k + r}$$

- $\alpha_k$  始终在  $0 < \alpha_k < 1$  范围内；
- 控制新信息和旧参数之间的平衡；
- 每个高斯分量不同，它取决于与特定混合对齐的帧总数（即零阶统计数据）和相关因子  $r$ （通常取0-20）。

# 最大后验（MAP）自适应（续）（了解）

## ■ 二维向量空间中MAP自适应图示

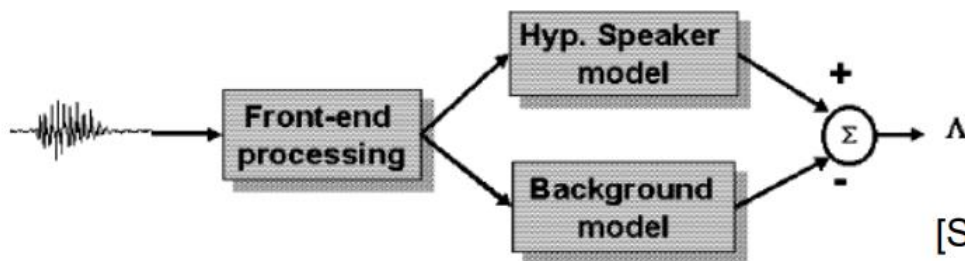
- 平均适应（Mean adaptation）导致高斯分量偏移；
- 协方差适应（Covariance adaptation）导致高斯分量的大小（不确定性）发生变化。



# GMM-UBM识别（掌握）

- 说话人验证（Verification）任务可以被转换为以下两项的似然比测试：
  - 零假设， $H_0$ ：测试段 $Y$ 来自假设说话人；
  - 备择假设， $H_1$ ：测试段 $Y$ 不是来自假设说话人。

$$\Lambda(Y) = \log \frac{p(Y|H_0)}{p(Y|H_1)} \approx \log \frac{p(Y|\theta_{\text{spk}})}{p(Y|\theta_{\text{UBM}})} = s(Y|\theta_{\text{spk}}) - s(Y|\theta_{\text{UBM}})$$



[Source: Reynolds et al, 2000]



# GMM-UBM识别（续）（掌握）

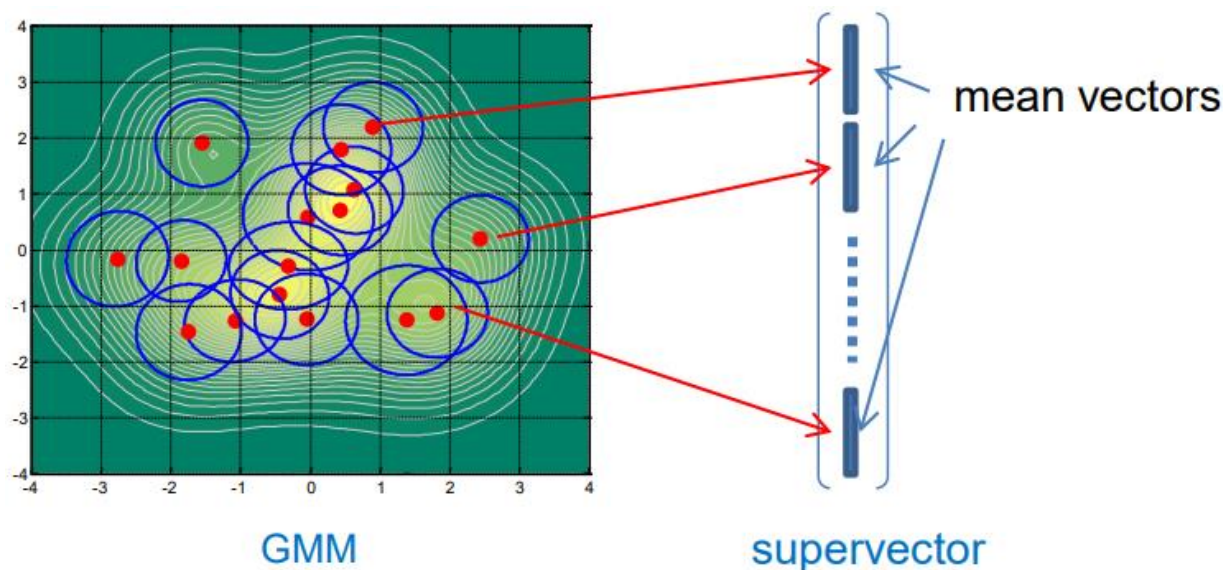
- 给定测试段和说话人模型 $\theta$ 的对数似然相似度：

$$s(Y|\theta) = \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{y}_t|\theta) = \frac{1}{T} \sum_{t=1}^T \log \underbrace{\sum_{k=1}^K w_k \mathbf{N}(\mathbf{y}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Speaker GMM or UBM}}$$

# GMM超向量

## ■ GMM-UBM范式

- 对于每个说话者，基于相同的 UBM 估计说话者相关的 GMM；
- 说话人相关的 GMM 由均值向量唯一表示，方法是仅调整均值向量，而权重和协方差矩阵保持与 UBM 相同。



# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念

- 6.1.2 声纹识别方法的回顾

- 6.1.3 声纹识别的典型应用

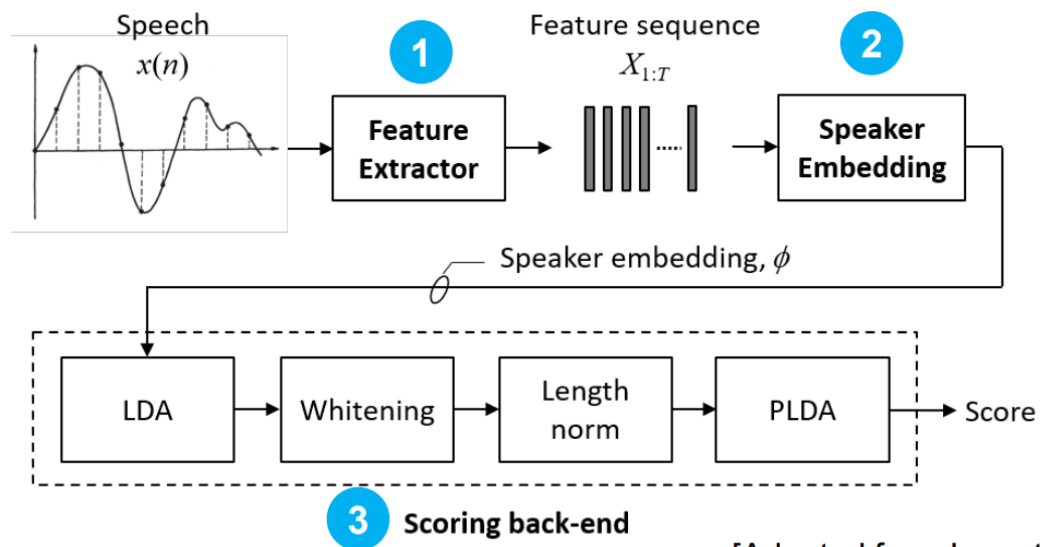
## ■ 6.2 传统声纹识别算法（GMM-UBM）

## ■ 6.3 基于深度学习的声纹识别算法（基本概念）

## ■ 6.4 声纹识别技术的展望

# 现代声纹识别系统

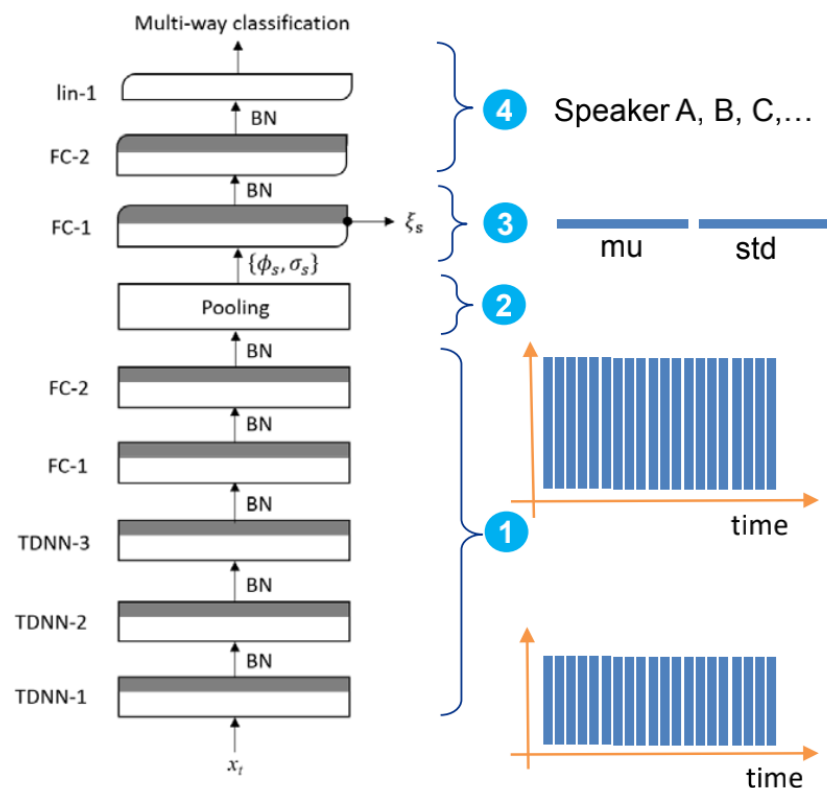
- 现代声纹识别系统主要基于说话人嵌入，包括：
  - 特征提取；
  - 说话人嵌入；
  - 打分。



[Adapted from Lee et al, CSL 2020]

# 基于x-vector的深度说话人嵌入（掌握）

- x-vector是一个包含四个函数块的深度神经网络
  - 使用多层神经网络实现帧处理器用以提取深度特征，例如延时神经网络(TDNN)
  - 池化层
  - 嵌入提取
  - 基于判别性损失的说话人分类



# x-vector提取器的总体架构（掌握）

## ■ 深度特征提取

- 获取一系列声学特征，例如MFCC或对数mel-filterbank系数；
- 采用多层深度神经网络从声学特征中提取深度说话人特征。

## ■ 时间池化

- 将特征向量序列转换为单个固定维向量；
- 时间池化（Temporal pooling）（取特征向量的均值和标准差）是最常见的池化方式；

# x-vector提取器的总体架构（续）（掌握）

## ■ 嵌入提取

- 由若干个相互连接的网络层组成；
- 其中一层被设计为瓶颈层（Bottleneck layer）；
- 瓶颈层的输出（非线性层之前）就是x-vector说话人嵌入。

## ■ 说话人分类

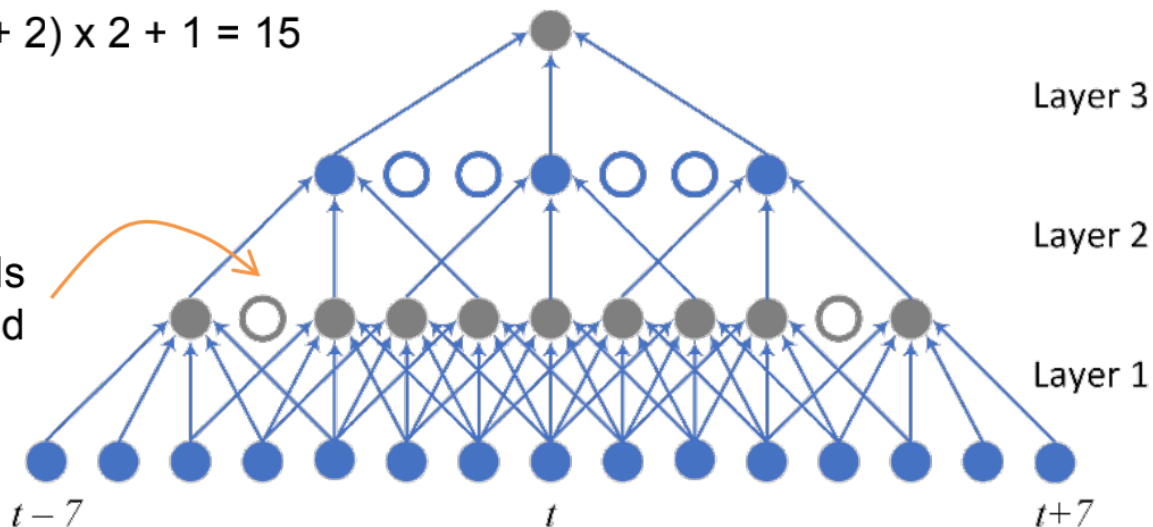
- 根据说话人标签对训练音频进行分类。

# TDNN中的时间上下文

- TDNN对时间上下文信息的提取取决于感受野的宽度
  - 感受野：影响网络特定单元的输入空间区域。

$$\text{Temporal context} = (3 + 2 + 2) \times 2 + 1 = 15$$

Dilated convolution expands the receptive field





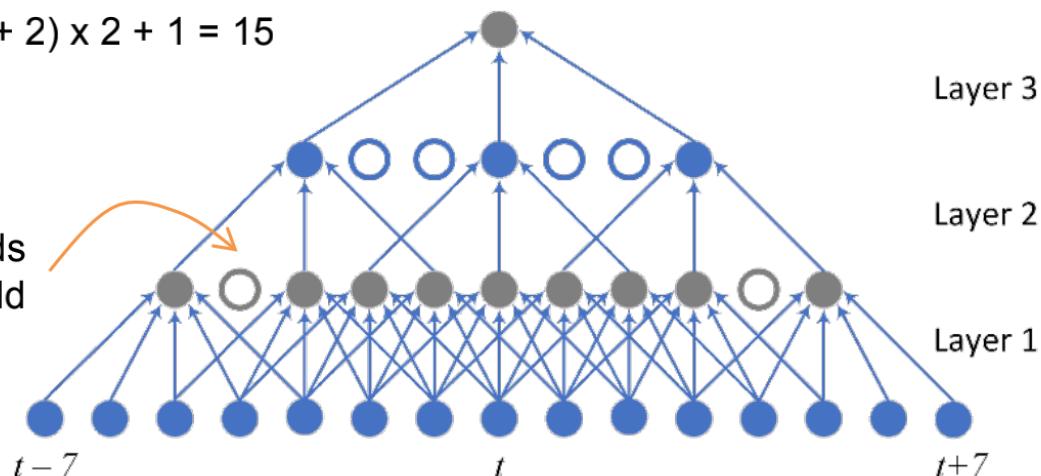
# TDNN中的时间上下文（续）

## ■ 更广泛的时间上下文视野：

- 空洞卷积（通过以一定的步长跳过输入值）；
- 堆叠更多的空洞一维卷积层。

$$\text{Temporal context} = (3 + 2 + 2) \times 2 + 1 = 15$$

Dilated convolution expands the receptive field



# 池化层

- 说话人嵌入通过特征的池化来形成固定长度的表征。
- 时间池化通过平均实现的（计算均值和标准差）。

$$\mu = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \quad \sigma = \sqrt{\frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \odot \mathbf{h}_t - \mu \odot \mu}$$

$$\mathbf{h}_t = g(\mathbf{W}^T \mathbf{f}_t + \mathbf{b})$$

- $\mathbf{h}_t$  - 在第t时刻步变换的特征向量
- $\mathbf{f}_t$  - 前馈层的输入，后跟非线性激活函数 $g(\cdot)$
- $\mathbf{W}$  - 权重矩阵
- $\mathbf{b}$  - 偏差向量

## 池化层（续）（了解）

### ■ 基于注意力机制的池化：

- 通过加权平均令模型更多地关注对说话人识别任务更重要的帧。

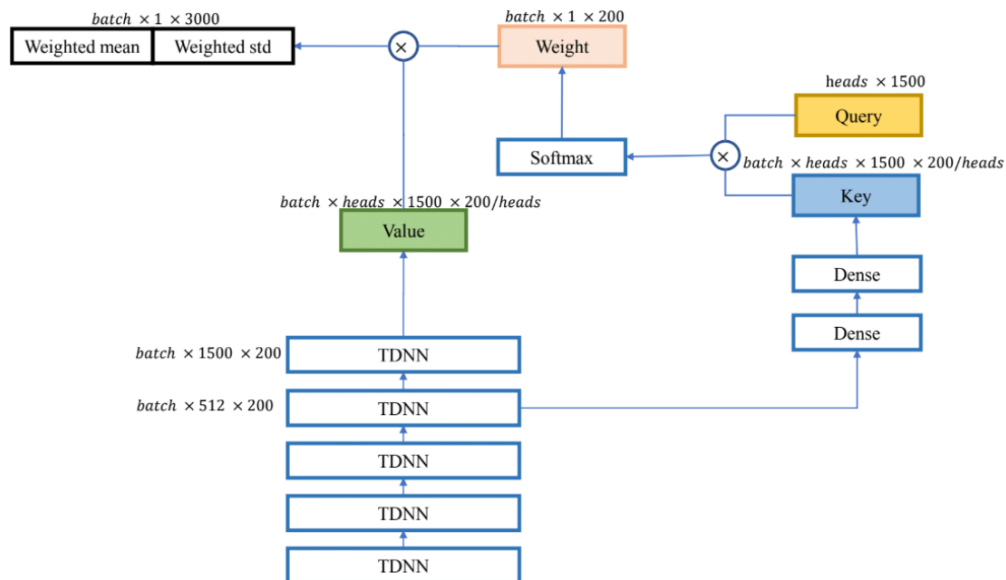
$$\begin{aligned}\mu &= \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \\ &= \sum_{t=1}^T \left( \frac{1}{T} \right) \cdot \mathbf{h}_t \\ &= \sum_{t=1}^T \alpha_t \cdot \mathbf{h}_t\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \odot \mathbf{h}_t - \mu \odot \mu} \\ &= \sqrt{\frac{1}{T} \sum_{t=1}^T \left( \frac{1}{T} \right) \cdot \mathbf{h}_t \odot \mathbf{h}_t - \mu \odot \mu} \\ &= \sqrt{\frac{1}{T} \sum_{t=1}^T \alpha_t \cdot \mathbf{h}_t \odot \mathbf{h}_t - \mu \odot \mu}\end{aligned}$$

# 池化层（续）（了解）

## ■ 自注意力池化：

- 找到一组权重，每个帧对应一个权重，加权求和；
- 类似于Transformer中基于{query, key, value}的自注意力池。

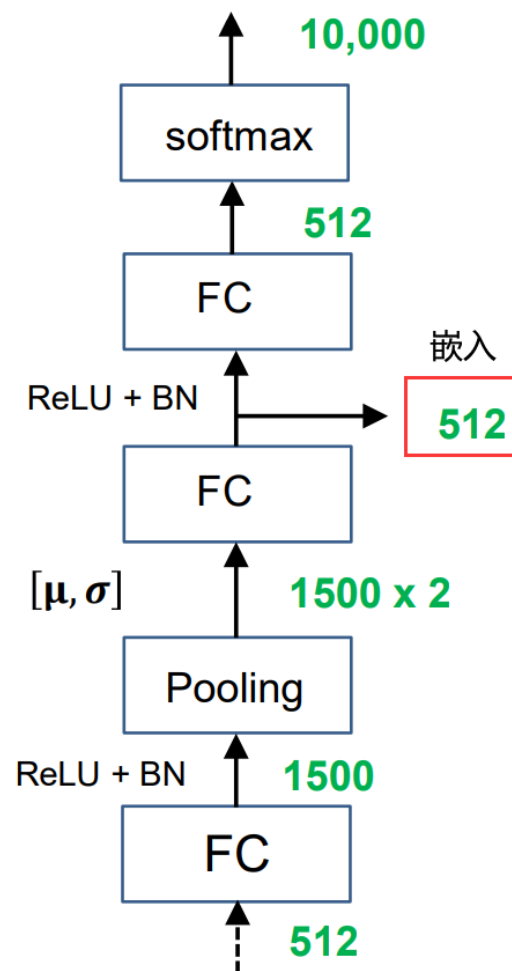


# 嵌入提取

- 说话人嵌入表征是一个固定长度的向量，它捕捉说话人的长期声纹特征；
- 聚合特征通过若干个全连接层进行处理；
- 将第一个线性层的输出作为嵌入表征。

$$\phi = W^T[\mu, \sigma]$$

$$\mu = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \quad \sigma = \sqrt{\frac{1}{T} \sum_{t=1}^T \mathbf{h}_t \odot \mathbf{h}_t - \mu \odot \mu}$$

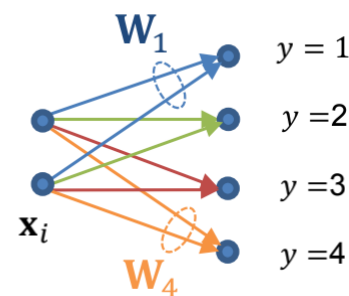
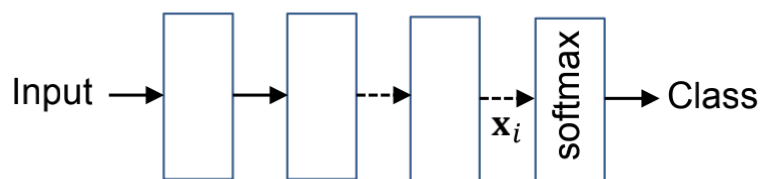


# 多类分类

- 神经网络的输出层被构造为每个节点代表一个类，其中  $y \in \{1, 2, \dots, C\}$  ,  $C$  表示训练集中说话人的数量；
- 使用 softmax 函数（归一化指数函数）激活输出层。

$$l_j(\mathbf{x}_i; \mathbf{W}, \mathbf{b}) = \frac{\exp(\mathbf{W}_j^T \mathbf{x}_i + b_j)}{\sum_{c=1}^C \exp(\mathbf{W}_c^T \mathbf{x}_i + b_c)}$$

- $\mathbf{x}_i$  - 第 $i$ 个训练样本的深度特征
- $\mathbf{W}_j$  - 权重矩阵 $\mathbf{W}$ （输出前馈层）的第 $j$ 列



# 交叉熵损失

- 交叉熵损失用于衡量网络输出和标签之间的差异：

$$\begin{aligned} L &= -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C 1\{y_i = j\} \cdot \log l_j(\mathbf{x}_i; \mathbf{W}, \mathbf{b}) \\ &= -\frac{1}{N} \sum_{i=1}^N \log l_{y_i}(\mathbf{x}_i; \mathbf{W}, \mathbf{b}) \end{aligned}$$

- $\mathbf{x}_i$  - 第 $i$ 个训练样本的深度特征
- $\mathbf{W}_j$  - 权重矩阵 $\mathbf{W}$ （输出前馈层）的第 $j$ 列
- $y_i$  - 第 $i$ 个样本的类标签
- $C$  - 类的数量
- $N$  - 批量大小
- $1\{y_i = j\}$  - 指示函数。如果第 $i$ 个样本属于第 $j$ 个说话者则取值为1，否则为0。

# 改进的softmax 函数（了解）

- 传统的softmax输出层的学习嵌入仅针对类间分离进行了优化，而未考虑类内紧凑性。
- 可以修改 Softmax 函数以促进类内紧凑性，例如，angular softmax和additive-margin softmax。

- Let the bias  $b_j = 0$ ,
- Weight normalization  $\|\mathbf{w}_j\| = 1$
- Feature normalization  $\|\mathbf{x}_i\| = s$
- The cross-entropy loss becomes:

$$\left. \begin{array}{l} \bullet \text{ Let the bias } b_j = 0, \\ \bullet \text{ Weight normalization } \|\mathbf{w}_j\| = 1 \\ \bullet \text{ Feature normalization } \|\mathbf{x}_i\| = s \end{array} \right\} \mathbf{w}_j^T \mathbf{x}_i = \|\mathbf{w}_j\| \|\mathbf{x}_i\| \cdot \cos \theta_{y_i}$$

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cdot \cos \theta_{y_i})}{\sum_{c=1}^C \exp(s \cdot \cos \theta_{y_c})}$$



# 数据增广（了解）

## ■ 添加噪音和混响

- 使用增广数据集（例如：MUSAN等）和房间脉冲响应
- MUSAN数据集（来自12种语言的噪音、音乐、语音）

## ■ 语音编码

## ■ 说话人增强

- 音频速度扰动
- 使用速度因子 $\alpha$ ， $x(\alpha^t)$ 的时间扭曲

# 第六章 声纹识别

## ■ 6.1 声纹识别概述

- 6.1.1 声纹识别的基本概念

- 6.1.2 声纹识别方法的回顾

- 6.1.3 声纹识别的典型应用

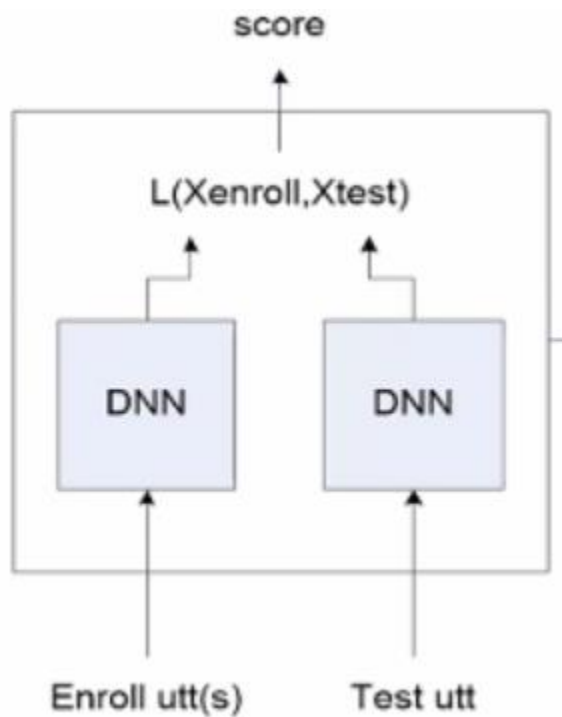
## ■ 6.2 传统声纹识别算法（GMM-UBM）

## ■ 6.3 基于深度学习的声纹识别算法

## ■ 6.4 声纹识别技术的展望（了解）

# 端到端声纹识别

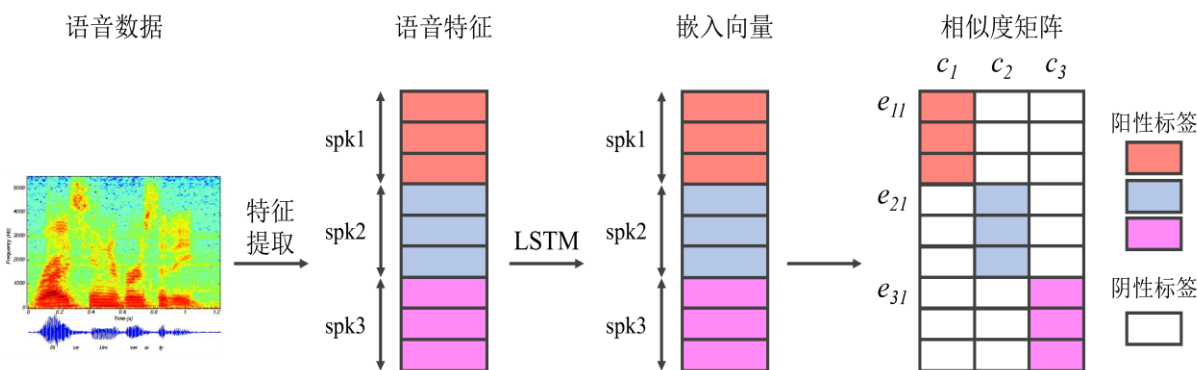
- 将语音输入模型，直接得到声纹识别结果。



# 端到端声纹识别（续）

## ■ Generalized End-to-End (GE2E)

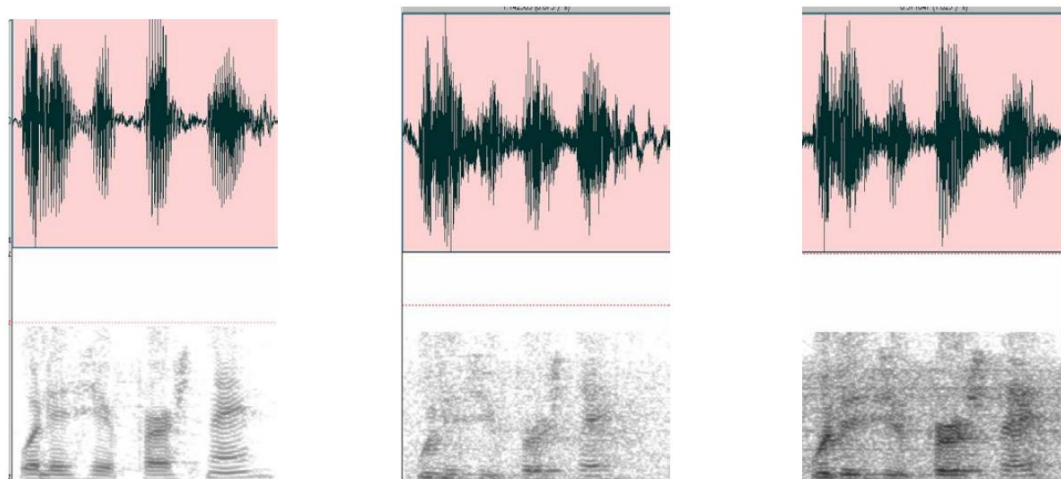
- 分别提取不同说话人的多条语音的嵌入向量；
- 计算不同嵌入向量间的相似度，并进行线性变换，来构造相似度矩阵。在相似度矩阵中，同说话人语音嵌入间的相似度要尽可能大，不同说话人语音嵌入间的相似度要尽可能小。以此为依据可以设计损失函数来优化模型参数。



基于GE2E的声纹识别模型示例

# 挑战 --- 跨信道

- 跨信道会严重降低声纹识别的性能
- 信道变量类型
  - 采录设备
  - 传输信道
  - 采录位置
  - .....

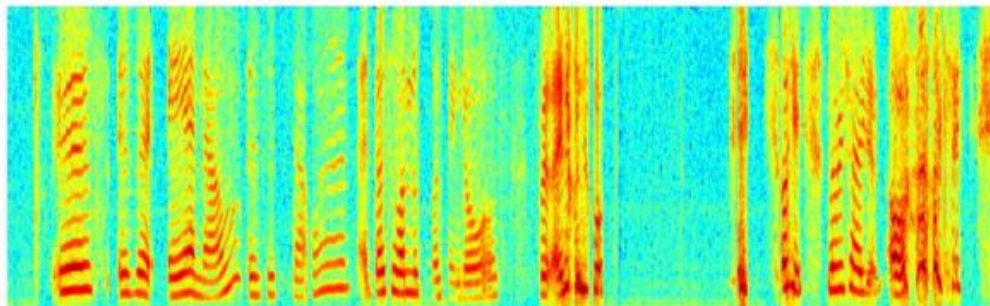


不同设备（相同原始语音）的语谱图

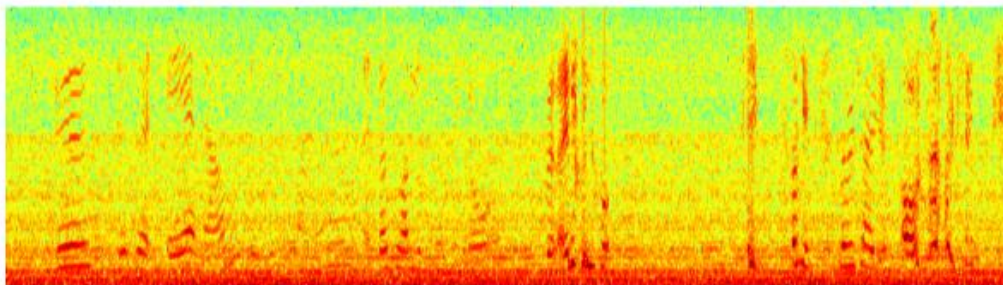
# 挑战 --- 复杂声学场景

- 加性噪声
  - ☐ 音乐
  - ☐ 杂音
  - ☐ 白噪声
  - ☐ .....
- 混响

Original utterance

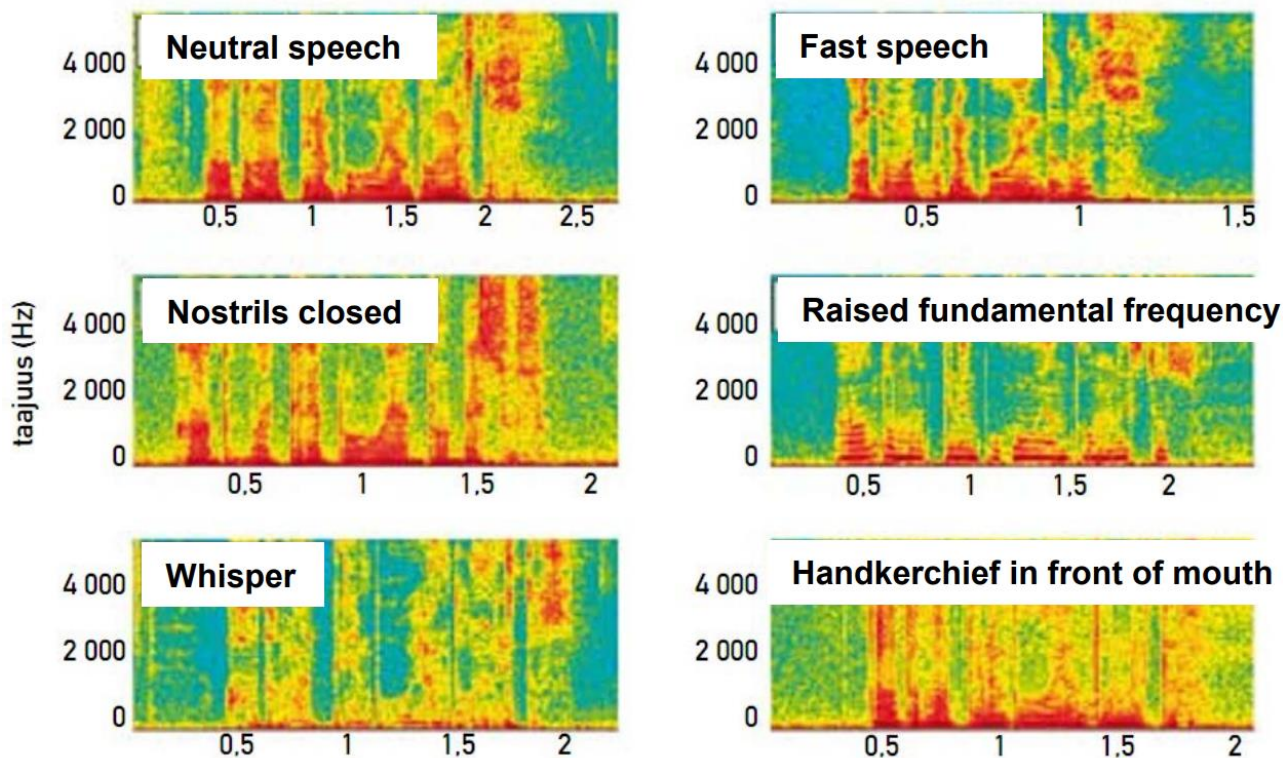


Simulated additive noise, signal-to-noise-ratio (SNR) = 6 dB



# 挑战 --- 时变及说话风格

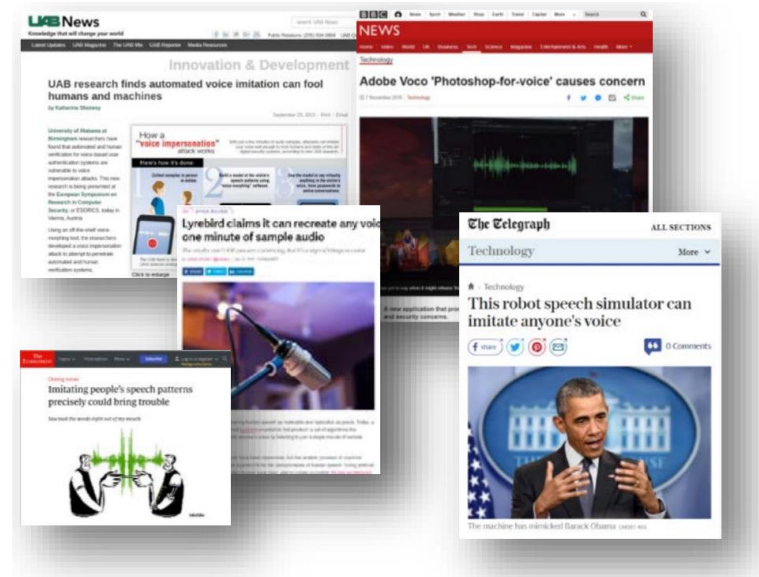
- 相同的说话人和相同的内容，由于说话风格、健康状况、发声器官等变化可能产生高度不同的声学特征。





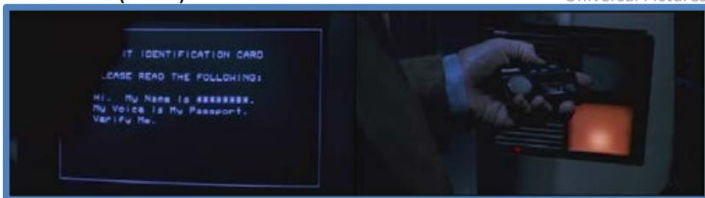
# 挑战 --- 声纹欺诈攻击

- 语音合成
- 语音转换
- 录音重放
- .....



*Sneakers (1992)*

Universal Pictures



Replay device



Device under attack



# 声纹识别前沿研究 (1/3)

## 2023年“声纹识别研究与应用学术研讨会”日程安排

大会主席： 天津大学教授 王龙标  
              科院声学所研究员 张鹏远

2023年10月21日 周六8:30-18:00

08:30-09:00 登记报道 地点:天津海河假日酒店5层宴会厅

09:00-09:10 开幕式 地点:天津海河假日酒店5层宴会厅

Session 1 主持人：王龙标

09:10-09:40 主旨报告 报告人：党建武 天津大学 教授  
慧言科技（天津）有限公司 首席科学家  
主题：说话人语音特征的编解码与识别

09:40-10:10 主旨报告 报告人：山世光 中国科学院计算技术研究所  
主题：视觉讲话人检测与唇语识别研究

10:10-10:30 邀请报告 报告人：李明 昆山杜克大学 副教授  
主题：时变与非言语语音声纹识别

# 声纹识别前沿研究 (2/3)

## Session 2 主持人：李明

10:40–11:00 邀请报告 报告人：何亮 清华大学副研究员  
新疆大学教授  
主题：稳定学习与可解释的声纹识别

---

11:00–11:20 邀请报告 报告人：张鹏远 中国科学院声学研究所研究员  
主题：基于一致性的伪造语音检测技术

---

11:20–11:40 邀请报告 报告人：王东 清华大学副研究员  
主题：Target speech extraction:  
Attributing to Speaker or Content

---

11:40–12:00 邀请报告 报告人：杜俊 中国科学技术大学副教授  
主题：多设备多场景远场说话人日志研究

---

## Session 3 主持人：洪青阳、王晓宝

13:30–13:50 邀请报告 报告人：钱彦旻 上海交通大学教授  
主题：Build a Strong Speaker Identification System:  
Lessons from SV Challenges

---

13:50–14:10 邀请报告 报告人：张晓雷 西北工业大学教授  
主题：噪声与对抗环境下的鲁棒声纹识别

---

14:10–14:30 邀请报告 报告人：谢磊 西北工业大学教授  
主题：说话人识别对抗攻击与说话人匿名化前沿进展

---

# 声纹识别前沿研究 (3/3)

## Session 4 主持人：张晓雷

15:30-15:50 邀请报告 报告人：洪青阳 厦门大学副教授  
主题：基于图结构建模的说话人日志研究

---

15:50-16:10 邀请报告 报告人：香港中文大学（深圳）副教授 武执政  
主题：基于深度学习的物理攻击模拟方法研究

---

16:10-16:30 邀请报告 报告人：易江燕 中科院自动化所副研究员  
主题：面向跨域数据的生成语音检测方法

---

## Session 5

16:30-17:30 主持人：清华大学 何亮  
专题讨论 声纹识别的机遇与挑战  
特邀学术界嘉宾：谢磊、李明、洪青阳、钱彦  
工业界嘉宾：穆向禹 王宇光

# 期末考试

时间： 11月19日10:00-12:00

地点： 46教A209

方式： 闭卷