学术新闻 / 研究进展

基于GMM-UBM的说话人识别算法

来自 admin | 已发表 十一月 1, 2018

近年来，音频识别作为个人信息验证领域的研究热点发展迅速。目前，语音识别系统在实验室环境中已经可以获得相当好的效果，但在现实场景中，由于噪声的干扰，系统的识别率将受到严重影响，这大大妨碍了语音识别技术在实际环境中的应用。

GMM-UBM作为概率统计模型，由于其能够很好地模拟说话人的声学特征分布，实现方法灵活有效，加上具有较高的鲁棒性，故提出后就迅速成为说话人识别中的重要建模方法。

一、特征参数的提取

对于说话人确认系统来说，从每一帧里提取出代表说话人信息的特征参数是之后所有步骤的基础。在文献中研究了多种特征参数。其中，线性预测参数（Linear Prediction Coefficients, LPC因为其直接由人的发声模型推导而来而受到了广泛关注。知觉预测参数（Perceptual Linear Prediction Coefficients, PLPC）也使用了同样的计算方法。这些参数是基于人体知觉，根据人的听觉过程而开发出来。但是近二十年来，应用最广泛的还是使用傅里叶变换得到的参数。其中由Davies和Mermelstein提出的Mel倒谱系数（MFCC）由于其

充分考虑人体知觉原理和较强的鲁棒性，以及在倒谱域的灵活运用，成为了在说话人识别领域使用最广，效果最好的特征参数。本文关于说话人确认的实验全部采用Mel倒谱系数作为特征参数。

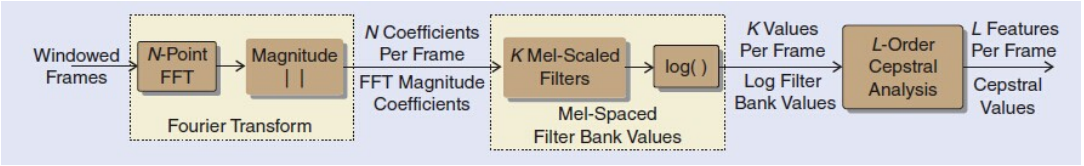

图1Mel倒谱系数的计算过程

Fig.1 Analysis block diagram for MFCC feature vectors

在说话人确认中，提取特征参数的一个重要进展是联合MFCC的一阶，二阶甚至三阶导数可以刻画帧间的动态联系。这些动态信息可以更形象地表现出说话人的特征。类似于与文本有关的语音识别中，这些动态联系可以刻画出说话人的表达习惯。如果定义为第t帧，则MFCC的一阶导数可表示为：

P通常取值为2。

同理可得，将换成可以得到二阶导数。按照同样的计算规则可以依次得到MFCC的更高阶导数。

这些高阶MFCC将联合原阶的MFCC一起构成说话人确认的特征参数。举例来说，一个13维的MFCC参数（L=12加上c_0）可以得到26维的联合一阶MFCC的特征参数，或39维的同时联合一阶和二阶MFCC的级联特征参数，详细阐释请见图2-5。

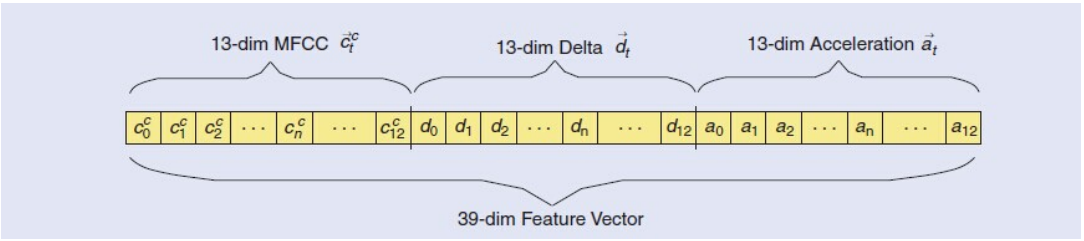

图2级联高阶导数的MFCC

Fig.2 Concatenate of MFCC features with temporal derivatives

二、联合背景模型的混合高斯模型

在说话人确认中，验证测试者是否为已经登记过的目标话者，需要将其的测试语音分别在说话人模型（GMM）和背景模型（UBM）中分别打分，以两者的比值作为最后的结果再与先前设置的阈值进行比较。UBM作为一个混合度较大的GMM（M=256或更多），通过大量说话人的语音数据的训练，平衡了不同人之间发音的差异，可用于任何话者的确认。虽然UBM的训练数据量大幅上升，但由于其表达了与话者无关的特征分布，是所有话者的“并集”，具有背景意义。

GMM-UBM主要用于开集的说话者辨认，因为GMM的性能足以应对普通的闭集测试集合。此外，UBM比单个说话人GMM更精确可靠的原因在于它在训练时调用了所有说话人的数据，因此UBM不会受到训练数据不足以及隐性数据（unseen data）的影响。

在GMM-UBM中，说话人的GMM模型是与UBM维度相同的一组混合高斯模型。不同于原本的GMM模型中训练数据全部来自于单一话者，GMM-UBM中的GMM模型是根据最大后验概率算法（Maximum A-Posteriori，MAP），利用每个说话人的数据在UBM的基础上进行调整修正（adaptation），得到与当前说话人对应的话者模型。所以在GMM-UBM的训练过程中，先要基于全体说话人的数据训练出背景模型UBM（类似于前文GMM的训练），然后在此基础上根据每个说话人的数据调整得到相对应的话者模型GMM。

根据MAP自适应算法可以对说话人的权重，均值，方差分别进行调整，但一般情况只对均值进行调整，实验表明只调整均值效最佳，其具体公式如下：

其中，代表了调整后的第m个高斯分布函数的均值；是调整时的规整因子，由先验知识得到；代表了话者的训练数据与原本UBM的相似度，越大，说明说话人的模型与UBM越相似；是原本UBM的第m个高斯分布函数的均值；代

表了由说话人自身数据训练得到的均值。

由式2-13可知，越大，就越接近1（为定值），越接近0， 表明话者模型进行修正时只改变与UBM中与说话人特征相似的高斯分布函数，使其更接近于目标话者的特征分布，体现了说话人的个性。反之，越小，就越接近0，越接近1，表明与UBM中与说话人特征相似度不高（即隐性信息）的部分几乎保持不变，这样在计算匹配度比值时就避免了低匹配度情况的出现，即原本的隐性信息被融合于背景模型中。此外，当关于话者个人的训练语音增多时，修正后的话者模型也就越远离原本的背景模型而接近真实话者模型的分布。

使用GMM-UBM的优势在于，因为UBM是由大量说话人的数据训练而成，单独训练每一个说话人模型时只需要少量数据进行修正即可，这样得到的话者GMM模型比原本直接训练GMM要可靠得多。此外，由于GMM-UBM中的GMM拥有与UBM相同多的维数，这比单独训练的GMM的维数要多出许多，正因为借助了UBM的优势，所以GMM-UBM在处理隐性数据方面的效果要比GMM好得多。

在GMM-UBM框架下，测试语音的匹配度计算是测试语音与说话人模型GMM和背景模型UBM匹配输出似然度的比值，在评分取对数的情况下，表现为两者的差值，如式2-14所示。

其中是测试语音的一帧的特征参数，和分别是目标模型和背景模型。由式2-14可知，在匹配度计算时，由于两者相减，使原本说话人模型中与背景模型相似的部分、背景噪音和通道的影响被消除，更加凸显说话人个性的同时，也增强了系统的鲁棒性。

假设有N个说话人，基于GMM-UBM的说话人确认系统的训练和测试的过程如图3所示。

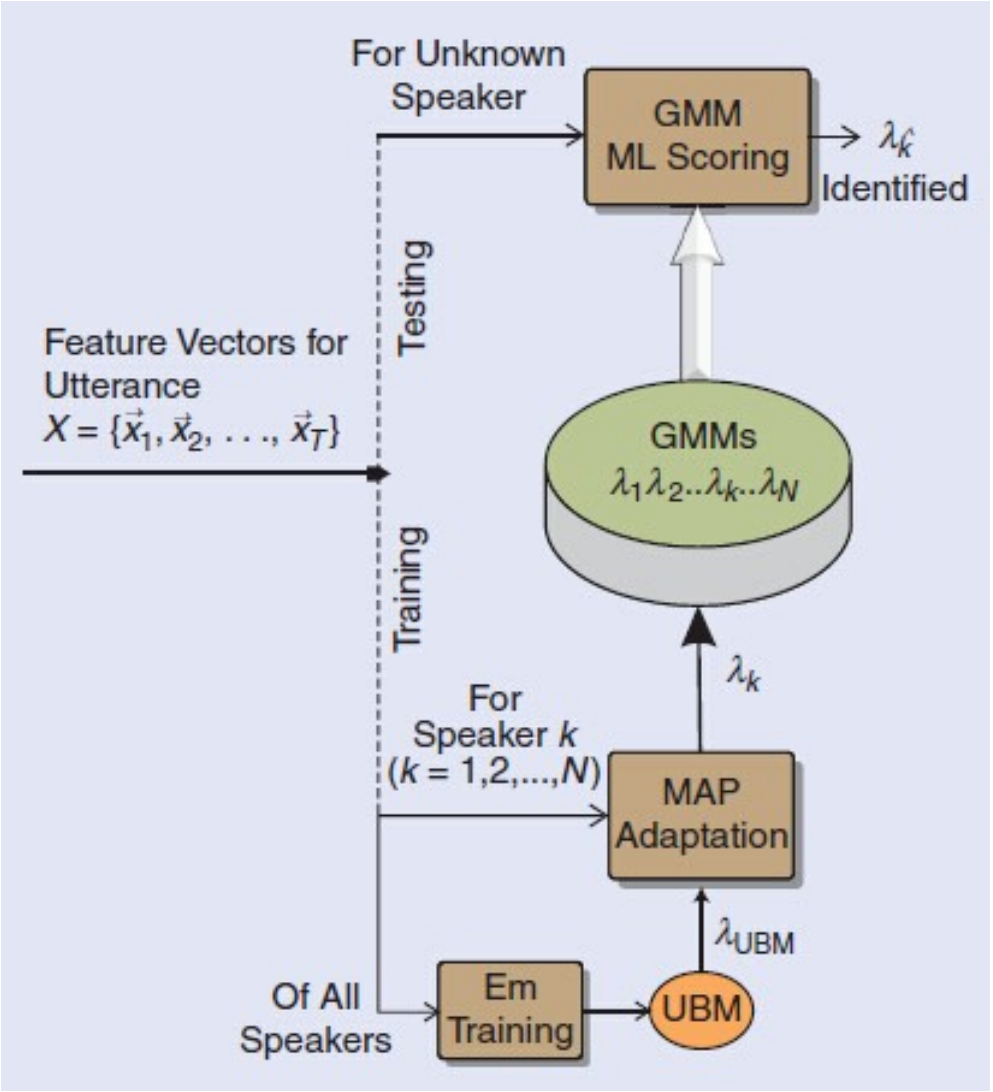

图3 GMM-UBM的训练与测试过程

Fig.3 GMM-UBM training and testing framework

三、效果总结

通过分析四种特征参数MFCC、LPC、LFS、LPCC和三种建模方式GMM、HMM、codebook的不同组合对噪声识别的影响，并分别从纯噪声和估计噪声两个方面进行实验。实验结果如图：

测试语句	噪声识别结果

噪声类型	airport	babble	car	exhibition	restaurant	station	street	train
airport	24	0	0	0	0	6	0	0
babble	2	27	0	0	0	1	0	0
car	1	0	29	0	0	0	0	0
exhibition	0	0	0	30	0	0	0	0
restaurant	2	0	0	0	22	6	0	0
station	1	0	0	0	0	29	0	0
street	2	0	0	0	0	5	23	0
train	0	0	0	0	0	0	0	30

实验结果证明MFCC联合GMM的方法是噪声识别中的最佳组合。最后还分析了该方法对不同的噪声类型的识别结果，大多数估计噪声的正确识别率都能达到70%及以上，说明了MFCC和GMM的组合在噪声识别的实际应用中具有一定意义。