Ivector

ivector是由Dehak，Kenny等人在2010年提出的模型，它将说话人空间与信道空间用一个包含了说话人信息以及信道信息的全差异空间T表示，将说话人模型映射到一个低维向量w，克服了GMM-UBM模型中各高斯分量互相独立的局限性，而且克服了JFA模型对语音数据依赖性大的缺点。截至今日，i-vector在大多数情况下仍然是文本无关声纹识别中表现性能最好的建模框架。

大致过程：

1，预处理：MFCC 提取，去静音 VAD 等等
2，训练GMM-UBM，训练出每个高斯分量的参数（主要是均值，权重和方差对于识别性能并未获得显著提高 ），当中是使用EM算法。

3，训练i-vector模型

4，信道补偿 （原因：总变化空间中，同时含说话人信息与信道信息 ），线性判别分析(LDA) 和类内协方差规整(WCCN) ，提升类间的差异。扰动属性投影（NAP）削 弱说话人特征空间中的信道子空间的分量来提高说话人之间 的“距离”，从而提高系统的识别性能。

5，测试：把测试说话人与目 标说话人的 i - vector 的余弦距离作为得分，或PLDA 得分 .

PS：在文本无关识别表现很好，但文本相关识别不如传统的GMM-UBM

3.3 i – vector 模型

以 JFA 为基础，Dehak 和 Kenny 提出了一种更为简化的基于因子分析（Factor analysis，FA）的说话人识别方法，称为 i – vector 模型。其中，i 是身份（Identity）的缩写，故 i – vector 相当于说话人的身份标识。i – vector 和 JFA 一样，也是一种基于统计特性的语音特征，衍生于 GMM 均值超矢量[15]，但却弥补了 JFA 对于语音数据依赖性大的缺点。

i – vector 模型不像 JFA 一样将均值超矢量空间划分两部分，而是用一个总变化空间（Total variability space）进行了代替。在这个总变化空间中，包含了说话人的语音信息以及信道信息。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，其中 M 是由所有的 C 个 GMM 均值矢量按照先后顺序串联在一起得到的。对于一段给定的语音，GMM 均值超矢量 M 定义如下：

$$M = m + Tw \tag{10}$$

其中，m 为 UBM 均值超矢量，T 为总变化空间矩阵（$CF \times R$），F 为 MFCC 特征向量的维数，w 为 i – vector。w 是一个 R 维的特征向量（$400 \leqslant R \leqslant 600$），并且服从标准高斯分布 $N(0, I)$；GMM 均值超矢量 M 服从高斯分布 $N(\mathrm{m}, TT^{*})$。

设说话人的一组特征序列为 $X(x_1, \cdots, x_t, \cdots, x_T)$，对每一个时刻 t，特征矢量 x_t 相对每个高斯分量 c 的状态占有率为：

$$\gamma_t(c) = \frac{a_c P(x_t \mid \mu_c, \Sigma_c)}{\sum_{i=1}^{C} a_i P(x_t \mid \mu_i, \Sigma_i)} \tag{11}$$

式中，$\gamma_t(c)$ 为语音 x_t 在 UBM 的第 c 个高斯分量上的后验概率。

利用 $\gamma_t(c)$ 可以求出每个说话人的权值和均值矢量对应的 Baum — Welch 统计量：

$$\begin{aligned} N_c &= \sum_{t=1}^{T} \gamma_t(c) \\ F_c &= \sum_{t=1}^{T} \gamma_t(c)\, x_t \end{aligned} \tag{12}$$

定义为 $\tilde{F}_c$ 一阶中心统计量：

$$\tilde{F}_c = \sum_{t=1}^{T} \gamma_t(c)(x_t - \mu_c) = F_c - N_c\mu_c \quad (13)$$

对每个说话人，令 $L = I + T^T \Sigma^{-1} N(X) T$，$w$ 的后验分布服从高斯分布 $N(L^{-1}T\Sigma^{-1}\tilde{F}(X), L^{-1})$ [16]，其中 $\tilde{F}(X)$ 为从 $\tilde{F}_c$ 拼接而来的 $CF \times 1$ 维的超矢量，$N(X)$ 为以 N_c 为对角块的 $CF \times CF$ 维对角矩阵。w 的后验均值用期望的形式可以表示为：

$$E(w) = L^{-1}T\tilde{F}(X) \quad (14)$$

由于总变化空间中，同时含说话人信息与信道信息，所以需要对上述过程中提取的初始 i-vector 做信道补偿。信道补偿技术有线性判别分析（Linear discriminant analysis，LDA）和类内协方差规整（Within-Class covariance cormalization，WCCN）[12] 等。

在 i-vector 说话人识别的测试阶段，把测试说话人与目标说话人的 i-vector 的余弦距离作为得分，若得分大于阈值则接受说话人，反之则拒绝。

（a_c， μ_c 和 Σ_c 分别表示第c个高斯分量的权重、均值和方差。

$P_c(x_t)$为每个x_t 在高斯分量c上的隐含类别的概率。 一阶统计量是数学期望）

T 与 w 的具体训练过程：

Ivec.py 中用 sidekit.FactorAnalyser().total_variability()来训练 T，在 sidekit 的相关源码中训练过程如下：

1. 先将均值向量和残差的协方差矩阵初始化为 0。
2. 再迭代地估计全变化空间矩阵：先将估计 factor analysis matrix 所需的三个参数_a,_c,_r 初始化为 0，再进行 E step：将每个 statserver 的统计量累加起来，通过运算得到_a,_c,_r，然后 M step：通过上一步得到的_a,_c,_r 解线性方程组得到类间矩阵 F，迭代到 divergence 最小时更新 F。

sidekit.FactorAnalyser().extract_ivectors()来提取 ivector(即 w)，在 sidekit 的相关源码中训练过程如下：

1. 将 w 的零阶统计量初始化为值均为 1 的向量，一阶统计量和方差初始化为值均为 0 的矩阵，
2. 分 batch 进行循环，每次循环更新零阶统计量，一阶统计量，通过这些值可以计算出一阶中心统计量，结合之前求出的 T 进而算出 w 的后验均值和方差。