改进的多窗谱 MFCC 在说话人确认中的应用①

曾 祺[1], 甘 涛[1], 曾红斌[2]

[1](电子科技大学 电子工程学院, 成都 610000)
[2](湖南广播电台 技术管理部, 长沙 410007)

摘 要: 语音中存在加性噪声降低了MFCC参数的鲁棒性, 使得说话人确认系统性能下降. 多窗谱MFCC引入了多窗谱估计技术在增强 MFCC 特征的噪声鲁棒性上取得了一定效果, 但改善的程度有限. 为了使 MFCC 参数对噪声具有更强的鲁棒性, 提出了一种改进的多窗谱 MFCC 提取算法. 改进算法在多窗谱 MFCC 的基础上引入谱减思想, 谱减法(Spectral Subtraction, SS)能够增强语音并降低噪音的干扰. 因此, 采用了 Multitaper+SS 组合的改进算法融合了两者的优势, 具备了更好的性能. 仿真结果表明, 当测试语音中含有加性噪声时, 与多窗谱 MFCC 提取算法相比, 采用改进的多窗谱 MFCC 的说话人确认系统性能在等错误率 EER 和最小检测代价函数值 minDCF 两项评测指标上都取得了更好的结果.

关键词: MFCC; 多窗谱; 谱减法; 说话人确认; 加性噪声

Improved Mutitaper MFCC and its Application in Speaker Verification

ZENG Qi[1], GAN Tao[1], ZENG Hong-Bin[2]

[1](School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China)
[2](Technology Management Department, Radio Hunan, Changsha 410007, China)

Abstract: Speech exists additive noise which is the main reason for reduced MFCC robustness. Then reduced MFCC robustness makes the speaker verification system identification rate decreased. To solve the problem, multitaper technology is introduced, which is a little effective in improving the noise robustness of MFCC. To get a better result, we propose a improved multitaper MFCC extraction algorithm which combines multitaper and spectral subtraction methods. Multitaper technology improve the robustness of MFCC, then spectral subtraction method enhance the speech and reduce the effect of noise. Simulation results show that when the test speech contains the additive noise, new algorithm could achieve better result in EER and minDCF which are the evaluation index of speaker verification system.

Key words: MFCC; multitaper; spectral subtraction; speaker verification; additive noise

特征参数的提取是说话人确认系统的重要组成部分, 对系统性能的优劣有重大影响. 理想的特征参数应当在各种复杂条件下具有较高的鲁棒性, 从不同说话人的语音提取出的特征参数差异应该比较大, 而从同一说话人发出的不同语音中提取出的特征参数应当比较相似. 短时谱特征是当前说话人确认研究领域应用最为广泛的特征, 其中最具代表性的是梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)和线性预测编码系数(Linear Prediction Coding Coefficient, LPCC). MFCC 是基于人耳听觉特性的特征参数, 有较好的区分度. 本文将重点对 MFCC 的提取算法展开研究, 并将其运用到说话人确认系统中, 通过说话人确认系统的性能指标来衡量 MFCC 区分度的强弱.

当测试语音纯净不含噪时, 采用传统 MFCC 提取算法的说话人确认系统一般也能达到较高的识别率. 然而, 当测试语音含有噪声时, 说话人确认系统的性能出现明显下降. 针对语音中存在加性噪声降低 MFCC 的鲁棒性, 从而使得说话人确认系统性能下降

① 收稿时间:2014-03-03;收到修改稿时间:2014-04-09

的问题, 已经有许多 MFCC 提取的改进算法. 这些有效的改进技术包括倒谱均值与方差规整[1](cepstral mean and variance normalization, CMVN), RASTA 滤波[2], 特征弯折(feature warping)[3]和 MVA 处理[4], 它们都曾被广泛的采用来提高 MFCC 的鲁棒性,然而它们都存在一个很大的缺点就是需要延迟处理. 本文首先对具有实时性的多窗谱 MFCC(Multitaper MFCC Features)[5]进行了研究. 多窗谱 MFCC 具有低方差性, 从统计学的角度来看, 谱估计值的低方差性提升了基于加窗离散傅里叶变换 MFCC 的性能[6]. 当测试语音含有加性噪声时, 与传统 MFCC 相比, 采用多窗谱 MFCC 的说话人确认系统性能有一定的提高, 但性能提升幅度不大. 为了进一步提高 MFCC 的噪声鲁棒性, 本文提出了一种多窗谱 MFCC 的改进算法. 该算法在多窗谱技术(Multitaper)的基础上, 引入了谱减法(Spectral Subtraction, SS). 由于谱减法能够增强语音并降低噪音的干扰, 采用 Multitaper+SS 组合改进后的 MFCC 应当具有更好的噪声鲁棒性.

1 多窗谱MFCC

MFCC 提取过程中一个很大的特色就是滤波处理时采用了梅尔频率刻度. 梅尔频率代表着一般人耳对于频率的感受度. 在低频部分, 人耳感受比较敏锐; 在高频部分, 人耳的感受就会越来越粗糙. 梅尔频率与赫兹频率的对应关系不是简单的线性关系. 研究表明, 当频率低于 1000Hz 时, 梅尔频率与赫兹频率对应关系近似为线性; 当频率高于 1000Hz 时, 二者的对应关系则近似为对数关系. 两者之间的这种对应关系可以用式(1)或(2)近似表示.

$$mel(f) = 2595 * \log_{10}(1 + f / 700) \quad (1)$$

$$mel(f) = 1125 * \ln(1 + f / 700) \quad (2)$$

1.1 传统 MFCC

MFCC 的提取过程一般都会包含以下步骤:

1)预加重: 目的是为了消除发声过程中声带和嘴唇的效应, 从而补偿语音信号受到发音系统所抑制的高频部分, 突显在高频的共振峰. 如式(1.3)所示, 其中介于 0.9 和 1.0 之间.

$$y(n) = s(n) - a * s(n-1) \quad (3)$$

2)分帧: 将 N 个采样点集合成一个观测单位, 称为一帧, 通常一帧的时间约为 10~30ms, 根据采样率可以计算出对应的 N 的大小. 为了避免相邻两帧变化过大, 一般我们会让相邻两帧之间有一段重叠区域, 此重叠区域包含了 M 个采样点, 通常 M 的值约是 N 的一半或三分之一, 我们称这样的重叠区域为帧迭.

3)汉明窗: 目的在于增加窗边界处信号的连续性, 减小吉伯斯效应. 如公式(4)所示.

$$w(n,\alpha) = (1-\alpha) - \alpha\cos(2\pi n/(N-1)), 0 \le n \le N-1 \quad (4)$$

4)离散傅里叶变换: 将时域信号转换成频域上的能量分布更易于观察信号的变化特性, 不同的能量分布, 代表不同的语音特性. 如公式(5)所示.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}, 0 \le k \le N \quad (5)$$

5)三角带通滤波: 三角带通滤波器有两个主要目的: 一是对频谱进行平滑化, 并消除谐波的作用, 突显原先语音的共振峰; 二是降低数据量的维数. 如公式(6)所示.

$$weigh(k) = \begin{cases} 0 & k < f(i-1) \ \text{or} \ k > f(i+1) \\ \dfrac{2(k-f(i-1))}{(f(i+1)-f(i-1))(f(i)-f(i-1))} & f(i-1) < k < f(i) \\ \dfrac{2(f(i+1)-k)}{(f(i+1)-f(i-1))(f(i+1)-f(i))} & f(i) \le k \le f(i+1) \end{cases} \quad (6)$$

6)离散余弦变换: 采用 DCT 变换是期望能转回类似时域, 又称为倒谱域. 因为之前采用梅尔频率来转换, 故称之为梅尔频率倒谱. 如公式(7)所示.

$$C_m = \sum_{k=1}^{N} \cos[m*(k-0.5)*\pi/N] * E_k, m = 1,2,\cdots,L \quad (7)$$

1.2 多窗谱 MFCC 的实现

在语音处理中, 经常要进行谱估计, 而谱估计时几乎都会采用加窗离散傅里叶变换. 用 $X = [x(0),\cdots,x(N-1)]^T$ 表示一帧含有 N 个采样点的语音数据, 则加窗傅里叶变换可用式(8)表示.

$$\hat{S}(f) = \left| \sum_{t=0}^{N-1} w_j(t) x(t) e^{-i2\pi tf/N} \right|^2 \quad (8)$$

其中, $w_j(t) = [w_j(0),\cdots,w_j(N-1)]^T$ 是时域下的窗函数, 该窗函数以单位帧为边界, 具有对称性.

多窗谱 MFCC 正是对加窗离散傅里叶变换方法进行改进, 从而提高 MFCC 的鲁棒性. 前面已经详细介绍了传统 MFCC 的提取步骤, 在离散傅里叶变换进行谱估计前对时域信号进行了加窗处理, 传统的 MFCC 提取算法一般采用汉明窗进行一次加窗, 这里的加窗处理减少了谱泄露. 当前还存在的问题是, 这种传统的加窗处理得到的谱估计值方差仍然较大, 而谱估计

值的高方差性降低了基于加窗离散傅里叶变换 MFCC 的性能. 为了进一步减小谱估计值的方差值, 提高 MFCC 的鲁棒性, 文献[5]采用多窗频谱估计[7]对传统 MFCC 进行了改进, 提出了多窗谱 MFCC(Multitaper MFCC), 多窗谱MFCC具有低方差性和更好的鲁棒性. 多窗谱 MFCC 的提取过程如图 1 所示.

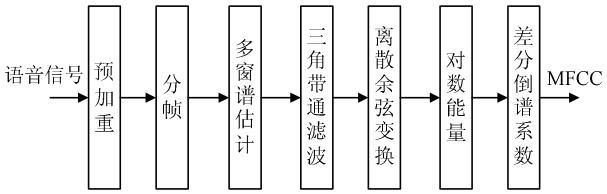

图 1　多窗谱 MFCC 提取过程框图

上图中多窗谱估计过程可简单地用式(9)表示.

$$\hat{S}(f)=\sum_{j=1}^{K}\lambda(j)\left|\sum_{t=0}^{N-1}w_j(t)x(t)e^{-i2\pi tf/N}\right|^2 \tag{9}$$

其中 K 表示窗的个数, $\lambda(j)$ 为加权系数, $w_j(t)=[w_j(0),\cdots,w_j(N-1)]^T$ 为加窗函数, $j=1,\cdots,K$. 由此可见传统的加窗傅里叶谱估计其实就是当 $\lambda=K=1$, 并选用汉明窗时的特殊情况. 多窗谱技术的精髓是采用了多次加窗处理.

2　算法改进

本文提出的改进算法正是在多窗谱 MFCC 的基础上进一步引入了谱减法, 形成了 Multitaper+SS 的组合改进法, 改进的多窗谱 MFCC 提取过程如图 2 所示.

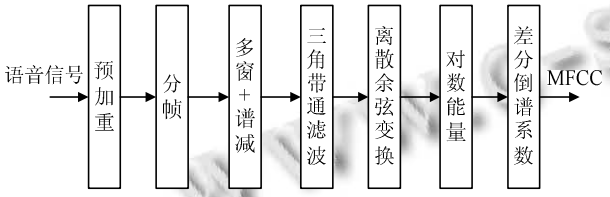

图 2　改进的多窗谱 MFCC 提取过程框图

一般认为语音数据的分布符合非平稳随机过程, 然而当取 $10\sim30$ ms 的语音帧数据进行分析时, 又可以认为其具有短时平稳性. 假定短时平稳的语音信号与加性噪声信号相互独立, 谱减法(Spectral Subtraction, SS)的基本思路就是用含噪语音信号的功率谱减去噪声信号的功率谱, 从而得到纯净语音的功率谱. 在这里我们假设语音信号 $y(k)$ 含有加性噪声, 用 $s(k)$ 表示纯净的语音信号, 用 $n(k)$ 表示加性噪声, 显然它们之间满足式(10)所示关系.

$$y(k)=s(k)+n(k) \tag{10}$$

假设 $y(k)$、$s(k)$、$n(k)$ 三者经过傅里叶变换后依次为 Y_k、S_k、N_k, 则有如下关系式.

$$Y_k=S_k+N_k \tag{11}$$

转换式(11)可得:

$$S_k=Y_k-N_k \tag{12}$$

式(12)表明, 只要能估计出各频率的噪声谱值, 我们就能用当前含噪语音谱值减去噪声谱值, 从而得到纯净语音的谱值估计.

在介绍多窗谱 MFCC 时, 我们提到多窗谱 MFCC 进行了多次加窗傅里叶变换, 每次加窗傅里叶变换得到的谱估计值乘以对应的权重, 将多次加窗傅里叶变换的结果之和作为最后各频率点的谱估计值. 因而, 本文的谱减法改进就是在每次加窗傅里叶变换得到谱估计后进行帧局部谱减. 这种帧局部谱减采用一帧数据频谱估计的最小值作为各频率点的噪声谱估计值, 将此估计值作为谱减值, 由公式(12)对一帧数据的所有频谱值进行谱减. 用 $X_{\min}$ 表示一帧数据频谱估计的统计最小值, 见式(13). 则多窗谱减频谱估计可用式(14)表示.

$$X_{\min}=\min\{X_i\},0\le i\le N-1 \tag{13}$$

$$\hat{S}(f)=\sum_{j=1}^{K}\lambda(j)\left(\left|\sum_{t=0}^{N-1}w_j(t)x(t)e^{-i2\pi tf/N}\right|^2-X_{\min}\right) \tag{14}$$

改进算法结合了多窗谱技术与谱减法两者的优势, 使得改进后的 MFCC 比多窗谱 MFCC 具有更好的鲁棒性, 在加性噪声测试中有了更好的表现.

3　实验与结果分析

3.1　评测指标

说话人确认系统的评估主要采用 2 个指标: 等错误率 EER 和最小检测代价函数值 minDCF. EER 和 minDCF 是分析错误接受率 R_{FA} 和错误拒绝率 R_{FR} 后所得结果. 两种错误率及 minDCF 分别定义如下:

$$R_{FA}=\frac{n_{FA}}{n_{imp}} \tag{15}$$

$$R_{FR}=\frac{n_{FR}}{n_{tar}} \tag{16}$$

$$\min DCF = \min\{C_{FR}P_{tar}R_{FR} + C_{FA}P_{non}R_{FA}\} \quad (17)$$

其中，n_{imp} 表示冒认者测试语音段总数，n_{FA} 表示错误接受次数；n_{tar} 表示目标说话人测试语音段总数，n_{FR} 表示错误拒绝次数；C_{FA} 和 C_{FR} 分别表示 R_{FA} 和 R_{FR} 的代价，P_{non} 和 P_{tar} 分别表示冒认者的先验概率和目标说话人的先验概率.

3.2 实验数据

实验采用由 MIT、TI 和 Stanford Research Institute International 等多个机构共同开发的 TIMIT 标准数据库, 其中训练背景模型采用 TIMIT 训练部分中的 dr4 所有女性的语音数据, 训练冒认者模型采用 TIMIT 测试部分中的 dr4 后 10 位女性的语音数据, 测试部分中的 dr4 前 6 位女性语音数据作为测试语音, 这 6 人每人有十段语音, 前 5 段用于训练说话人模型, 后 5 段进行定量加噪后作为测试语音. 我们将使用开源工具包 FaNT(Filtering and Noise Adding Tool)[8]对测试语音进行定量加噪. 噪声文件将从网站 FreeSound[9]中免费获得. 对干净的测试语音定量加噪过程是随机的, 信噪比分别为 20dB、10dB、0dB 和-10dB.

3.3 实验结果

为了验证本文提出的改进算法比文献[5]提出的多窗谱 MFCC 在噪声环境下具有更好的鲁棒性, 下面对传统 MFCC、多窗谱 MFCC、以及本文的改进 MFCC 进行加噪测试实验, 通过在说话人确认系统上的表现进行验证. 本文的实验平台为基于 GMM-UBM 的说话人确认系统, 该系统在纯净语音测试上最好性能达到了 EER 和 minDCF 分别为 3.33%和 2.67%. 分别采用上述 3 种 MFCC 特征进行加噪测试实验, 实验结果如表 1、表 2 所示.

表 1 EER 评测结果

MFCC 提取算法	EER(%)			
	20dB	10dB	0dB	-10dB
传统	24.00	33.33	40.00	43.33
Multitaper	17.33	30.61	38.52	42.10
Multitaper+SS	16.67	26.67	33.33	39.30

表 2 minDCF 评测结果

MFCC 提取算法	minDCF(%)			
	20dB	10dB	0dB	-10dB
传统	19.00	31.67	37.00	43.00
Multitaper	15.67	29.00	35.00	42.33
Multitaper+SS	14.67	23.00	28.00	38.00

从以上实验结果可以看出, 采用 TIMIT 作为加噪测试语料库时, 说话人确认系统未能达到比较理想的性能. 随着测试语音信噪比的降低, 系统性能下降幅度较大. 但相比较而言, 多窗谱 MFCC 的性能好于传统 MFCC, 而本文提出的改进 MFCC 又比文献[5]提出的多窗谱 MFCC 在 EER 和 minDCF 两项评价指标表现更好, 当信噪比依次为 20dB、10dB、0dB、-10dB 时, EER 的表现依次提高 3.8%、12.9%、13.5%、6.7%, minDCF 的表现依次提高 6.4%、20.7%、20.0%、10.2%, 从中可以看出当测试语音加噪程度不是特别严重时, 改进算法较之多窗谱 MFCC 性能上有较大提高, 因而可以说本文的改进是有成效的.

4 结语

针对加性噪声降低说话人确认系统性能的问题, 本文提出了具有更强鲁棒性的 MFCC 提取算法. 首先介绍了传统的 MFCC 提取算法, 随后仿真实现了多窗谱 MFCC(Multitaper MFCC). 经过加噪测试实验, 测试结果表明多窗谱 MFCC 较之传统的 MFCC 具有更强的鲁棒性, 在 EER 和 minDCF 两项评价指标都有更好的表现. 本文提出的在多窗谱 MFCC 基础上引入谱减的方法, 进一步改善了多窗谱 MFCC 的性能, 取得了更好的 EER 和 minDCF 值.

参考文献

1 Huang X, Acero A, Hon HW, et al. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001.

2 Hermansky H, Morgan N. RASTA processing of speech. IEEE Trans. on Speech and Audio Processing, 1994, 2(4): 578–589.

3 Pelecanos J, Sridharan S. Feature warping for robust speaker verification. Proc. Speaker Odyssey: Speaker Recognition Workshop (Odyssey 2001). Crete, Greece. 2001. 213–218.

4 Chen CP, Bilmes JA. MVA processing of speech features. IEEE Trans. on Audio, Speech, and Language Processing, 2007, 15(1): 257–270.

5 Kinnunen T, Saeidi R, Sedlák F, et al. Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Trans. on Audio, Speech, and Language Processing, 2012, 20(7): 1990–2001.

6 Percival DB. Spectral analysis for physical applications. Cambridge University Press, 1993.

7 Wu CH, Hsieh CH. Multiple change-point audio segmentation and classification using an MDL-based Gaussian model. IEEE Trans. on Audio, Speech, and Language Processing, 2006, 14(2): 647–657

8 Filtering and Noise Adding Tool. http://dnt.kr.hsnr.de/download.html

9 FreeSound. http://www.freesound.org