基于MFCC和GMM的说话人识别

黄伟俊

复旦大学

13307130251@fudan.edu.cn

June 29, 2016

I. 摘要

本项目基于课程提供的数据集（关注其中的说话人特征），通过对语音数据端点检测提取有效语音部分后进行MFCC，提取出MEL系数作为GMM的训练数据来训练GMM模型。报告中介绍了说话人识别问题(Speaker recognition)，比较了不同分类器的分类结果，并对训练结果进行了集外测试。

II. 简介

说话人识别是使用计算机利用语音中特定参数来自动识别说话人身份的技术。不同于内容识别,说话人识别关注的是内嵌于一个人发声系统中的特征，而不在意其说话的具体内容。一直以来识别正确率和鲁棒性都是说话人识别的研究重点。由于不同情景下人发音特点的变化（窃窃私语时，打电话时等等），对不同情景下说话人的识别也一直是该领域的难点。基于MFCC的特征提取算法在说话人识别领域被广泛使用，对于识别模型，高斯混合模型（GMM）以及隐马尔科夫模型（HMM）也被广泛使用,GMM–UBM于1994被 提 出[1]。SVM作为一个非常流行的监督式学习算法，在2006年有研究发现利用GMM得到的超向量使用SVM来做说话人识别的效果很好。[2]针对SVM-based的说话人识别的鲁棒性的改进算法WCCN同年被提出。[3]当前,I-vector算法是公认的最好算法之一。

随着网络和只能手机以及机器学习的发展，说话人识别近年来被广泛使用，特别的有Microsoft的Cortana和Apple的Siri都是十分成功的商业实践。未来，随着人们产生的语音信息爆炸式的增加，相信会使得说话人识别领域有更长远的发展与应用。本项目根据在数字信号与语音处理课程上学到的知识以及课程提供的数据集，在报告的第三部分收分探索提供的数据，有一个直观的理解和感受。第四部分详细阐述研究的问题。第五部分针对所要研究的问题对数据进行清理，产生所需的训练集和测试集。第六部分和第七部分别分介绍了核心算法MFCC和GMM，第八部分对实验的过程详细的进行描述。

III. 数据探索

本课程提供的数据集包含30个课程同学的语音数据，其中我使用了27位同学的数据（有三位同学因为语音数据质量欠佳故没有使用），每个同学的数据集都包含了20个单词，每个单词重复二十次，每次以8000HZ的频率采样两秒，对于数据中异常的部分（空语音，背景噪音太大），

我进行了处理，去除异常的语音用该单词正常的语音重复一次来替代（去掉的数量很少，一共不超过10个）。单个人数据集的构成如下表1所示：

由于项目感兴趣的是对说话人进行识别，关注的是说话人的发声特征而不是其说话的内容，训练时的数据集使用了两种方案，一种是把每个人说的每个单词的前19遍作为训练数据，最后一遍作为测试数据，另一种是把除了“语音"，“*start*"这两个词外的其他词作为训练数据，这两个词作为测试数据进行测试。对于前一种情况，每个人训练集大小为20*19=380,总训练集大小为20*19*27=10260,总测试集大小为*540*。对于后一种情况，每个人的训练集大小为18*20=360,总训练集大小为360*27=9720,总测试集大小为*1080*。两种情况所需建的GMM模型数量都是27个。

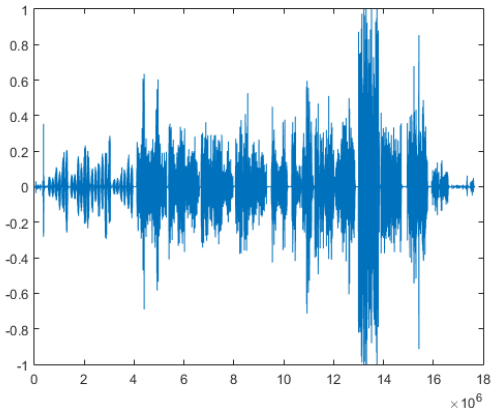

Figure 1: 语音震动的时域图

IV. 研究问题描述

基于对数据集的探索，本项目的目的在于从已有的语音数据出发，提取出个体的语音特点，对集合内的语音（出现过的单词的新语音）和集合外的语音（从未出现过单词的语音）进行识别。

具体来讲，对于每一段待测语音，这是一个类别大小为27的分类问题，对于每个模型得出的预测值我们选用预测值最大的那个类。

V. 数据清洗

对于一段语音，我们要先对语音进行端点检测，去除语音中的无效部分。

i. 端点检测

在进行端点检测前，我们要对声道进行小波去噪，去掉语音中的噪声并对幅度进行归一化，之后计算出短时能量和过零率(如图2)，依据短时能量与过零率进行端点检测(如图3)。

具体的，我们

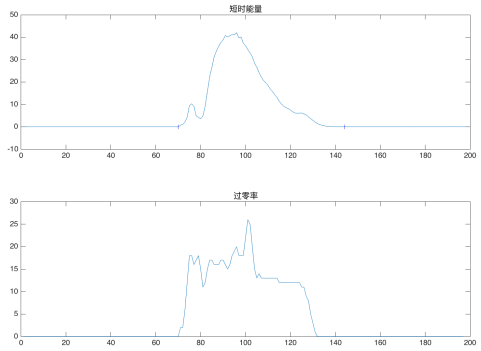

Figure 2: 短时能量与过零率

VI. 特征提取(MFCC)

对于经过端点检测后的语音数据，使用*MFCC*提取特征。

梅尔倒谱系数（Mel-scale Frequency Cepstral Coefficients，简称MFCC）是在Mel标度频率域提取出来的倒谱参数，Mel标度描述了人耳频率的非线性特

Table 1: 单数据集一览

包含单词	北京	背景	分析	复旦	商行	上海	识别	数字	信号	语音
单人数据量	20	20	20	20	20	20	20	20	20	20
包含单词	close	file	happy	lucky	open	sound	speech	start	stop	voice
单人数据量	20	20	20	20	20	20	20	20	20	20

性，它与频率的关系可用下式近似表示:

$$f_{mel} = 2595 * \log 1 + \frac{f}{700Hz}$$

其中f为频率(Hz),图5展示了Mel频率与线性频率的关系

i. 预加重

预加重操作讲信号通过高通滤波器

$$H(Z) = 1 - \mu z^{-1}$$

式中μ值介于0.9-1.0之间，通常选取0.97。

ii. 分帧

对于本实验我们选取240个采样点为一帧(30ms)，并且帧中心的移动幅度为80个采样点(10ms)。

iii. 加窗

将每一帧用汉明窗进行加窗。假设分帧后的信号为S(n), n=0,1...,N-1, N为帧的大小，那么乘上汉明窗后$S'(n) = S(n)XW(n)$,W(n)形式如下

$$W(n,a) = (1-a) - aXcos[\frac{2\pi n}{N-1}, 0 \leq n \leq N-1]$$

不同的a值会产生不同的汉明窗，一般情况下a取0.46

iv. 快速傅里叶变换(FFT)

对时域信号进行FFT得到频域信号

$$X_\alpha(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi k/N}, 0 \leq k \leq N$$

v. 梅尔滤波器

将能量谱通过一组Mel尺度的三角滤波器(如图7) 对时域信号进行FFT得到频域信号

$$X_\alpha(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi k/N}, 0 \leq k \leq N$$

vi. 计算滤波器组输出的对数能量并获得MFCC系数

$$s(m) = ln(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k), 0 \leq m \leq M$$

$$C(n) = \sum_{m=0}^{N-1} s(m)cos(\frac{\pi n(m-0.5)}{M}, n = 1, 2, ..., L$$

这里L指的是MFCC系数阶数，通常取12-16。M为滤波器个数。

vii. 一阶与二阶差分

标准的MFCC系数只反应了语音的静态特征，可以增加一阶和二阶差分这两维系数

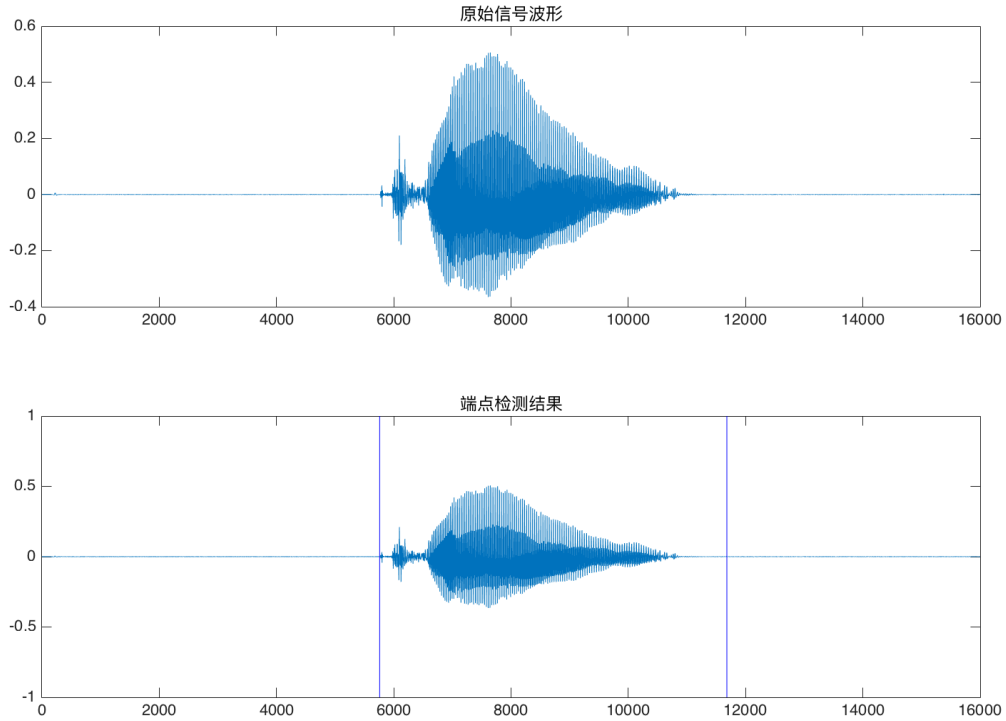

Figure 3: 端点检测结果

来描述动态特征。

$$d_t = \begin{cases} C_{t+1} - C_t, t < K \\ \frac{\sum_{k=1} kk(C_{t+k} - C_{t-k}}{\sqrt{2\sum_{k=1}^{K} k^2}}, v \\ C_t - C_{t-1}, t \geq Q - K \end{cases} \quad (1)$$

VII. 模型训练(GMM)

对于提取完的特征数据，使用高斯混合模型（GMM）进行建模，对于每一个同学（类），都建一个模型勇于分类。
高斯混合模型有多个单高斯模型混合（GSM）而成，其基于一个简单的想法，高斯分布在自然界中普遍存在，其认为人发声的系统的概率分布也可以看作是多个高斯分布混合而成。

$$\begin{aligned} p(x) &= \sum_{k=1}^{K} p(k)p(x|k) \\ &= \sum_{k=1}^{K} \pi k N(x|\mu_k, \sum k) \end{aligned} \quad (2)$$

GMM的log-likelihood function为

$$\sum_{i=1}^{N} log \sum_{k=1}^{K} \pi k N(x_i|\mu_k, \sum k)$$

我们的任务即是确定每个组件中的参数

i. EM算法

1.对于每个数据x_i，有

$$y(i,k) = \frac{\pi_k N(x_i|\mu_k, \sum k)}{\sum_{j=1}^{K} \pi_j N(x_i|u_j, \sum_j)}$$

2.

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} y(i,k) x_i$$

$$\sum k = \frac{1}{N_k} \sum_{i=1}^{N} y(i,k)(x_i - \mu_k)(x_i - \mu_k)^T$$

设定参数的初始值然后重复上述两个步骤直到似然函数收敛。

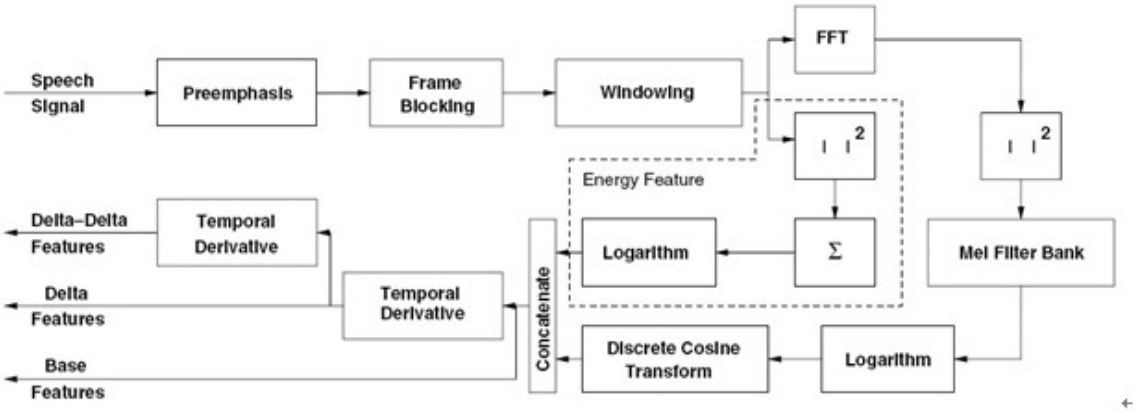

Figure 4: *MFCC-FLOW*

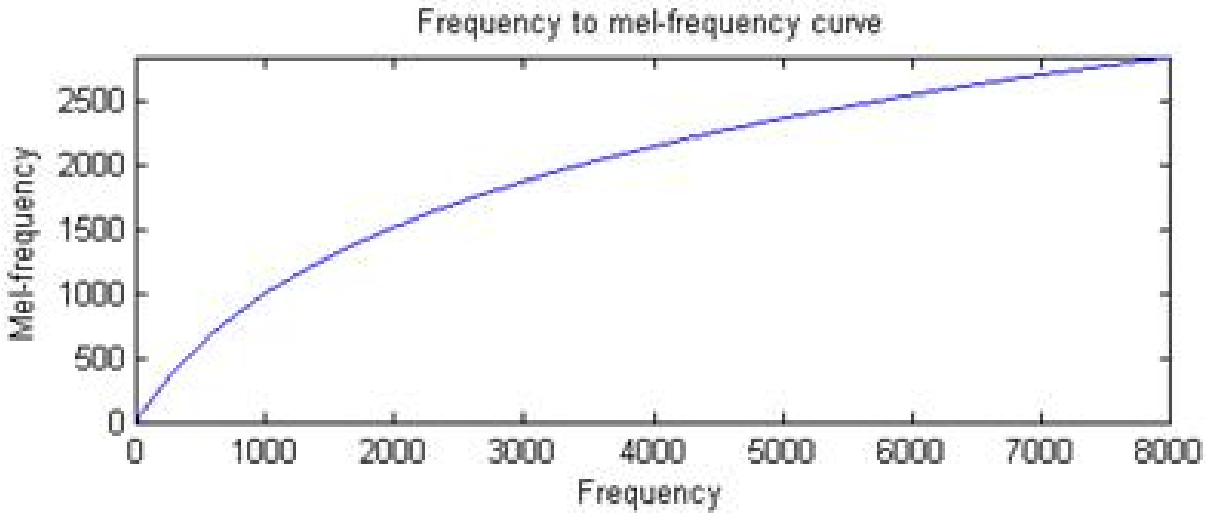

Figure 5: *MEL-linear*

训练数据的获取是把一个人（类）的待训练样本的MFCC系数序列全部连在一起进行训练。

ii. 判别方法

对于特征序列X,其由说话人S产生的概率为

$$Pr(\lambda_S|X) = \frac{X|\lambda_S Pr(\lambda_S)}{p(X)}$$

$p(X|\lambda_S)$是模型λ_S产生序列X的概率密度；$Pr(\lambda_S)$是说话人S的先验概率，一般都假定先验概率是相等的；$p(X)$是X来自任意说话人的先验概率。

VIII. 实验分析

i. 实验方法

实验在几种条件下进行

基础情况MEL系数的个数为16个，混合模型的数目为14个,MFCC低频限制为300HZ，高频限制为3700HZ。

Table 2: 各条件下的实验结果

实验条件(默认mix为14,MFCC维数为16)	集合内识别率	集合外识别率（测试集251）
(每个单词前19遍训练，最后一遍测试)	99.6%	缺失
(除了语音与start外都训练)	89.4%	82%
(梅尔系数的数量为15其于同上)	88.9%	74%
(增加一阶差分,其于同上)	90%	83%
(混合数为16)	89.9%%	77.4%
(全训练)	无	88.7%

i.1 1

对于每个人的每个单词，取前19个进行训练，最后一个拿来测试，此时每个人有380个训练数据，测试数据一共有540个。

i.2 2

对于每个人，取其除“语音”和“*start*”外的全部数据集进行训练,此时每个人有360个训练数据，测试数据一共有1080个。

i.3 3

在上述基础上，修改MEL系数的数目为15个。

i.4 4

在3的基础上训练全部的数据集。
识别结果的获得方法是，对于一段语音提取出的MFCC特征序列，每对MFCC系数都用每一个GMM进行预测然后投票，总票数最多的那个类即识别为说话人。

ii. 测试性能

测试性能的结果如表2所示,应当注意不同条件下的测试集和训练集都并非完全相同，不同训练集和测试集下的结果不好直接比较。
但是也足以在一定程度上表明问题，大体来看，在相同测试集和训练集的条件下在一定范围内随着训练数据的增加，MEL系数维数的增加，高斯模型的混合数增加，三角滤波器数量的增加，性能是在上升的（由于测试时部分条件下的结果并未记录，所以在表格中没有体现出来），但是在一个合理的区间内(如MEL系数维数为12-16,三角滤波器数量为22-26)，增长对性能的提升并不显著。
具体来讲在MEL系数维数为16，滤波器数量24个，混合数为14的情况下，采用实验方法1的识别率为99.6%(注意此时测试集为每个单词的第二十遍,即为集内测试）。
除了“语音”和“start”这两个词其他数据都拿来训练，把这两个作为测试集（集外

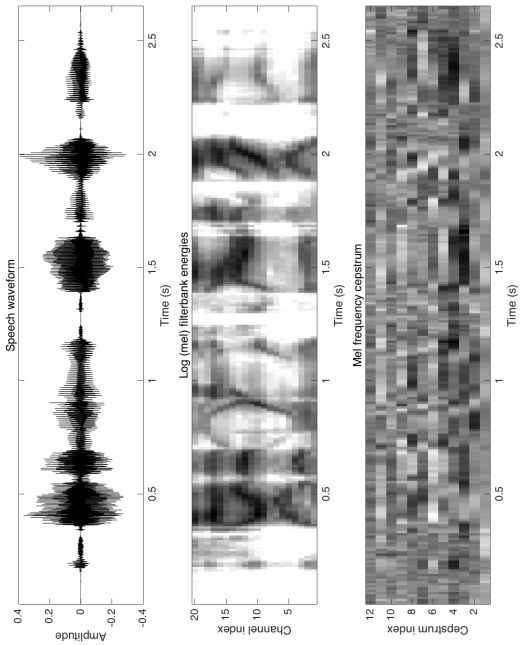

Figure 6: 梅尔图谱

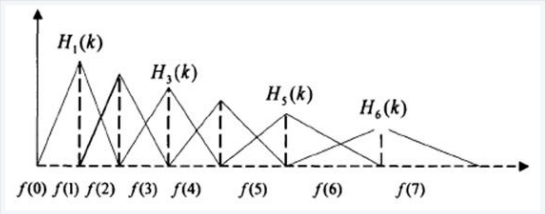

Figure 7: 三角滤波器

测试），识别率为89.4%，另外，我自己录了一些集外的单词（大小为50个）（下记为测试集251），在其上的识别率为82%。

除了“语音”和“start”这两个词其他数据都拿来训练，把这两个作为测试集（集外测试）,且MEL系数维数调整为15个，识别率为88.9%，在测试集251上的识别率为74%。

除了“语音”和“start”这两个词其他数据都拿来训练，把这两个作为测试集（集外

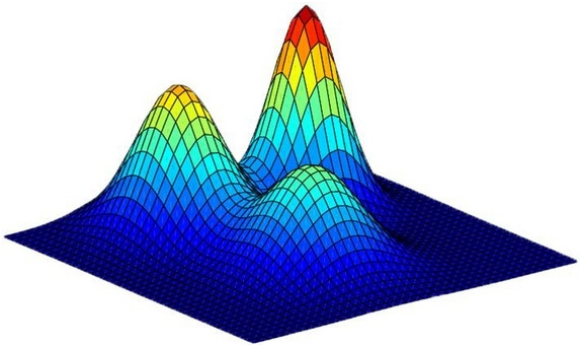

Figure 8: 高斯混合模型*(GMM)*

测试），且MFCC系数增加一阶差分，识别率为90%，在测试集251上的识别率为87%，另外此时还测试了另一位同学的大小为100的集内测试（即数据集中出现过单词的新录音，每个单词5遍）（记为测试集498)识别率为83%。除了“语音”和“start”这两个词其他数据都拿来训练，把这两个作为测试集（集外测试），且把高斯混合数改为16，识别率为89.9%,测试集251识别率77.4%,测试集498识别率85%。把数据集全部拿来训练，在测试集251上的识别率为88.7在测试集498上的识别率为88%。

iii. 实验结果分析

由于本实验的数据集大小比较有限，不同的词种类只有20，考虑到文本无关的说话人识别应该有足够丰富的数据集使得包含各种不同的语音特征，对于本项目受限的数据集大小，可能有一部分音节没有出现过，这对集外测试时可能会有较大的影响，且受限于时间没有充分地在各种不同参数的条件下的进行实验。不过，根据在已实验部分的观测结果上可以发现，模型参数合理的范围内越大识别率是越高的，但是随着混合数的增加，MFCC系数维数的增加，算法训练、预测的时间也显著的增长了，同时对于本数据集有限的语音长

度（2S），过高的混合度可能会造成性能的下降。

IX. 结论

在文本无关条件下的说话人识别问题是具有很重要意义的问题。本项目讲用短时语音信息(2s,8000Hz),利用端点检测提取处音频的说话内容部分，用MFCC进行特征提取，最后用GMM进行模型训练并对一段未知语音信息进行说话人识别 。
在报告中我们展示了一个完整的实验流程，从粗略的数据处理，到精细的特征提取和标准化，再到算法的研究和讨论，最后比较结果的优劣并在新的测试集上进行验证。
对于数据集中的集外测试（即测试集为“语音”和“start”这两个词），识别率是较好的，我想这有赖于数据集的录制经过了同学们的删选，并且录制的环境比较理想。对于完全的集外测试，识别率则较差，这一方面可能是由于两次录音间隔几个月，个体当时的发声系统状态不完全一致，另一方面也是录音的设备上可能发生了变化，并且本次使用的MFCC-GMM识别算法的鲁棒性较差，在这一方面，已知的GMM-UBM,GMM-SVM,I-VECTOR算法都有着更优良的变现，其中I-VECTOR算法的性能十分优良，这有待于进一步的实验。[4]

参考文献

[1] J. Gauvain and C. Lee
Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains IEEE Trans. Speech Audio Processing,vol. 2, no. 2, pp. 291–298, 1994

[2] W. M. Campbell, D. E. Sturim, and D. A. Reynolds *Support vector machines using GMM supervectors for speaker verification.* IEEE Signal Process. Lett., vol.13, no. 5, pp. 308–311, 2006

[3] A. O. Hatch, S. S. Kajarekar, and A. Stolcke *Within-class covariance normalization for SVM-based speaker recognition.* in Proc. Interspeech, Pittsburgh,PA, pp. 1471–1474, 2006.

[4] John H.L. Hansen and Taufiq Hasan *Speaker Recognition by Machines and Humans.* IEEE SIGNAL PROCESSING MAGAZINE [74]november 2015.