文本无关的声纹识别 验证

By Dake

Dake的专栏：www.glade.tk

一、声纹识别简介

声纹是指能惟一识别某人或某物的声音特征，是用电声学仪器显示的携带言语信息的声波频谱。虽然人的发音器官生理构造总的是相同的，但人的语言产生是人体语言中枢与发音器官之间一个复杂的生理物理过程，人在讲话时使用的器官——舌、牙齿、喉头、肺、鼻腔在尺寸和形态等方面，每两个人之间的差异会很大（见图2-1所示）所以任何两个人的声纹图谱都有差异，而对于每个人而言，从十几岁发育变声后直到五十多岁，其声纹基本保持不变。声纹识别技术正是利用这一特点，将声音输入到声谱仪中，使声音不同频率的机械振动变成频谱图像，显示在荧光屏或记录在纸上，这种图像就是声纹。

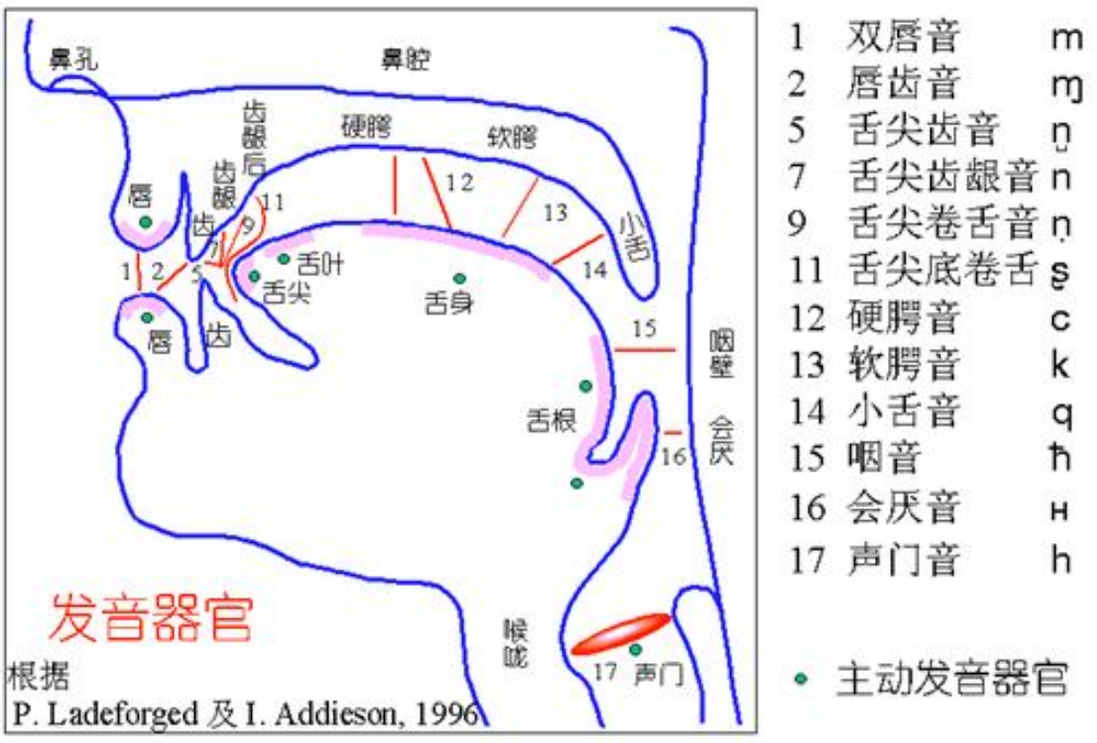

图2-1 发音器官

声纹识别(Voiceprint Recognition，即VPR)，通常也被称为话者识别(Speaker Recognition)，分为两类，即话者辨认(Speaker Identification)和话者确认(Speaker Verification)。前者用以判断某段语音是若干人中的哪一个所说的，是“多选一”问题；而后者用以确认某段语音是否是指定的某个人所说的，是“一对一判别”问题。不同的任务和应用会使用不同的声纹识别技术，如缩小刑侦范围时可能需要辨认技术，而银行交易时则需要确认技术。不管是辨认还是确认，都需要先对说话人的声纹进行建模，这就是所谓的“训练”或“学习”过程[7]。声纹识别过程如图2-2所示：

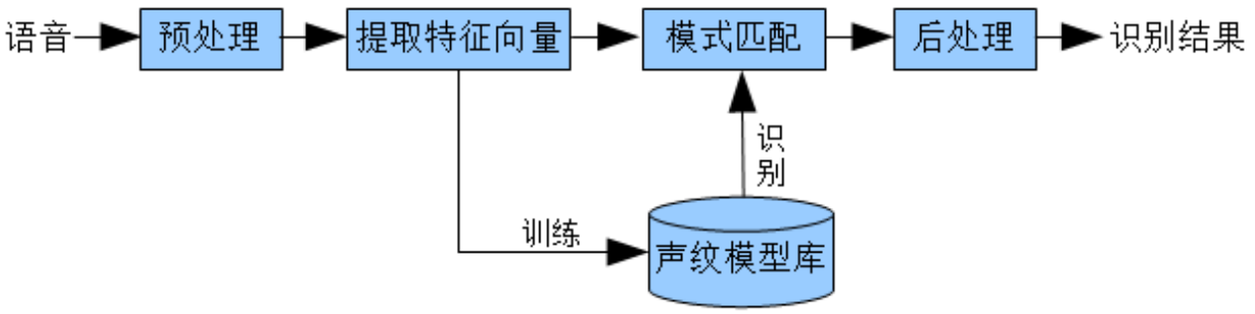

图2-2 声纹识别过程

声纹识别可以说有两个关键问题，一是特征提取，二是模式匹配。特征提取的任务是提取并选择对说话人的声纹具有可分性强、稳定性高等特性的声学或语言特征。与语音识别不同，声纹识别的特征必须是“个性化”特征，而说话人识别的特征对说话人来讲必须是“共性特征”。虽然目前大部分声纹识别系统用的都是声学层面的特征，但是表征一个人特点的特征应该是多层面的，包括： (1)与人类的发音机制的解剖学结构有关的声学特征(如频谱、倒频谱、共振峰、基音、反射系数等等)、鼻音、带深呼吸音、沙哑音、笑声等；(2)受社会经济状况、受教育水平、出生地等影响的语义、修辞、发音、言语习惯等；(3)个人特点或受父母影响的韵律、节奏、速度、语调、音量等特征。从利用数学方法可以建模的角度出发，声纹自动识别模型目前可以使用的特征包括：(1)声学特征(倒频谱)；(2)词法特征(说话人相关的词n-gram，音素n- gram)；(3)韵律特征(利用n-gram

描述的基音和能量“姿势”)；(4)语种、方言和口音信息；(5)通道信息(使用何种通道)；等等。

对于模式识别，主要有这几大类方法：(1)模板匹配方法：利用动态时间弯折(DTW)以对准训练和测试特征序列，主要用于固定词组的应用(通常为文本相关任务)；(2)最近邻方法：训练时保留所有特征矢量，识别时对每个矢量都找到训练矢量中最近的K个，据此进行识别，通常模型存储和相似计算的量都很大；(3)神经网络方法：有很多种形式，如多层感知、径向基函数(RBF)等，可以显式训练以区分说话人和其背景说话人，其训练量很大，且模型的可推广性不好；(4)隐式马尔可夫模型(HMM)方法：通常使用单状态的HMM，或高斯混合模型(GMM)，是比较流行的方法，效果比较好；(5)VQ聚类方法(如LBG，K-均值)：效果比较好，算法复杂度也不高，和HMM方法配合起来更可以收到更好的效果；(6)多项式分类器方法：有较高的精度，但模型存储和计算量都比较大。 其中模板匹配法的的要点是，在训练过程中从每个说话人的训练语句中提取相应的特征矢量来描述各个说话人的行为，在测试阶段，从说话人的测试语音信号中用同样的方法提取测试模板，主要有动态时间规整方法和矢量量化方法。

对说话人确认，还面临一个两难选择问题。通常，表征说话人确认系统性能的两个重要参数是错误拒绝率和错误接受率，前者是拒绝真正说话人而造成的错误，后者是接受集外说话人而造成的错误，二者与阈值的设定相关。在现有的技术水平下，两者无法同时达到最小，需要调整阈值来满足不同应用的需求，比如在需要“易用性”的情况下，可以让错误拒绝率低一些，此时错误接受率会增加，从而安全性降低；在对“安全性”要求高的情况下，可以让错误接受率低一些，此时错误拒绝率会增加，从而易用性降低。前者可以概括为“宁错勿漏”，而后者可以“宁漏勿错”。我们把真正阈值的调整称为“操作点”调整。好的系统应该允许对操作点的自由调整。

声纹识别有文本相关的(Text-Dependent)和文本无关的(Text-Independent)两种。与文本有关的声纹识别系统要求用户按照规定的

内容发音，每个人的声纹模型逐个被精确地建立，而识别时也必须按规定的内容发音，因此可以达到较好的识别效果，但系统需要用户配合，如果用户的发音与规定的内容不符合，则无法正确识别该用户。而与文本无关的识别系统则不规定说话人的发音内容，模型建立相对困难，但用户使用方便，可应用范围较宽。根据特定的任务和应用，两种是有不同的应用范围的。比如，在银行交易时可以使用文本相关的声纹识别，因为用户自己进行交易时是愿意配合的；而在刑侦或侦听应用中则无法使用文本相关的声纹识别，因为无法要求犯罪嫌疑人或被侦听的人配合。

二、MFCC参数（Mel倒谱系统）的提取说明

1、**预增强（Pre-Emphasis）**：差分语音信号。

2、**音框化（Framing）**：对语音数据分帧。

3、**汉明窗（Hamming Windowing）**：对每帧信号加窗，以减小吉布斯效应的影响。

4、**快速傅立叶变换（FFT）**：将时域信号变换成为信号的功率谱。

5、**三角带通滤波器（Triangle Filters）**：三角滤波器覆盖的范围都近似于人耳的一个临界带宽，以此来 模拟人耳的掩蔽效应。

6、**离散余弦转换（DCT）**：去除各维信号之间的相关性，将信号映射到低维空间。

三、声纹提取、识别过程

话者的声纹提取过程总的分4步：

1、 对输入的语音数据序列（PCM 码流）进行预处理。

目的：a) 去除非语音信号 和 静默语音信号；

b) 对语音信号分帧，以供后续处理。

2、 提取每一帧语音信号的MFCC 参数 并保存。

3、 用第2 步提取的 MFCC 参数训练话者的 GMM （高斯混合模型），得到专属某话者的 GMM 声纹模型。

4、 声纹识别。提供输入话音与GMM 声纹模型的匹配运算函数，以判断输入话音是否与声纹匹配。

一）、 语音数据预处理（去除静寂声音）

输入语音流采用单声道、8bit、16KHz采样。

以256个采样点为一个音框单位（帧），以128为音框之间的重迭单位，对输入语音流进行分帧。

计算各帧语音数据的累积能量E（最大值为256^3=16777216，用int表示足够），

$$E=\sum_{n=1}^{N}x^2(n), N=256$$

，

如果连续语音帧累积能量E大于预设静音阈值（连续数>100），则采纳该段连续语音帧为训练语音；

保留所有可供训练的语音。

二）、MFCC参数提取

图1.显示了MFCC参数提取流程

具体6步：

1) 预增强（Pre-Emphasis ）

2) 音框化（Framing ）

3) 汉明窗（Hamming Windowing ）

4) 快速傅立叶转换（FFT ）

5) 三角带通滤波器（Triangle Filters ）

6) 离散余弦转换（DCT ）

1） 预增强（Pre-Emphasis ）(对原始采样数据处理，所以N 不是256)

以S1(n)（n：0..N-1）表示时域信号，预增强公式为：

S(n) = S1(n) – a×S1(n-1) (0.9 < a < 1.0)----------*每字节做差分*

该过程可以达到在音框化阶段对静音数据的判断，因为静音数据的值是几乎不变的

所以在做差分以后值会很小，接近于0，而有声音的数据则会保留较大的值

S(n)=(S1(n)-128)/128

此时还是不分帧的好，这样就只需要做完帧数据大小一半的差分

差分后必须以short以及比它字节大的有符号类型，因为差分结果可能为负，且超过char的范围，造成溢出

2） 音框化（Framing ）

音框化即预处理阶段的语音信号分帧。

3） 汉明窗（Hamming Windowing ）

假设音框化的信号（M帧共N点）为S(n)，n=0，1，...，N-1。那么乘上汉明窗后为:

S′(n) = S(n) ×W(n)

$$w(n) = w(n,a) = (1-a) - a\cos(\frac{2\pi n}{N-1}), 0 \le n \le N-1$$

, a = 0.46

//即使是重叠处的样值在汉明窗以后也会不同，因为n不同

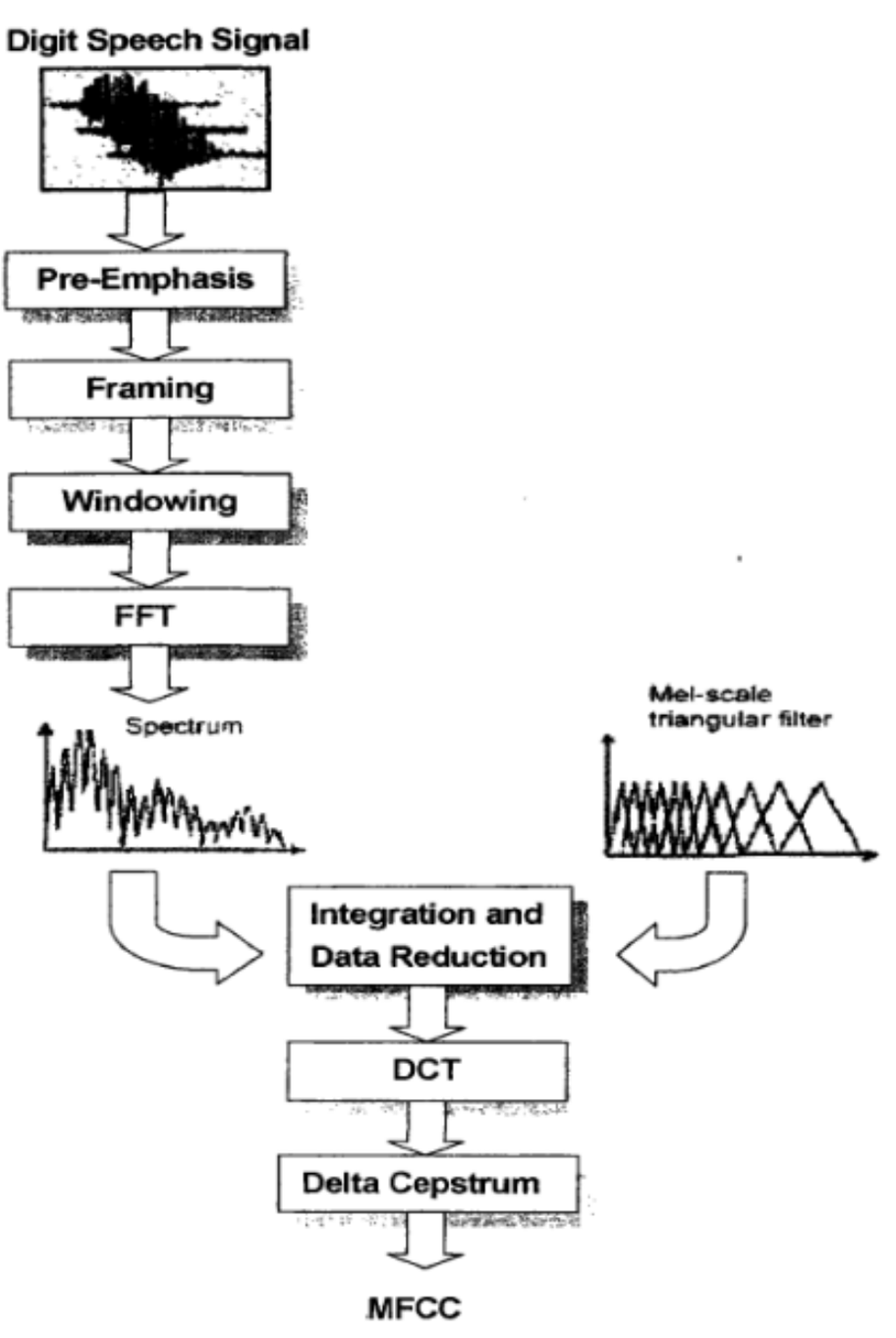

图1. MFCC参数提取流程

4） 快速傅立叶转换（FFT ）

对S′(n) 的每帧实施基2 FFT时域变换（代码到网上找）

得到X(n), n = 0..N-1（N=256）

注意X(n)为复数，所以S′(n) 也要在3）以后转化为复数作为FFT的输入

5） 三角带通滤波器（Triangle Filters ）

设定16KHz 和 8KHz条件下的滤波器数组 melf16[]，melf8[]。

melf[]滤波器数组为20×129的稀疏矩阵，以结构化数组的方式存储。

计算每个滤波器输出的对数能量z[20]，计算公式为：

z[ ] = log (melf[]*(|X(0:128)|.^2))----------log(m)是以e 为底m为真数的对数

同理exp(m)是以e为底m为指数的指数

melf[]数组见melf16，melf8

6） 离散余弦转换（DCT ）

对上一步所获得的对数能量进行DCT变换，获得DCT系数数组r[20]

r[] = dct (z[ ])；

dct()变换公式为

$$r[k]=\sum_{l=0}^{D-1} z[l]\cos\left(\frac{\pi(2l+1)k}{2D}\right)$$

，D = 20

r[] 即一帧语音信号的MFCC参数

计算并保存所有各帧语音信号的MFCC参数。

三）、 训练话者的GMM 模型

GMM模型主要公式为：

（1)

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

----------(1-1)

$\vec{x}$为D维随机矢量，与 r[ ] 对应；

$$b_i(\vec{x}), i=1,\ldots,M$$

是m组D维高斯概率密度函数；

$$b_i(\vec{x}), i=1,\ldots,M$$

是M组高斯向量的混合数，

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

。

（2）D维高斯概率密度函数公式

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

；

（3）一个话者的GMM模型由其参数组唯一表示

$$\lambda=\{p_i,u_i,\sum_i\},i=1,\ldots,M$$

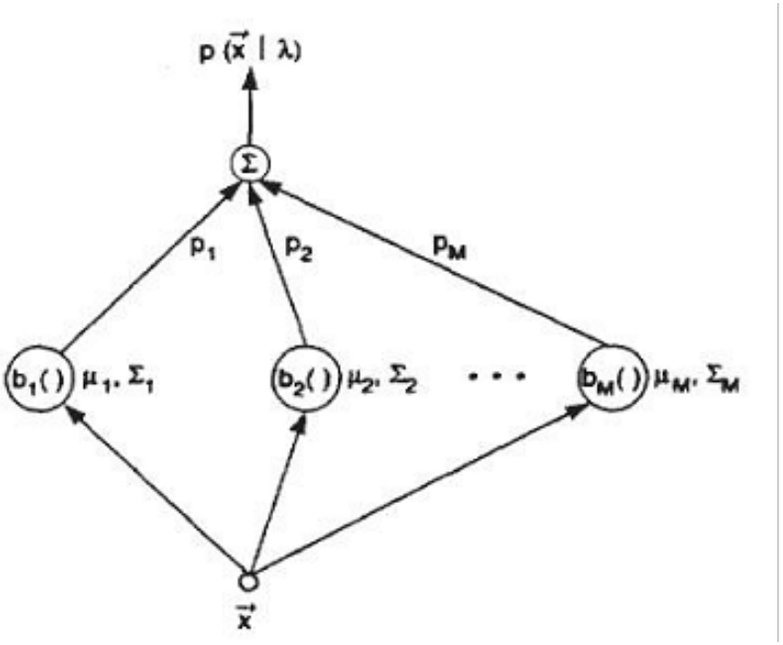

图2. GMM模型图

GMM模型训练的目的即得到特定话者的GMM参数组

$$\lambda=\{p_i,u_i,\sum_i\},i=1,\ldots,M$$

。

步骤为：

1）、读入训练语音的MFCC参数序列，即

$$x_i[]=r_i[],i=1,\ldots,T$$

；T =训练语音的总帧数。

2）、设定起始参数值

$$\lambda_0 = \{\vec{p}_{0i}, \vec{u}_{0i}, \sum_{0i}\}, i = 1, \ldots, M$$

3）、用期望值最大化算法（简称EM），迭代计算

$$\tilde{\lambda} = \{\tilde{p}_i, \tilde{u}_i, \sum_i\}, i = 1, \ldots, M$$

，直至

$$\left\| \tilde{\lambda} - \lambda \right\| < \varepsilon$$

，算法停止。得到的 λ 即为特定话者GMM参数组。

步骤2）具体算法为：

$$\vec{p}_{0i} = \frac{1}{M}$$

；//这里表示M个值都取1/M

$\vec{u}_{0i}$ 由k-均值算法获取；用以训练k-均值的向量数量为1..T

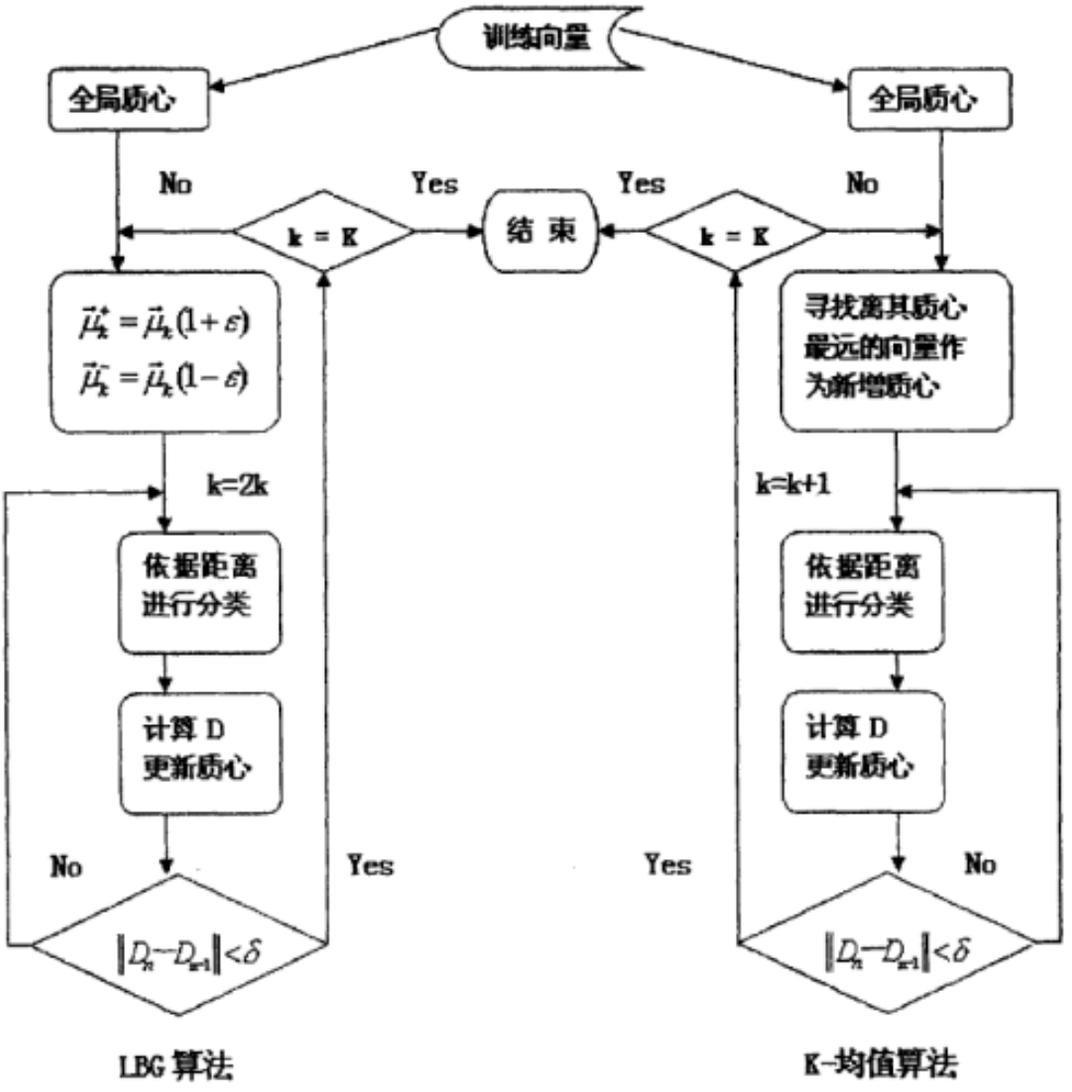

图3. k-均值算法示意图

$$\Sigma_{0i} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_D \end{pmatrix}$$

为协方差矩阵，i = 1,...,M，D是MFCC参数矢量维度=20，为计算方便，假设其为对角阵。

$$\vec{\sigma}_{0i}^{\,2} = \frac{\sum_{t=1}^{T} \vec{x}_t^{\,2}}{T} - \vec{u}_{0i}^{\,2}$$

，

$$\vec{\sigma}_{0i}^{2}$$

为

$$\left(\sigma^{2}{}_{i1},\ldots,\sigma^{2}{}_{iD}\right)$$

的一组矢量，共M组。

k-均值算法一次性得到了所有的$\vec{u}_{0i}$，1≤i≤M

步骤3）具体算法为：

a) 准备好T 个训练向量，记为

$$X=\{\vec{x}_1,\ldots,\vec{x}_T\}$$

b) 计算事后概率

$$p\left(i\left|\vec{x}_t,\lambda\right.\right)=\frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)},i=1,\ldots,M$$

，λ为上一轮迭代后获得的GMM参数组。

公式中

$$\vec{\chi}_t$$

是表示要计算每个训练向量的事后概率，共计算T个M组的事后概率

也就是说每个训练向量都对应一个M组的事后概率

c) 计算

$$\vec{p}_i=\frac{1}{T}\sum_{t=1}^{T} p(i\left|\vec{x}_t,\lambda\right.)$$

这里的pi是M维向量

d) 计算

$$\vec{u}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)}$$

e) 计算

$$\vec{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \vec{u}_i^2$$

f) 计算

$$\left\|\lambda - \tilde{\lambda}\right\| < e ?$$

若是，则迭代训练结束，得到话者GMM参数组模型，

若否，则令

$$\lambda = \tilde{\lambda}$$

，返回b）步继续计算。

注：

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

D = 20；

$$|\sum_i| = \sigma_{i1}^2\sigma_{i2}^2.....\sigma_{iD}^2$$

$$\sum_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_{iD}} \end{pmatrix}$$

四）、声纹识别

假设已训练了S个（>2）GMM声纹模型

$$\lambda_1, \lambda_2, \ldots, \lambda_S$$

，现输入一位话者的语音序列 X（已经过mfcc参数提取），要求判断该话者是谁，即语音序列与哪一个声纹模型匹配。

用后验概率计算

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(\lambda_k | X) = \arg\max_{1 \le k \le S} \frac{p(X | \lambda_k) \Pr(\lambda_k)}{p(X)}$$

由于假定先验概率相同，故上式可简化为求下式：

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(X | \lambda_k)$$

该式又近似于下式。。故实际计算中以下式为准。

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t | \lambda_k)$$

此过程中：

$$p(\vec{x}_t | \lambda_k)$$

只对大于0的值取log，否则为0的值将导致最后的累加值可能出现无穷小

$$p(\vec{x}_t | \lambda_k)$$

即公式（1-1），计算即可。

四、验证实现

采用标准C语言实现：MFCC参数提取，K-means聚类，GMM建模及识别

源码地址如下：

https://github.com/dake/openVP

如果觉得这个项目有用，欢迎以资鼓励：

- 支付宝钱包扫码：

五、参考文献

[1] Douglas A. Reynolds, Richard C. Rose. Robust Text-Independent Speaker Identification Using Gaussian

Mixture Speaker Models.
IEEE Transactions on Speech and Audio Processing, Vol.3, No.1, January 1993.

[2]郭慧娟.声纹识别系统研究[D]．西华大学硕士学位论文，2006.

[3] 魏凯．声纹识别中关键技术的研究[D]．华中科技大学硕士学位论文，2006.

[4] ZhiQiang Wang, Yang Liu, Peng Ding, Xu Bo .Covariance-tied Clustering Method

In Speaker Identification[J].National Laboratory of Pattern Recognition.

Institute of Automation, Chinese Academy of Science Beijing 100080.

[5] 郭皓婷．基于声纹识别技术的应用难点研究[R]．第十四届全国青年通信学术会议，2009.

[6] 张万里，刘桥．Mel频率倒谱系数提取及其在声纹识别中的作用[J]．贵州大学学报，第22卷第2期.

[7] 张广兰．声纹识别的关键技术及发展趋势[J]．黑龙江科技学院，黑龙江，哈尔滨,150027.