文章编号:1671-4598(2016)04-0155-04 DOI:10.16526/j.cnki.11—4762/tp.2016.04.046 中图分类号:TN912.34 文献标识码:A

基于改进噪声估计的谱减法应用于说话人识别

李哲军[1]，周 萍[1]，景新幸[2]
(1. 桂林电子科技大学 电子工程与自动化学院，广西 桂林 541004；
2. 桂林电子科技大学 信息与通信学院，广西 桂林 541004)

摘要：针对语音信号中存在加性噪声使MFCC的鲁棒性和识别系统的性能下降的问题，基本谱减法的引入在增强MFCC抗噪性上取得的效果有限，为了使MFCC具有更好的抗噪性，提出了一种改进算法，在谱减法的基础上引入谱熵的思想，利用谱熵值的分布逐帧进行噪声估计，可更精确地谱减去噪；实验结果表明，当语音中含有加性噪声时，与基本谱减法相比，改进谱减法的说话人识别系统抗噪性与鲁棒性更好。

关键词：说话人识别；谱减法；谱熵；噪声估计；梅尔频率倒谱系数

Speaker Recognition Using Spectral Subtraction Method Based on Improved Noise Estimation

Li Zhejun[1], Zhou Ping[1], Jing Xinxing[2]
(1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China;
2. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Aiming at the problem that additive noise in speech signal makes the performance of speaker recognition system degradate when using MFCC. The introduction of traditional spectral subtraction achieved some effect on enhancing noise immunity of MFCC, but the improvement is limited. To get a better result, a novel algorithm of spectral subtract is proposed in this paper. The concept of spectral entropy is introduced based on the spectral subtraction, the noise of each flame is estimated more accurately according to its spectral entropy and subtracted to get better denoising effect. Experimental results show that when there is additive noise in the test speech, compared with traditional spectral subtraction, the speaker recognition system of novel algorithm has better noise immunity and robustness.

Keywords: speaker recognition; spectral subtraction; spectral entropy; noise estimation; MFCC

0 引言

声纹识别[1]是通过语音识别说话人的身份，与指纹识别、文字密码等认证技术相比，其具有不会遗失、无须记忆、实现简单等特点，是一种非接触识别方式。有效特征参数[2]的提取是其关键问题，常见的特征参数有线谱对参数（LSP）、线性预测倒谱参数（LPCC）、Mel频率倒谱系数（MFCC）等，其中MFCC因能充分描述人耳的感知特性而应用广泛[3]。

语音纯净不含噪时MFCC的鲁棒性及系统识别效果都比较好，但系统在语音含噪时的识别性能下降明显。针对语音中存在的加性噪声降低识别性能的问题，已经有许多改进算法[4]，有倒谱均值与方差规整（CepstralMean and Variance Normalization，CMVN）、特征弯折、RASTA滤波等，都曾被用来提高MFCC的鲁棒性，但它们都存在需要延迟处理的缺点。

首先，本文研究了语音增强中的谱减法[5]（Spectral Subtraction，SS），相比传统MFCC，加入谱减法的系统处理含有加性噪声的语音时性能有提高但程度有限，于是提出了改进算法以进一步提高MFCC的抗噪性。在基本谱减法基础上引入谱熵[6]的概念，根据谱熵的定义和性质分析噪声与语音信号的谱熵分布规律，用以动态更新噪声谱值，使谱减更精确、所提取的MFCC抗噪性更好。此外，实验采用GMM-UBM模型[7]代替GMM模型以弥补样本的不足。实验结果表明改进谱减法的说话人识别系统抗噪性改善明显。

1 MFCC特征参数

常用特征参数可分为时域和频域两类，时域中有幅度、平均过零率等参数；频域中有线谱对参数（LSP）、线性预测倒谱系数（LPCC）、共振峰频率、Mel频率倒谱系数（MFCC）等，其中MFCC因反映了人耳听觉特性而具有较好的鲁棒性。

MFCC采用的是梅尔频率，代表着人耳对不同频率声音的感受程度[8]：在1 000 Hz以下人耳感知较为敏锐，与频率近似成线性关系；在1 000 Hz以上人耳感知与频率成对数关系。梅尔频率与赫兹频率的转换公式为：

$$f_{mel} = 2595 lg(1 + f_{hz}/700) \quad (1)$$

其提取过程如图1所示。

1）预加重：滤除低频干扰，补偿受发音系统所抑制的高频部分，其传递函数为：

$$H(z) = 1 - kz^{-1} \quad (2)$$

其中：k介于0.9和1.0之间，本文实验中取0.95。

2）分帧：将N个采样点集合成一个观测单位，称作帧，

收稿日期:2015-10-10； **修回日期**:2015-11-08。
基金项目:广西研究生教育创新计划资助项目(YCSZ2015152)；国家自然科学基金(61363005)。
作者简介:李哲军(1990-)，男，湖北天门人，硕士研究生，主要从事语音识别方向的研究。
周 萍(1961-)，女，河北唐山人，教授，硕士研究生导师，主要从事智能控制及语音信号处理的研究。
景新幸(1960-)，男，湖北武汉人，教授，硕士研究生导师，主要从事语音识别及其混合集成电路的研究。

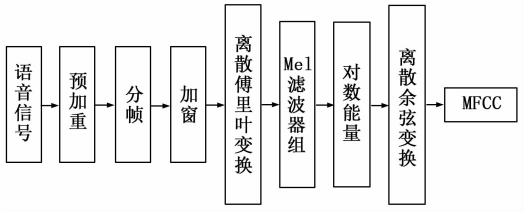

图 1 MFCC 提取流程

为避免相邻两帧间变化过大，相邻帧间有一段重叠区域，称作帧移，常为 N 的 1/2 或 1/3。

3）汉明窗：增加窗边界处信号的连续性，减小吉伯斯效应：

$$\omega(n)=0.54-0.46\cos\left[\frac{2\pi n}{(N-1)}\right],0\leqslant k\leqslant N-1 \quad (3)$$

4）离散傅里叶变换：将信号的时域分布变换为频域上的能量分布：

$$X(k)=\sum_{n=0}^{f_m+1}x(n)e^{-j2\pi nk/N},0\leqslant k\leqslant N \quad (4)$$

5）Mel 滤波：消除谐波，降低数据维数，将离散谱 $X(k)$ 通过 M 个 Mel 滤波器组，得到 M 个 $h(m)$ 参数：

$$h(m)=\sum_{k=f_{m-1}}^{f_m+1}W_m(k)X(k),\quad m=1,2,\cdots,M \quad (5)$$

6）离散余弦变换：将经过对数运算的滤波输出变换到倒谱域，得到 MFCC 参数：

$$C_{mfcc}(n)=\sqrt{\frac{2}{N}}\sum_{m=1}^{M}\cos\left[(m-0.5)\frac{n\pi}{M}\right]\mathrm{In}[h(m)],\quad n=1,2,\cdots,M \quad (6)$$

由以上步骤得到的静态 MFCC，经差分运算可得到一阶差分倒谱系数 ΔMFCC，ΔMFCC 作为动态特征参数，能更加完整地表征说话者的动态语音特征[9]，描述语音信号帧间变化即说话者的动态特征。本文实验中采用 MFCC 与 ΔMFCC 的组合参数。

2 基于谱减法的语音增强

语音增强是从带噪语音中消去或减小其中的噪声以获得较纯净的语音，使提取的特征参数接近于无噪声的情况。语音增强主要有谱减法、Wiener 滤波法、最小均方误差估计法等，其中谱减法具有计算量小、引入约束条件少等优点而应用广泛。

2.1 基本谱减法

基本谱减法中，假定且噪声和语音不相关且噪声为加性噪声，记为 $z(n)$，纯净语音信号为平稳信号，记为 $s(n)$，则带噪信号为：

$$y(n)=s(n)+z(n) \quad (7)$$

设 $y(n)$、$s(n)$、$z(n)$ 的傅里叶变换分别为 Y_k、S_k、Z_k，则对（7）两边进行傅里叶变换有：

$$Y(k)=S(k)+Z(k) \quad (8)$$

于是可得：

$$|Y_k|^2=|S_k|^2+|Z_k|^2+S_kZ_k^*+S_k^*Z_k \quad (9)$$

由于 $s(n)$ 与 $z(n)$ 相互独立，则 S_k 和 Z_k 独立，且 Z_k 满足高斯分布且均值为零，则有：

$$E\|Y_k|^2|=E\|S_k|^2|+E\|Z_k|^2| \quad (10)$$

记为无语音时的统计平均值，则对于分帧内的短时平稳过程有：

$$|Y_k|^2|=|S_k|^2+\lambda_z(k) \quad (11)$$

于是，增强后的语音信号为：

$$|S_k|=[|Y_k|^2-E\|Z_k|^2|]^{1/2}=[|Y_k|^2-E\|\lambda_z(k)]^{1/2} \quad (12)$$

基本谱减法的核心是以无语音帧中噪声的统计均值替代整段语音的噪声估计，但以不变的统计均值替代随机变化的噪声进行谱减就会产生很大误差，出现残留噪声即音乐噪声。为了改善音乐噪声问题而出现了许多改进的谱减法：有人将听觉掩蔽模型用于谱减法[10]，但要人为设定参数，会增加系统复杂度和引入新的失真；有人提出在谱减法计算谱值时引入修正系数[11]，但人为确定的系数并没有改变以偏概全的本质；还有人提出添加语音活性检测[12]步骤，但在低信噪比时效果较差。本文在基本谱减法的基础上引入谱熵的概念，用以更为精确地进行噪声估计以获得更好的去噪效果。

2.2 谱熵与频谱的关系

针对短时平稳的语音信号，将其分成若干短时帧，然后经傅里叶变换得到的短时频谱并进行归一化处理，其概率密度函数如下：

$$p_i=s(f_i)/\sum_{k=1}^{N}s(f_k),i=1,2,\cdots,N \quad (13)$$

其中：$s(f_i)$ 是频率分量 f_i 的频谱值，对应的概率密度值为 p_i，N 为 FFT 的频率点数，每帧谱熵定义为：

$$H=-\sum_{k=1}^{N}p_klog_2(pk) \quad (14)$$

谱熵是熵的一种形式，具有熵的基本性质[13]：熵值不因各分量的次序改变而变化；熵值在集合中的事件等概率发生时达到最大值，例如在式（14）中有 $H\leqslant\log_2(N)$。由谱熵的定义和性质可知，每帧谱熵值仅与频谱的分布有关，与频谱值不直接相关，且语音谱熵值随频谱分布的变化有如下规律：

纯静语音的频率分布的范围较小，频谱 $s(f_i)$ 及其概率分布 p_i 较为集中，可表示为 $p_{i1}=(p_1,p_2,\cdots,p_s,0,0,\cdots,0)$，$i=1,2,\cdots,N$，$s\ll N$；噪声的频谱较为丰富，频谱 $s(f_i)$ 及其概率分布 p_i 也较为分散，可表示为 $p_{i2}=(p_1,p_2,\cdots,p_n,0,0,\cdots,0)$，$i=1,2,\cdots,N$，$n\approx N$；对于 $H(p_{i1})$ 和 $H(p_{i2})$，由于 $s\ll n$，根据谱熵的性质可以知 $H(p_{i1})<H(p_{i2})$，即噪声的谱熵值总是大于纯净语音的。

综上可知，谱熵值受频谱分布影响且与频谱幅度不直接相关，于是可根据谱熵值更准确地区分噪声帧和语音帧使提取的特征参数具有更好的鲁棒性。

2.3 基于谱熵的谱减法改进

噪声值的估计不准会使谱减去噪时产生音乐噪声，且噪声值随机变化，但其谱熵值变化不大，本文根据各帧的谱熵变化来确定并动态的更新噪声值，每一帧都减去更新后的噪声值，由信号的短时平稳性可知，这样进行谱减更为准确[14]。

基于谱熵噪声估计改进的谱减法（Improved Spectral Subtraction，ISS）分为 3 个部分：

1）初始噪声估计，将谱熵值最大的一帧作为噪声帧并将该帧频谱值更新为初始噪声值；

2）噪声更新，根据判断新一帧与前一噪声帧谱熵值的比值是否大于设定阈值 r（根据实验，取为 0.95）：是则判定此帧为新噪声帧并更新其频谱值为噪声谱值，否则当前帧的噪声值等于前一帧的噪声值；

3）谱减，每一帧减去更新后的噪声值完成消噪。

加入改进谱减法后的 MFCC 提取算法过程如下：

1）输入含噪语音；

2）对每一语音帧进行 FFT 变换，得到语音频谱 S_i，其中，$i=1, 2, \cdots, N$；

3）计算每一帧的谱熵值 $h(S_i)$，将谱熵值最大的一帧 m 作为初始噪声帧，即 $Noise=S_m$；

4）若新的一帧的谱熵值与前一纯噪声帧的比值大于阈值 γ（取为 0.95），即 $h(S_n)/h(S_m)>\gamma$，$n=1, 2, \cdots, N$，此时便更新噪声估计 $Noise=S_n$；

5）利用前面已得到的语音谱 S_i 以及更新后的噪声帧估计 Noise 进行谱减；

6）输出消噪后的增强语音频谱。

3　GMM-UBM 模型

3.1　GMM 模型

GMM[15]模型原理是若干高斯函数的线性组合可逼近任意曲线，其作为一种概率统计模型能精确地描绘说话人特征参数的概率分布。对于混合度为 M、模型参数为 λ 的 GMM，特征矢量为 X，则 X 在该 GMM 模型下的似然度为：

$$p(X \mid \lambda)=\sum_{i=1}^{M}\omega_i N_i(X) \tag{15}$$

式中，ω_i 为混合权值，满足 $\sum_{i=1}^{M}\omega_i=1$；$N_i(X)$ 表示第 i 个混合高斯分量的高斯密度函数：

$$N_i(X)=\frac{1}{(2\pi)^{D/2}\mid\sum_i\mid^{1/2}}exp\left\{-\frac{1}{2}(X-\mu_i)^T\sum_i^{-1}(X-\mu i)\right\} \tag{16}$$

式中，μ_i 表示均值向量，$\sum_i$ 表示协方差矩阵，本文 $\sum_i$ 采用对角阵的形式以方便计算。

GMM 模型参数包含混合权值、均值矢量及协方差矩阵，即 $\lambda=\{\omega_i, \mu_i, \sum_i\}$，$i=1, 2, \cdots, M$，$\lambda$ 可通过 EM 算法[16]估计得出。

3.2　GMM-UBM 模型

GMM 模型在训练和测试语音都足够长且语音较纯净的情况下，其识别效果比较理想。当训练语音只有数十秒、测试语音只有几秒时，GMM 模型就不能很好地刻画说话人特征。GMM-UBM 模型的原理是先利用所有的语音训练得到一个 UBM，然后基于 MAP（Maximum A Posteriori）自适应 UBM 得到目标说话人的 GMM 模型，可用来弥补数据的不足。UBM 是一个大型的高斯混合模型，可反映所有说话人语音特征以及环境通道的共性，通过大量说话人在各种环境下的数据训练获得。

在 GMM-UBM 模型中，对于测试语音的特征矢量序列 $X=\{X_i\}$，$i=1, 2, \cdots, M$，每个说话人的对数概率得分计算公式如下：

$$S(X)=\frac{1}{M}\sum_{i=1}^{M}[log\ p(X_i \mid \lambda_s)-log\ p(X_i \mid \lambda_{UBM})] \tag{17}$$

式中，λ_s 为目标说话人的 GMM 模型参数，λ_{UBM} 为 UBM 模型参数。

训练阶段利用大量的语音进行训练得到 UBM，在 UBM 的基础上通过 MAP 自适应得到目标说话人的 GMM 模型。测试阶段根据已经训练好的 UBM 模型和 GMM 模型，利用公式（17）计算出对数概率得分，找到最大的得分者即目标说话人。基于 GMM-UBM 模型的说话人识别原理框图如下：

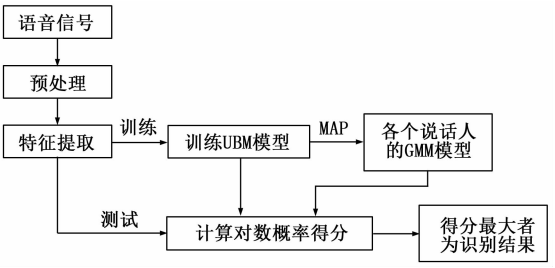

图 2　GMM-UBM 识别模型框图

采用似然比打分的方式是一种归一化处理，可对不同的目标话人设置统一的判决阈值。识别时分别计算似然度得分，选取最大者对应的目标说话人即为识别结果[17]。

4　实验结果与分析

4.1　实验设置

硬件环境：PC 个人计算机（Intel（R）Core（TM）i5－3210M CPU@2.5 GHz）。

软件环境：Windows 7 操作系统、MATLAB R2010a 和 CoolEditpro-v2.0 录音软件。

实验采用的语音库为自建小型普通话语音数据库。语音文件在普通研究室环境下录制，采样频率为 8 kHz，量化精度为 16 bit。60 名录音者（34 名男性、26 名女性）随机朗读 5 分钟（文本无关）。从每人语音中截取 UBM 训练语音（1 min）、GMM 训练语音（10 s）和测试语音（5 s）。为提高本文后续实验的有效性，进行截取时避免所截取的语音段重复。

实验采用 13 维 MFCC 与 13 维 ΔMFCC 组成的组合参数，按帧长 256 个采样点、帧移 128 个采样点逐帧提取语音特征参数。训练阶段依次训练 UBM 模型（高斯混合度为 128）和 GMM 模型，之后通过 MAP 自适应得到目标说话人的 GMM 模型。测试阶段从语料库中选取 50 个说话人构成测试集，每个人有 5 段测试语音。

4.2　实验结果与分析

实验一：不同信号的幅值及谱熵值的对比：

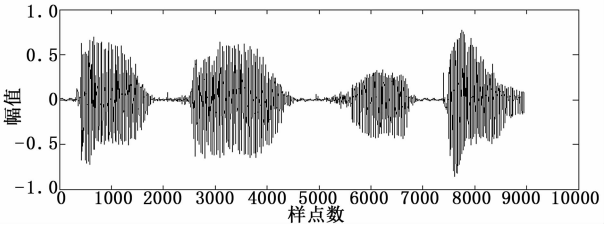

图 3　纯净语音信号 S

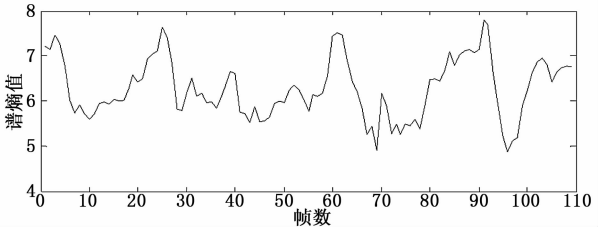

图 4　S 的谱熵值

从图 4 和图 5 可以看出，纯净语音信号中语音帧的谱熵值

安装并配置 JDK1.8、Apache1.9、Android SDK、Android NDK 这 4 个软件，配置完成后即可编译生成后缀为 Apk 的 Android 程序，然后在手机上进行调试[5-6]。

项目采用的软件开发工具 QT 是一种跨平台的开发工具，因此只需对客户端软件进行部分修改，绘制 APP 软件界面，在配置好的环境下重新编译，即可生成 APP 软件。项目开发的 APP 软件具有与客户端软件完全相同的功能，由于手机屏幕较小无法同时显示所有功能，分别绘制了参数设置界面、参数和状态查询界面、实时通信界面，这 3 个界面可实时切换，便于用户使用。

4 系统测试

本文设计与实现的系统需要在恶劣环境下长期工作，为了验证系统的可靠性和实用性，进行了大量的数据采集测试、应急模式测试和稳定性测试。

4.1 数据采集测试

利用标准信号源提供信号，接到 ADC 的 16 个通道上，通过主控板的 AD 子程序读取采样数据，转发到客户端软件进行显示，结果如表 2 所示。

表 2 ADC 数据采集测试结果 V

真实值	测量值	误差
−5	−4.994	0.006
−4	−3.998	0.002
−3	−2.997	0.003
−2	−1.998	0.002
−1	−0.998	0.002
0	0	0
1	0.998	0.002
2	1.998	0.002
3	2.997	0.003
4	3.997	0.003
5	4.993	0.007

由表 2 可以看出，获得的采样值与标准信号源的值误差在 0.01V 以内，采样率误差较小，系统可以准确地采集数据，且数据转换精度高、误差小，符合使用要求。

4.2 应急模式测试

应急模式是系统监测到某个采样值超过设定的阈值时紧急启动的一种机制，触发后立刻进入发射模式，将当天的采样数据发送出来。

模拟实际情况进行应急模式测试，在系统平稳运行采集数据时，给通道一个超过阈值的信号，观察客户端软件发现系统立刻进入了应急模式，完成了数据发送，事后通过回看系统日志也证实在该时刻系统进入了应急模式。在不同采样时间段分别进行应急模式测试，从触发阈值到进入应急模式均在 1 秒内，大量测试表明应急模式满足实际应用需求。

4.3 系统稳定性测试

稳定性是系统最重要的指标，设置好各项参数后，启动系统并进行了 72 小时不间断测试，在每天固定时间启动 AD 子程序进行数据采集，每天固定时间进行数据发送，每 1 小时进行一次系统校时确保与卫星时间同步。

通过分析操作日志以及观察客户端的实时数据，发现系统运行稳定，功能完善，在 72 小时内未出现错误，达到了预期的效果。

5 结论

本文提出了一种基于 ARM 和 Linux 的通用数据采集系统方案，设计了主控板卡硬件电路，搭建了嵌入式 Linux 系统采集平台，实现了 16 路精确定时模拟量采样、16 路数字量采样，具备 RS232/485 接口、以太网通信功能，开发的系统软件运行稳定，客户端软件具备良好人机交互界面，扩展的手机 APP 软件功能完善，组建完成的系统功能完善、运行稳定，在工农业生产监控、地质水文环境监测、森林灾害预防等领域有着广阔的应用前景。

参考文献：

[1] 柯新宇. 基于 ARM 的数据采集卡研制 [D]. 武汉：华中科技大学，2008.
[2] 韩雪川. 基于 ARM 嵌入式 Linux 的数据采集监控终端设计 [D]. 哈尔滨：哈尔滨工程大学，2010.
[3] 闫广续，袁纵横，等. 基于 ARM 嵌入式 Linux 的数据采集系统设计 [J]. 计算机测量与控制，2015，23 (5)：1724－1727.
[4] 陆文周. Qt5 开发及实例 [M]. 北京：电子工业出版社，2014.
[5] 王 森. 一种基于 Android 的远程控制工具的设计与实现 [D]. 西安：西安电子科技大学，2012.
[6] 王 峰，宣伯凯，等. 基于 Android 的家庭移动医疗监护系统的设计 [J]. 计算机测量与控制，2015，23 (5)：1586－1588.

(上接第 158 页)

[9] 吴 迪，曹 洁，王进花. 基于自适应高斯混合模型与静动态听觉特征融合的说话人识别 [J]. 光学精密工程，2013，21 (6)：1598－1604.
[10] 马义德，邱秀清，陈显莅，等. 改进的基于听觉掩蔽特性的语音增强 [J]. 电子科技大学学报，2008，37 (2)：255－25.
[11] 茅正冲，王正创，龚 熙. 一种低信噪比下的说话人识别算法研究 [J]. 计算机应用与软件，2014，31 (12)：218－220，251.
[12] Kitaoka N，Yamamoto K，KusamizuT，et al.. Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance [A]. Automatic Speech Recognition & Understanding [C]，Kyoto，Japan，2007：607－612.
[13] 李振静，王国胤，杨 勇，等. 基于谱熵噪声估计的改进谱减法 [J]. 计算机工程，2009，35 (18)：164－166.
[14] 杜志然，周 萍，景新幸，等. 基于谱熵的耳语音增强研究 [J]. 传感器与微系统，2012，31 (6)：69－72.
[15] 蒋 晔，唐振民. GMM 文本无关的说话人识别系统研究 [J]. 计算机工程与应用，2010，46 (11)：179－182.
[16] 赵立辉，毛 竹，霍春宝，等. 基于 GMM-SVM 的说话人识别系统研究 [J]. 工矿自动化，2014，40 (5)：49－53.
[17] 侯 珏，刘 轶，郑 方，等. 基于 VP 树结构的多层匹配算法在哼唱识别中的应用 [J]. 清华大学学报 (自然科学版)，2009，49 (S1)：1419－1424.