声纹识别

洪青阳 副教授

厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

声纹识别

- 什么是“声纹识别”
  - 声纹识别（又称说话人识别，国外也叫Speaker ID），就是从某段语音中识别出说话人的身份的过程。
  - 与指纹类似，每个人说话过程中蕴涵的语音特征和发音习惯等也几乎是唯一的。

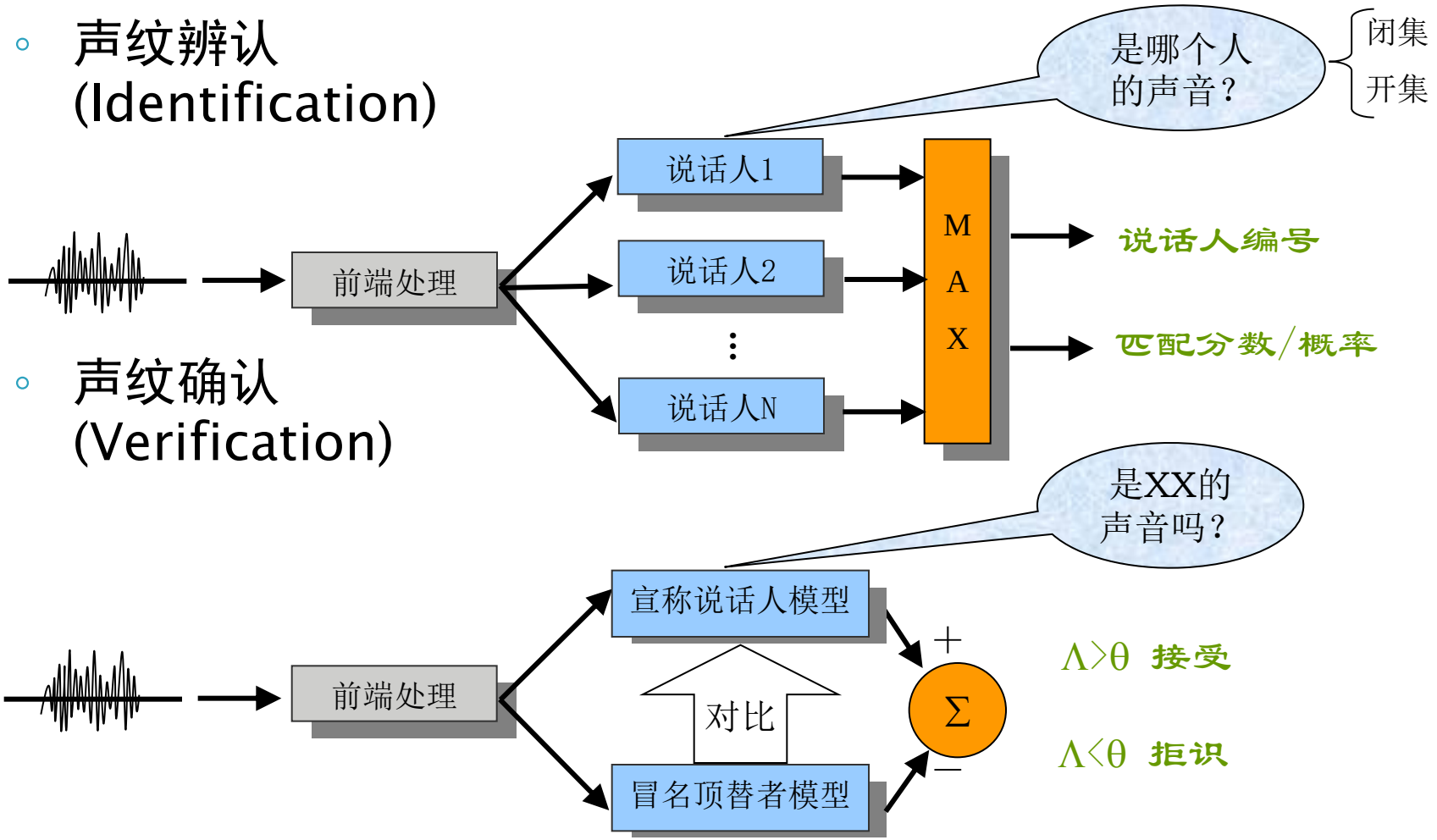
- 与“语音识别”的不同
  - “语音识别”是共性识别，判定所说的内容(说的什么)。
  - “说话人识别”是个性识别，判定说话人身份(是谁说的)。

声纹识别独特优势

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

分类方式一

- 按识别任务分类
  - 声纹辨认 (Identification)
  - 声纹确认 (Verification)



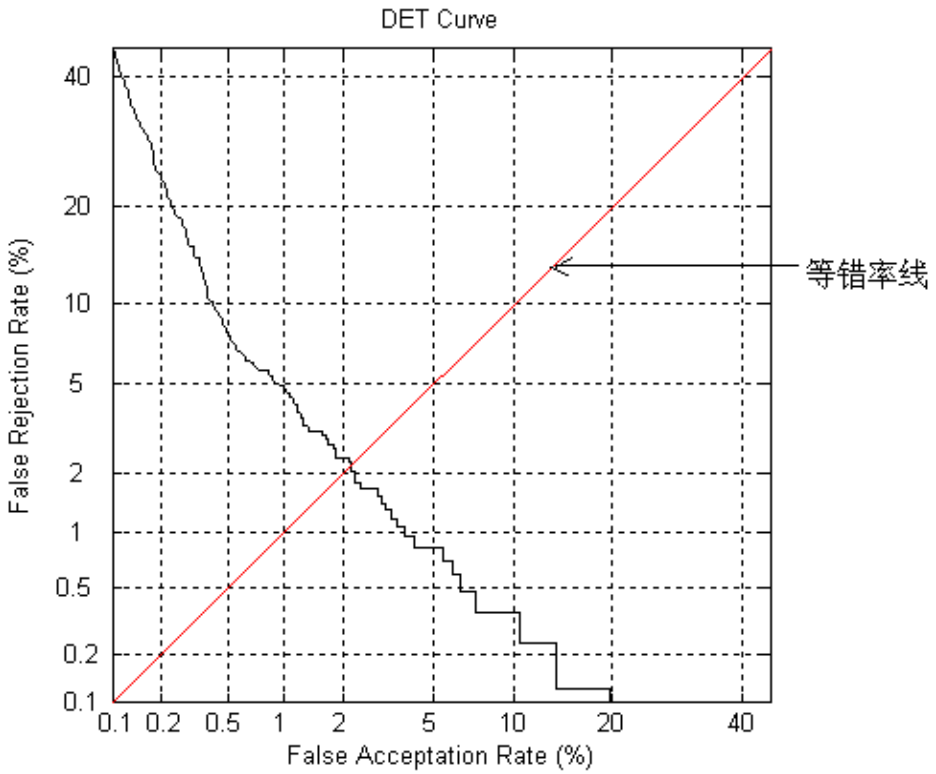
分类方式二

- 按说话内容分类
  - 文本无关（Text-Independent）
    - 不限定说什么文本
    - 语种无关（Language-Independent）
语种相关（Language-Dependent）
  - 文本相关（Text-Dependent）
    - 要求说特定的文本（与训练阶段一致，或现场提示）
    - 必定是语种相关的

性能评价标准

- 对于说话人辨认系统，其性能的评价标准主要是正确识别率。
- 对于说话人确认(SV)系统，其最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“认假率”，两者均与阈值的设定相关。
- 等错率(EER)：FRR与FAR相等。

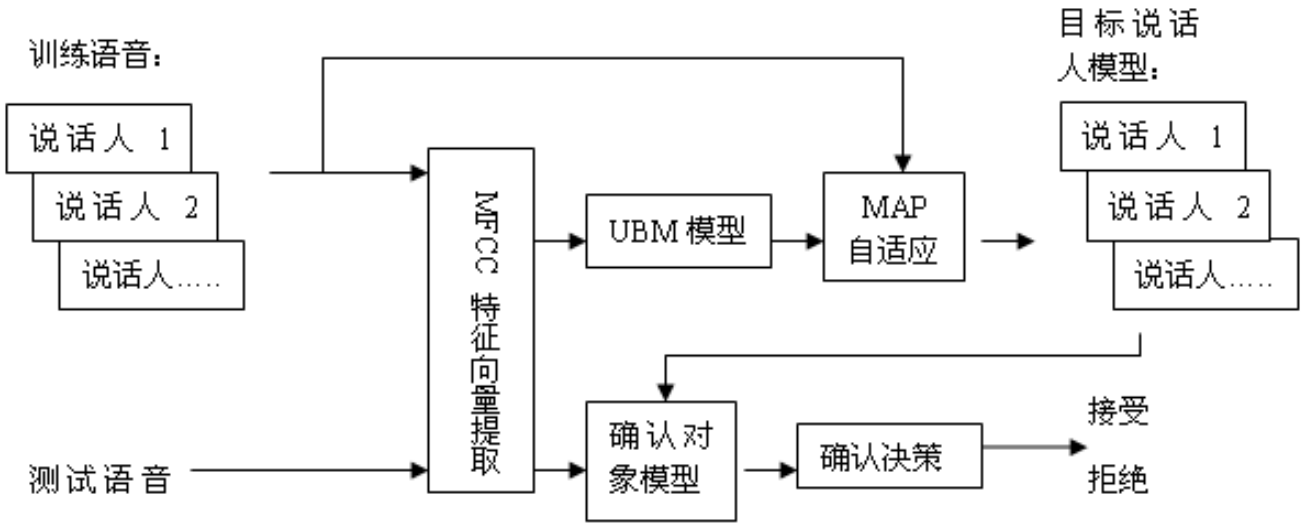
DET曲线图


DET Curve
False Rejection Rate (%)
False Acceptation Rate (%)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
等错率线

声纹建模方法

类型	主要算法
文本无关	GMM-UBM (D.A. Reynolds, 2000) GMM-SVM (W.M. Campbell, 2006) JFA(联合因子分析, P. Kenny, 2007) i-vector/PLDA (N. Dehak, 2011) DNN i-vector (Y. Lei, 2014) Deep Embedding (2017)
文本相关	GMM-UBM HMM-UBM TMM-UBM (TMM - Tied Mixture Model) i-vector DNN-ivector

经典方法(GMM-UBM)

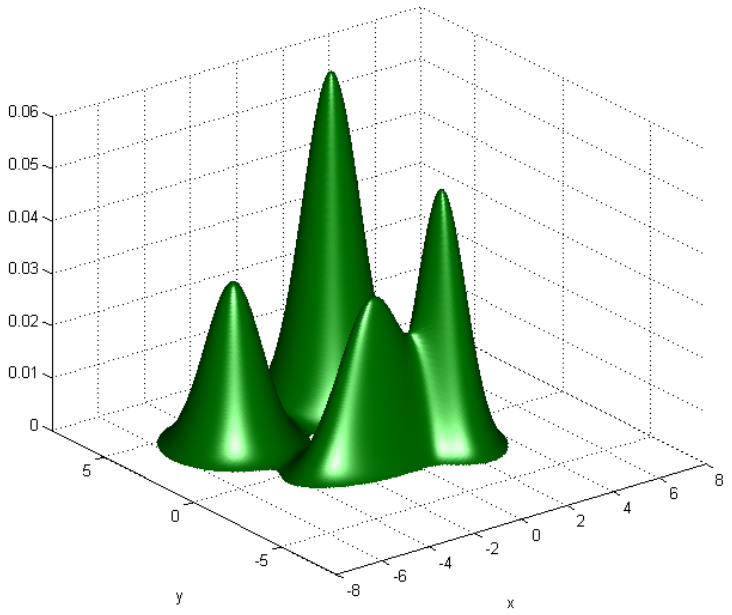


GMM-UBM说话人确认系统

➢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也是一个GMM，只是这个GMM需要用大量的不同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。



MAP自适应训练

- 只适应均值

$$\hat{\mu}_c = \frac{\gamma_c}{\gamma_c + r} E_c(\boldsymbol{X}) + \frac{r}{\gamma_c + r} \mu_c$$

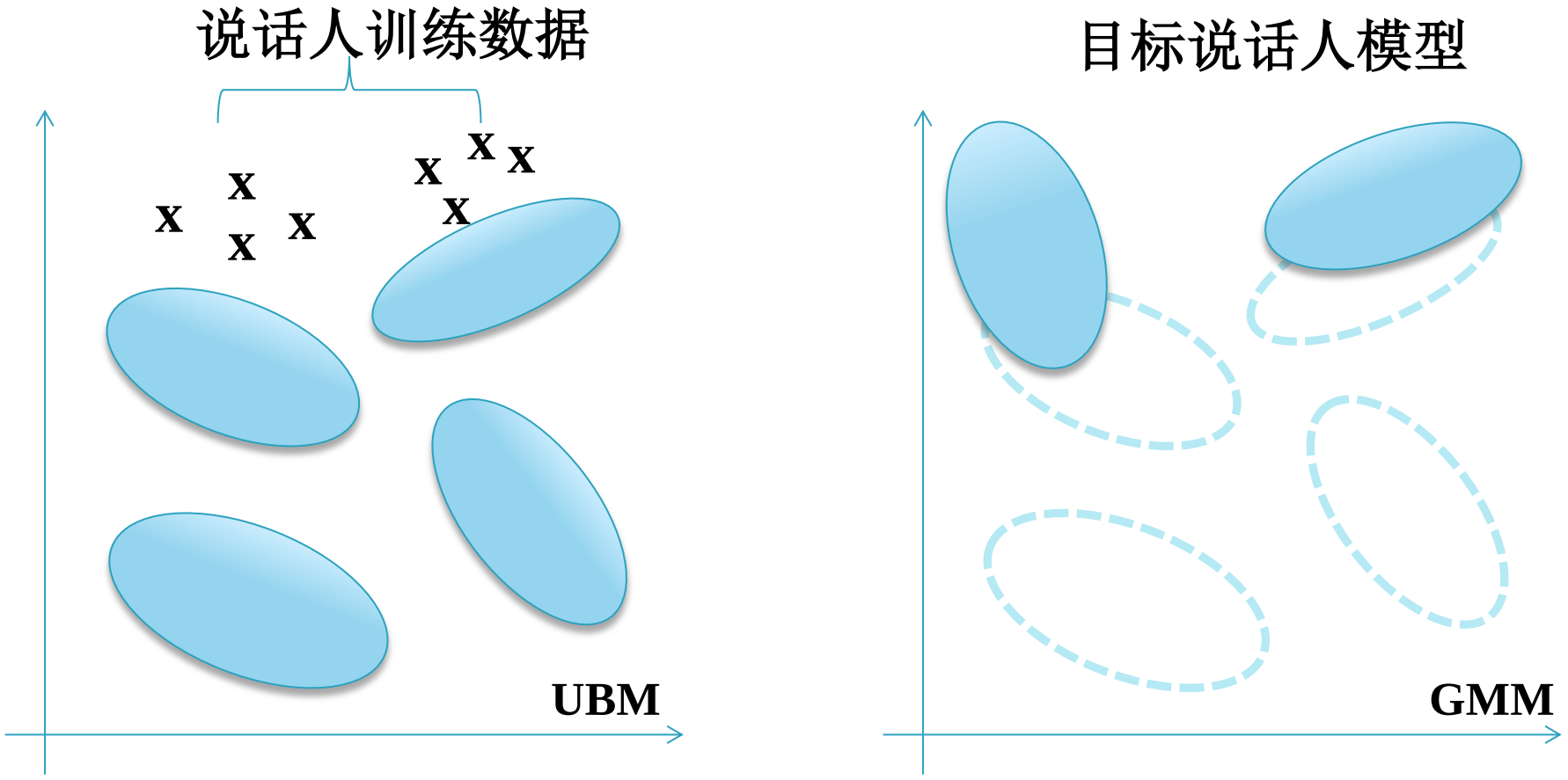
- 其中$\boldsymbol{X}$是目标说话人的特征向量集合(T帧)，μ_c是UBM的第c个高斯均值。r是个相关系数，对所有高斯都一样，是个常量，文本无关系统一般设为16。γ_c和$E_c(\boldsymbol{X})$是用EM算法得到的期望：

$$\gamma_c = \sum_{t=1}^{T} \gamma_t(c)$$

第c个高斯对特征向量x_t的后验概率

$$E_c(\boldsymbol{X}) = \frac{1}{\gamma_c} \sum_{t=1}^{T} \gamma_t(c)\, x_t$$

UBM=>GMM



GMM优缺点

GMM优点：

- 概率统计模型，通过大量训练语音数据集的统计分布进行描述，可较好地刻画目标话者不同情况下的特点，具有良好的鲁棒性。
- 同信道效果很好，已可实用。

GMM缺点：

- 有限的数据不一定能充分代表说话人的真实特征分布；只考虑某一类的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之间的区分性。
- 跨信道性能急剧下降！

i-vector

- i-vector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。

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

$$\lambda_c = \{w_c, \mu_c, \sigma_c^2\}, c = 1,2,\cdots,C$$

- 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：

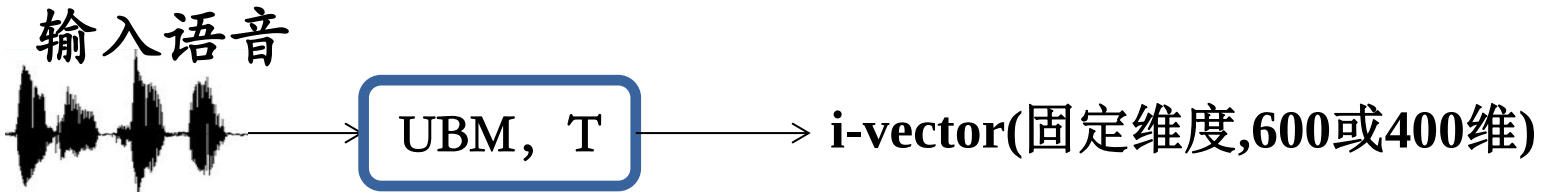
$$N_c = \sum_{t=1}^{T} P(c|o_t, \lambda_c)$$

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

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。

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

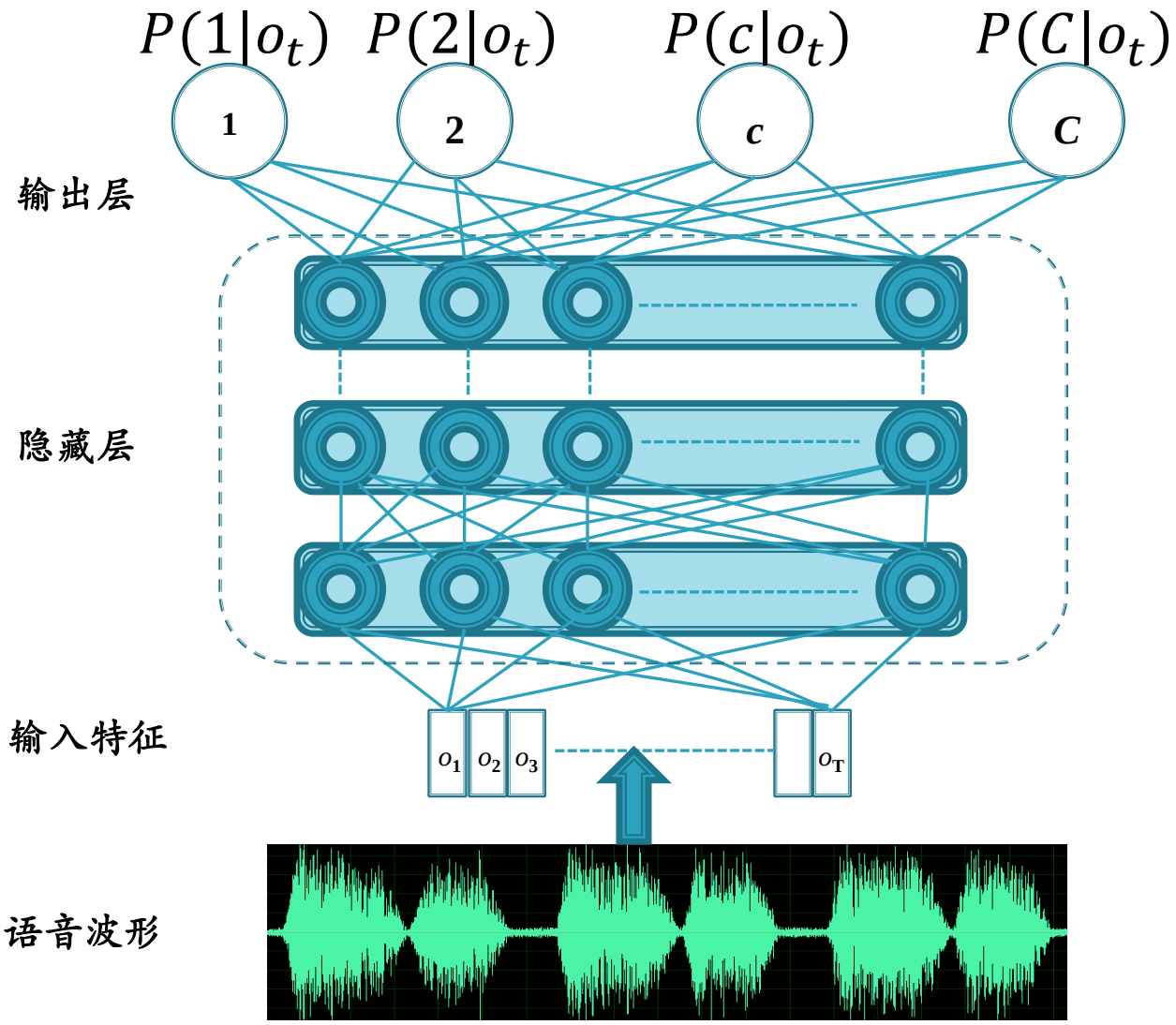


DNN/i-vector

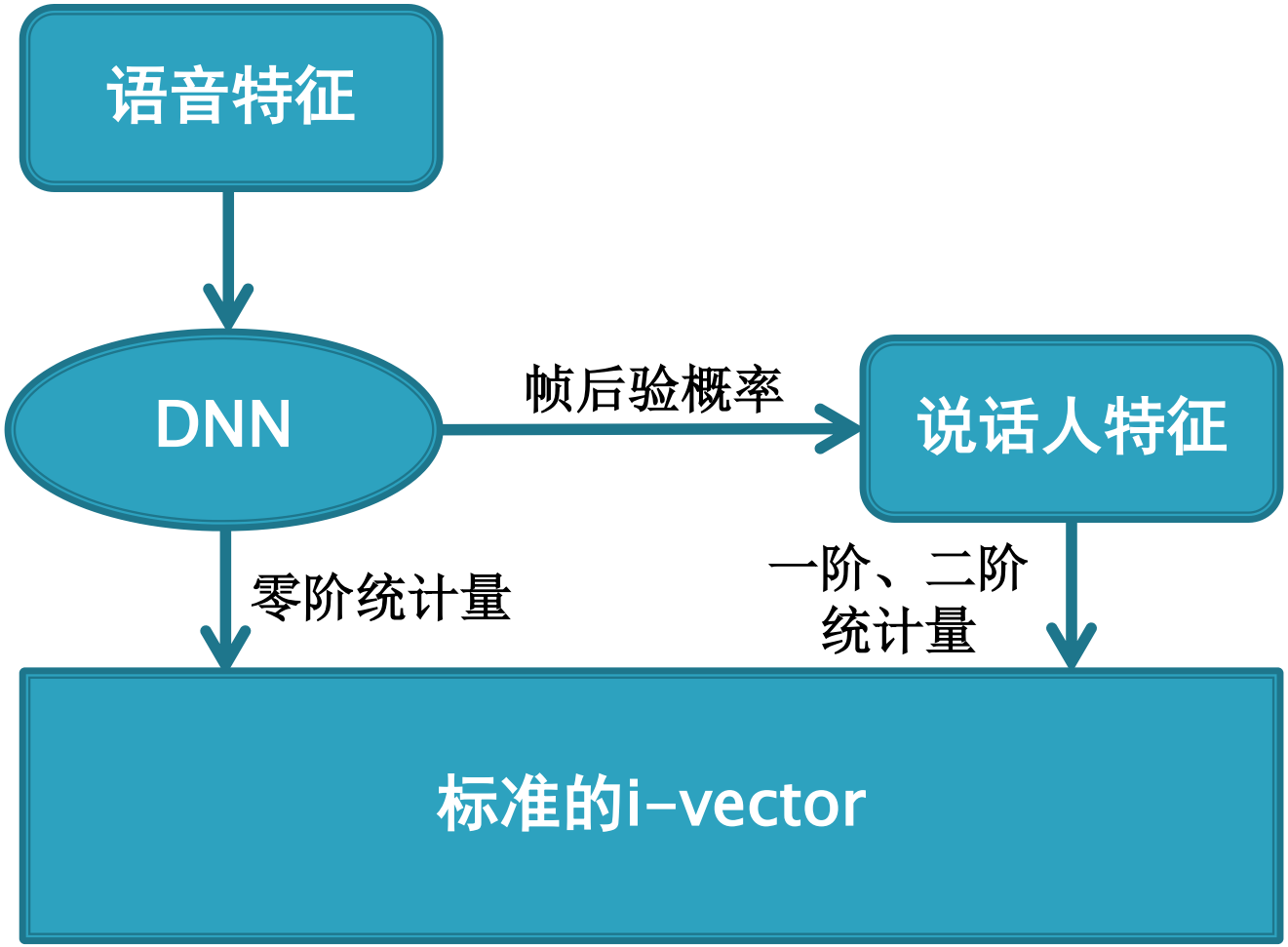
- UBM的作用
  - 包含C个高斯，用来做分类器；
  - 给出每一帧对每个高斯的后验概率；
  - UBM分类器缺乏语义信息(音素区分)；

- 类似语音识别GMM-HMM=>DNN-HMM，说话人识别的UBM也可替换为DNN。

DNN帧后验概率输出



DNN i-vector

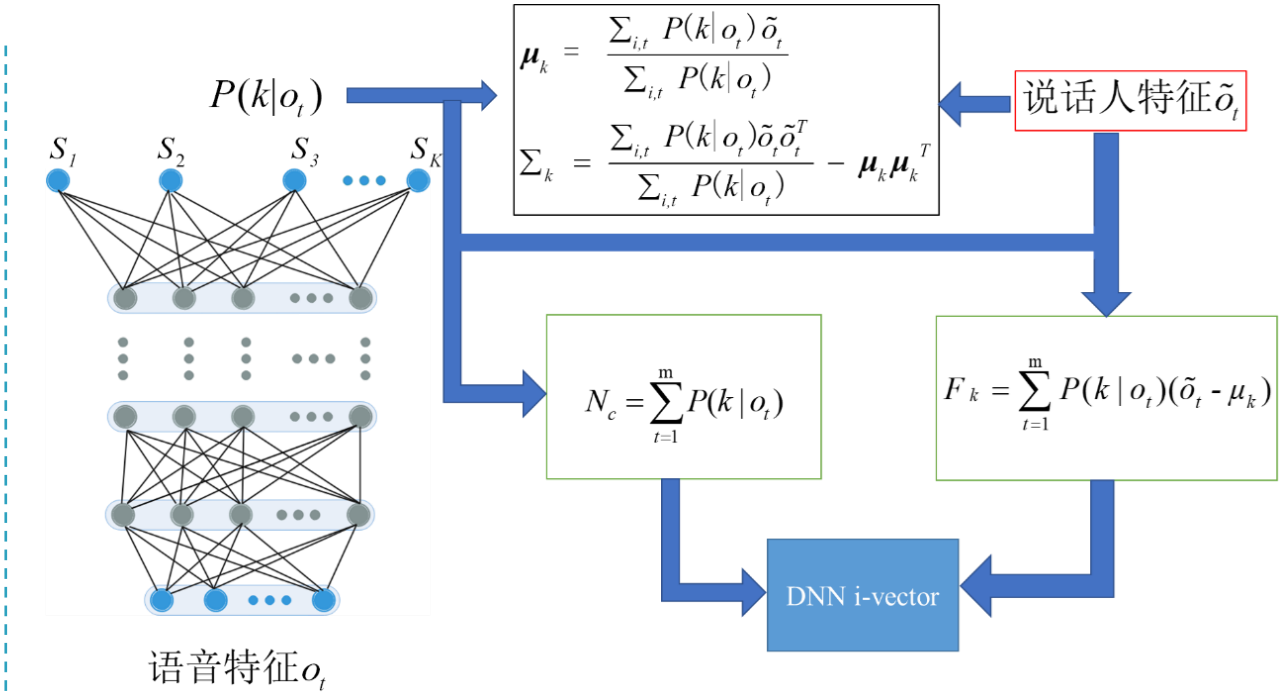


DNN i-vector

$$N_k = \sum_{t=1}^{m} P(k \mid o_t)$$

$$F_k = \sum_{t=1}^{m} P(k \mid o_t)(o_t - u_k)$$

传统 i-vector统计量的估计



DNN i-vector模型框架

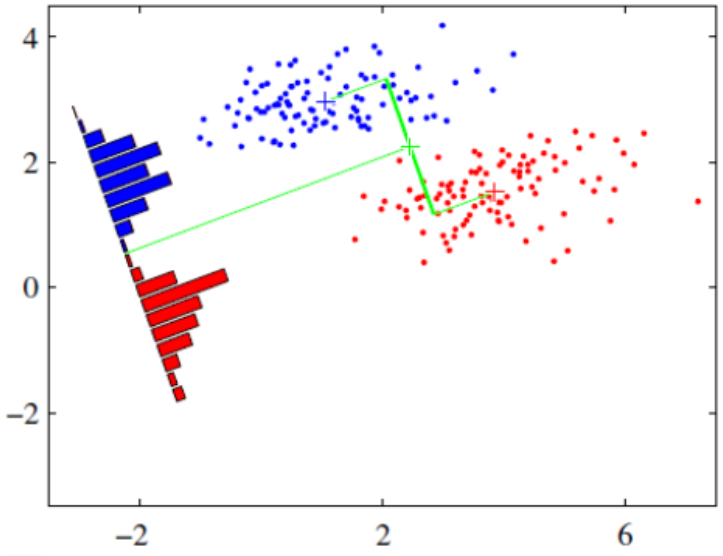
Cosine距离

- 在计算得到i-vector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。

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

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。

PLDA因子分析

PLDA对i-vector做进一步的因子分析，具体表示如以下公式：

$$x_{ij} = \mu + \phi\beta_i + Gw_{ij} + \varepsilon_{ij}$$

其中x_{ij}是第i个说话人的第j个i-vector，μ是所有训练i-vector的均值，ϕ是说话人空间矩阵(EigenVoice)，用于描述说话人的特征，β_i是符合$N(0,I)$分布的说话人因子，G是信道空间矩阵(EigenChannel)，用于描述信道的特征，w_{ij}是符合$N(0,I)$分布的信道因子，ε_{ij}是残差因子。

在实际应用中，往往把后面两项，即信道因子和残差因子$(Gw_{ij} + \varepsilon_{ij})$合并，简化为以下公式：

$$x_{ij} = \mu + \phi\beta_i + \varepsilon_{ij}$$

其中，ε满足高斯分布$N(0,\Sigma)$，β满足高斯分布$N(0,1)$。**要估计PLDA模型的参数，实际就是根据训练数据去估计参数ϕ和Σ。**

PLDA参数估计过程

首先，我们根据x_{ij}来定义一个统计量，如下：

$$\begin{cases} \widetilde{x_i} = \sum_{j=1}^{M_i} x_{ij} \ ; (sufficient\ stastics\ for\ ith\ speaker) \\ F_i = \frac{\widetilde{x_i}}{M_i} \ ; \quad F_i \sim N(\phi\beta_i, \frac{\Sigma}{M_i}) \end{cases}$$

F_i的分布用概率表示即$\mathrm{P}(F_i|\beta_i) = \mathrm{N}(\phi\beta_i, \frac{\Sigma}{M_i})$

根据贝叶斯法则，有：

$$\mathrm{P}(\beta_i|F_i) = \frac{P(F_i|\beta_i)P(\beta_i)}{P(F_i)}$$

即：

$$\mathrm{P}(\beta_i|F_i) \propto P(F_i|\beta_i)P(\beta_i)$$

由于右边的两项都是高斯分布，故左边的一项一定满足高斯分布。

PLDA参数估计过程

E-step：

求在给定观测数据和当前参数下对未观测数据β_i的条件概率分布$\mathrm{P}(\beta_i|F_i)$的期望$E(\beta_i|F_i)$（简写为$E(\beta_i)$），即上式中高斯分布的均值。即：

$$E(\beta_i) = [I + \phi^T \Sigma^{-1} M_i \phi]^{-1} \phi^T \Sigma^{-1} M_i F_i$$

又由期望相关公式可以得到：

$$E\left(\beta_i \beta_i^T\right) = E(\beta_i) E\left(\beta_i^T\right) + [I + \phi^T \Sigma^{-1} M_i \phi]^{-1}$$

M-step：

在这里，根据最大似然估计的原理，我们要最大化的是$\prod_{i=1}^{N} \prod_{j=1}^{M_i} P(x_{ij}, \beta_i)$，取对数，有：

$$\max \sum_{i=1}^{N} \sum_{j=1}^{M_i} \log[\mathrm{P}(x_{ij}|\beta_i)\mathrm{P}(\beta_i)]$$

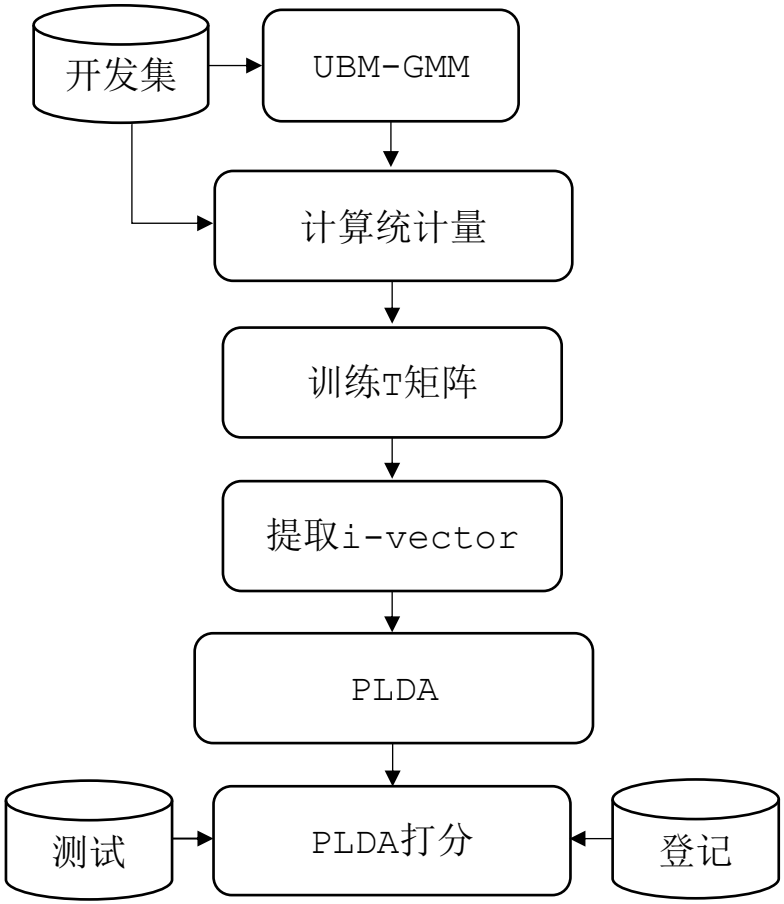
PLDA参数估计过程

重估计公式

$$\phi = \left(\sum_{i=1}^{N} M_i F_i E(\beta_i)^T \right) \left(\sum_{i=1}^{N} M_i E(\beta_i \beta_i^T) \right)^{-1}$$

$$\Sigma = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} [x_{ij}(x_{ij}^T - E(\beta_i)^T \phi^T)]}{\sum_{i=1}^{N} M_i}$$

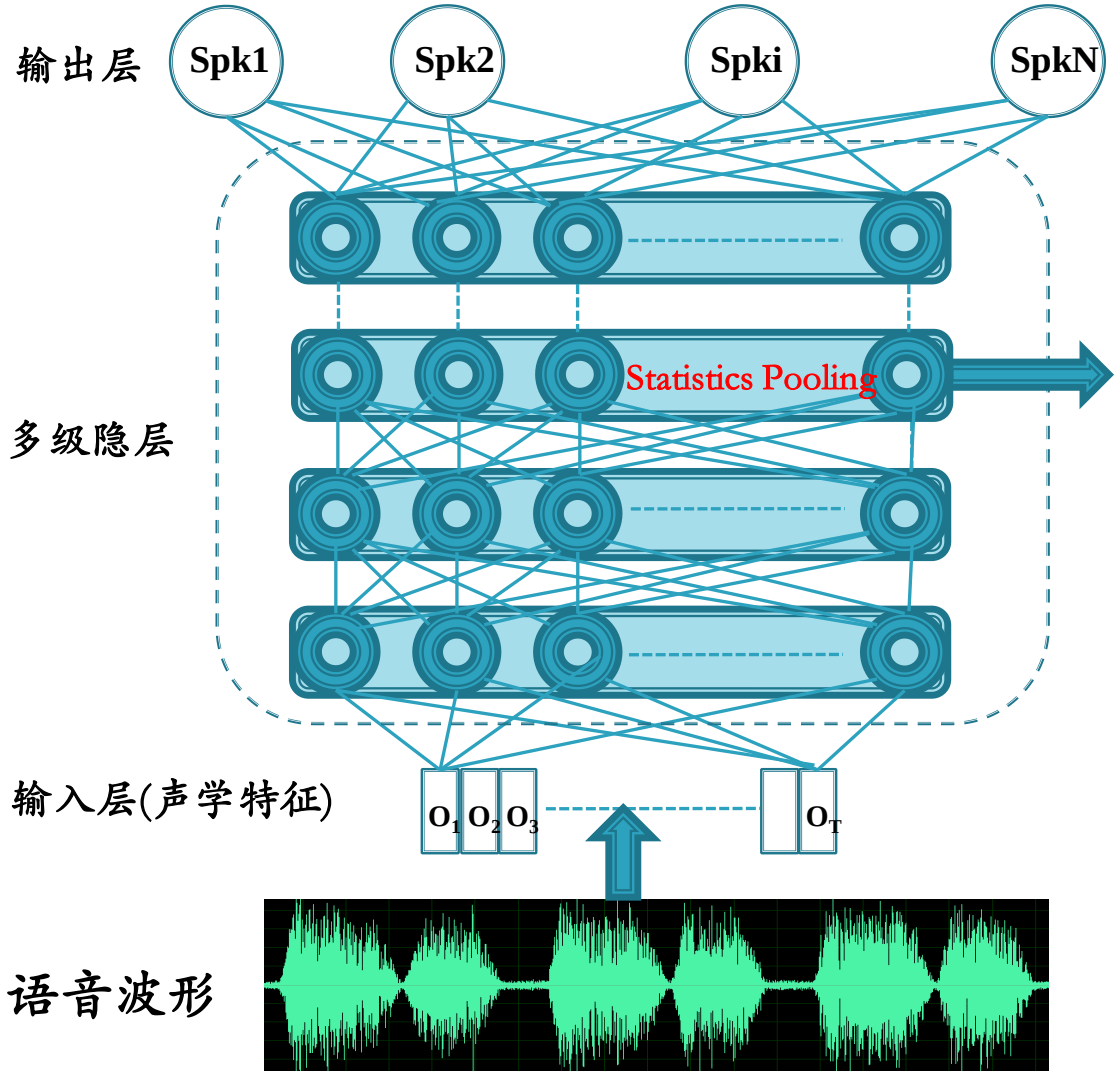
i-vector/PLDA系统



Deep Embedding

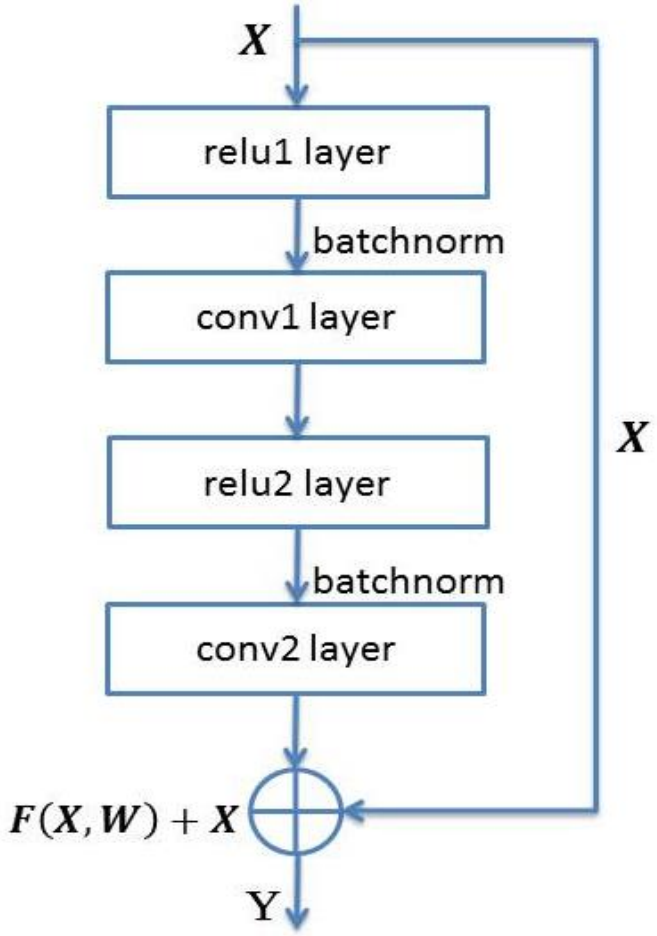
- 采用大数据，模拟噪声和远场环境
- 输出标签与说话人ID对应
- 基于深度模型，提取更有效的说话人特征—Deep Embedding (x-vector)

Deep Embedding

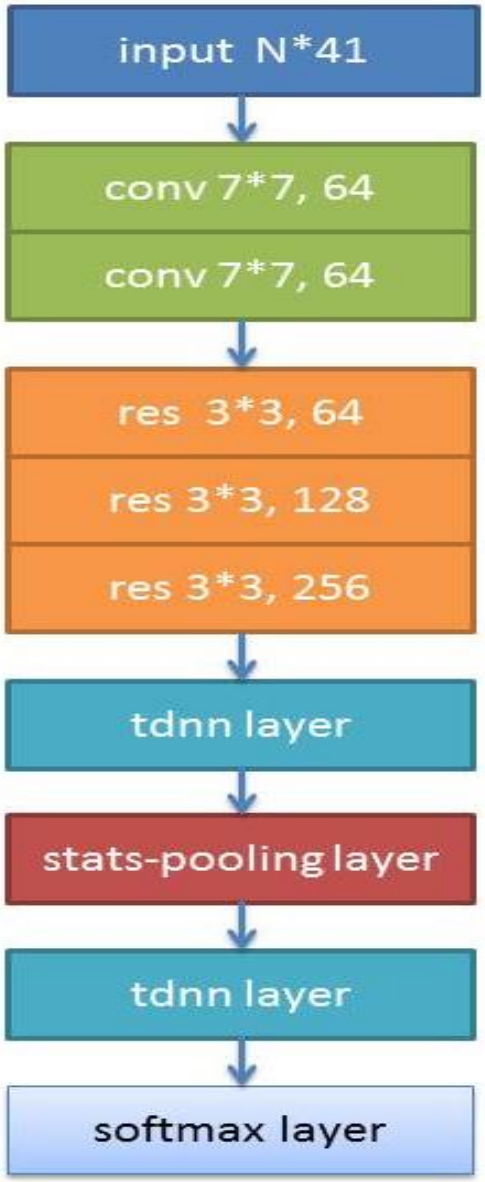


Statistics Pooling:
不同帧数的输入特征在这里聚合平均，形成段级别的特征，输出后再用LDA进行降维。

残差网络ResNet


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

融合CNN、TDNN和ResNet的网络



总结及展望

- **总结**
  - 经典模型：GMM-UBM
  - 技术难题：跨信道、噪声、短语音
  - 文本无关说话人识别主流技术：i-vector/PLDA
  - 深度学习：DNN i-vector, Deep Embedding (x-vector)

- **展望**
  - 更精准的声纹算法
  - 短语音条件下的可靠说话人识别
  - 噪声/远场环境鲁棒性

Thank you!

Any questions?