基于基音频能值和梅尔参数的语音识别设计与实现

Design and Implementation of Speech Recognition by Mel – Cepstrum Integrated with Pitch

相　征　尹成俊　（安徽工程科技学院 安徽省电气传动与控制重点实验室安徽 安徽 芜湖 241000）

摘　要：根据语音发声特点，在分析了语音信号中的基音频率和梅尔参数之间的关系，论文提出一种在强噪声环境下提高语音识别率的实现方法，并对基于基音频能值和梅尔参数的方法和传统语音识别方法进行比较。实验结果表明该方法能够有效提高语音识别率，同时具有计算量小，实时性高，鲁棒性好等特点。

关键字：基音 频能值 梅尔参数 语音识别

1　引言

从感官和语音的发声特点上，人们往往从音长、音强、音高和音色四个方面来描述语音特征；而从声学特性上，通常用时长、幅度、基音频率和频谱来表征上述的语音特征[1]。基音频率描述音高，MFCC 参数是语音信号处理中描述语音频谱特征最常用参数，本文主要应用这两种参数在强噪声环境下提高语音的识别率。由于本论文主要进行在强噪声情况下，不同端点检测和特征参数的提取对语音识别率的影响，因此在模板匹配和模式识别过程中采用较为简单的改进 DTW 识别算法进行识别率比较。

2　基音频率和梅尔参数（MFCC）

2.1　基音频率

浊音是种准周期信号，因此浊音信号的周期为基音周期，其倒数即为基音频率。我们以 10KHZ 的采样率上看，基音周期的持续时间会从高音调的女性或儿童的约 20 个采样点变化到很低音调的男性的 250 个采样点，这就意味着人的语音信号基音频率主要分布在 40 – 500Hz，而男性和女性的声音在这个频率段上又呈现不同的频率范围，一般而言男性的基音频率分布在 60 – 200Hz，女性的基音频率分布在 200 – 450Hz[2]。如图 1，分别对不同男性和女性进行录音（去掉清音和静音），仅保留基音可以了解基音频率的分布范围，从图中可以看到人的基音频率变化规律和上述相同。

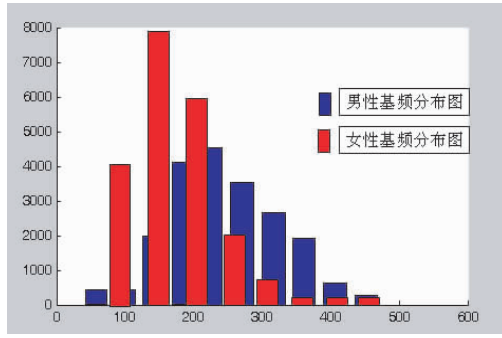

图 1　男性和女性的基音频率分布图

2.2　梅尔参数（MFCC）

人耳对不同频率的语音具有不同的感知能力，在 1000HZ 以下，感知能力与频率呈线性关系，而在 1000HZ 以上，感知能力与频率成对数关系，为了模拟人耳对不同频率语音的感知特性，人们提出了 Mel 频率概念，将频谱转化为基于 Mel 频率的非线性频谱，然后转换到倒谱域上[2]。由于充分模拟了人的听觉特性，而且没有任何前提假设，则 MFCC 参数具有识别性能和抗噪能力，实验证明在语音识别中 MFCC 参数性能明显优于早期的 LPCC 参数。

3　基于基音能频值的端点检测算法

为了在强噪声环境正确截取语音信号，滤除多余噪声，需要对语音信号加入端点检测。

3.1　基音频率的提取

基音频率提取采用 Rabiner 提出的基音检测算

法[1]。首先对语音数据进行 60 – 900Hz 的带通滤波，对原始信号进行滤波，截止频率取 900Hz，可以去掉大部分共振峰的影响，同时还可为频率低于 450Hz 以下的基音频率保留一两次谐波。然后取帧移 10ms，帧长 30ms，每帧语音信号进行中心消波，求取消波以后信号的自相关。然后从落入 60 – 500Hz 范围内的点找最大值 Rmax，如果 Rmax > 0.55R(0)，则基音周期为使 R(k) 最大值 Rmax 时的位置的 k 值。用语音的采样频率除以基音周期即为基音频率。对于 Rmax < 0.55R(0) 的语音帧，认为是清音或静音。通过验证，如图 2 发现图中红线区对应语音信号浊音，周期相应变化较慢，这就是基音的周期，而其他的部分周期变化快，对应的就是清音的部分，没有周期规律。

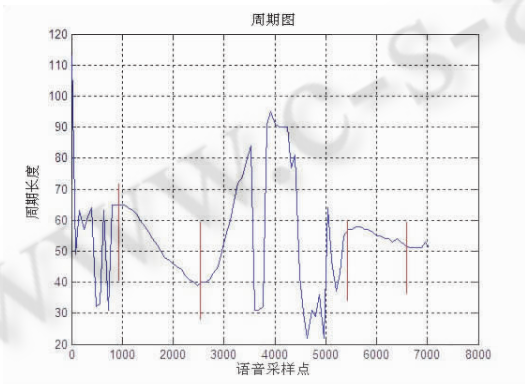

图 2　基音周期仿真

3.2 基音频能值算法的实现

该算法的具体步骤如下：

1) 首先对抽样频率为 10KHZ 的语音信号 x(n) 进行归一化，零点漂移校正，分帧取帧移 10ms，帧长 30ms 每帧数据首先用滤波器 $H(Z) = 1 - 0.98Z^{-1}$ 预加重处理，然后加 hamming 窗，定义语音信号最高频率为 f_{max}，语音能量为 SE，噪声能量为 NE，由上述基音频率提取方法确定语音信号基音频率范围为 $f_{starl} \sim f_{end}$，即 60 – 500Hz，则第 i 帧的基音能量为：

$$SE(i) = \sum_{f_{start}}^{f_{end}} SE(n) = \sum_{f_{start}}^{f_{end}} [x(m)\omega(n-m)]^2 \quad (1)$$

$$NE = \frac{1}{10}\sum_{i=1}^{10}\sum_{f_{start}}^{f_{end}} SE(n) \quad (2)$$

2) 确定自适应端点检测的判断阈值，确定条件：首先确定语音前 10 帧频率在 60 – 500Hz 的能量值的平均值为噪声能量 NE，设定进入语音段的阈值范围 TSL – TSH：

$$TSL = \alpha^* NE; \qquad TSH = \beta^* NE; \quad (3)$$

根据对大量不同噪声环境下，如工业现场噪声，交通噪声（如 f16 等），建筑噪声等语音信号进行大量的实验，可以得出：调整参数 α, β 的范围为：

$$\alpha \in [1.40, 1.80]; \ \beta \in [1.01, 1.09]; \quad (4)$$

由于短时语音信号的分析中认为语音信号是平稳的，但是仍需要对噪声能量进行实时更新，使整个阈值的判断具有自适应性，同时可以提高端点检测的精确性。

$$\begin{cases} NE = \xi^* NE + (1-\xi)^* SE(i) & SE(i) < TSL \\ NE = \xi^* NE + (1-\xi)^* SE(i) & TSH > SE(i) \geqslant TSL \end{cases} \quad (5)$$

式中 ξ 和 ξ 为调节系数，笔者经过大量的实验得到经验值，取 $\xi = 0.1, \xi = 0.9$。

搜索基音信号时，有以下判断步骤：

① 从头开始检测基音频能值，若连续得到频能值低于 TSL，则认为噪声段，同时更新阈值。

② 若连续得到频能值 TSL < SE(i) < TSH，则认为检测的信号进入语音段（或为清音段）。

③ 若连续得到频能值高于阈值 TSH，前提下的后 3 – 5 帧小于 TSL，则证明仍然是噪声信号，与此同时进一步更新阈值。否则认为检测的信号为真正的语音信号。

3) 对真正的语音段进行优化判断，在得到各帧的频能值之后，对其进行中值滤波。

$$Vad(i) = \begin{cases} 0 & Vad(i) < 0.5 \\ 1 & Vad(i) \geqslant 0.5 \end{cases} \quad (6)$$

$$Vad(i) = \begin{cases} \text{语音帧} & Vad(i) = 1 \\ \text{非语音帧} & Vad(i) = 0 \end{cases} \quad (7)$$

值得注意的是，笔者在进行中值滤波的时候发现，由于分帧过程中帧长和帧移选取的不同，会导致真正的语音帧出现漏检，因此只要将 Vad(i) = 0.5 的情况算入进去就可以大大提高端点检测的准确性。整个基于基音频能值的端点检测算法如图 3 所示：

4 基于改进 MFCC 参数提取算法实现

由于语音信号数据量大，为了减小数据量，必须进行特征提取。目前比较有效的识别特征有 MEL 频率倒谱系数（MFCC），MFCC 参数符合人耳的听觉特性，

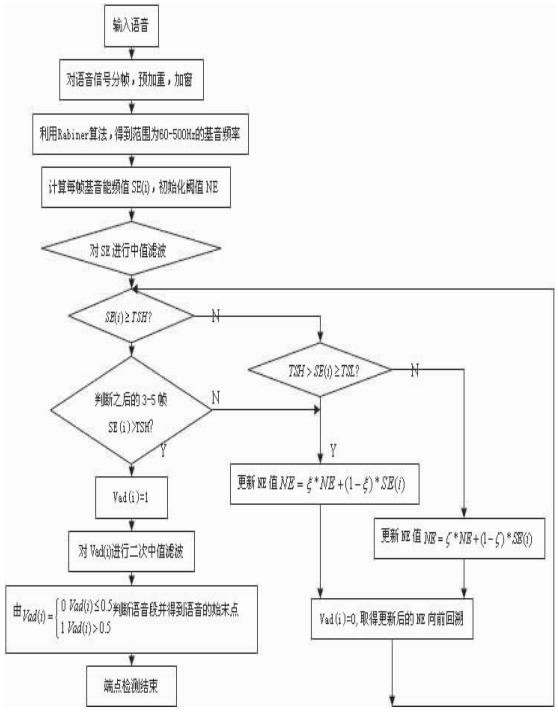

图 3　基于基音频能值的端点检测算法

而且在有信道噪声和频谱失真情况下表现比较稳健[3]。但对于 MFCC 来说主要体现了语音信号的静态特性，因此还可以用一阶差分系数和二阶差分系数近似描述语音信号的帧间相关性，反映语音信号的动态特征。经过研究表明，在特征参数提取中去掉二阶差分系数后，对于识别率的降低是很轻微即

对于模型精度影响不大，但是由此带来的是模型存储规模则大大的降低。

该算法的具体步骤如下：

① 首先计算经过预处理之后语音信号 x(n)(0≤n<N)的离散幅度谱，主要利用 FFT 算法。

$$|X(k)| = \sum_{n=0}^{N-1} x(n) * e^{-j*2\pi k/N} 0 \leqslant k < K \quad (8)$$

② 按照 *Bark* 刻度对频谱等分，并记中心频率为 $f(d)$，*Bank* 刻度和 *HZ* 刻度 f 间转换公式为：

$$b(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (9)$$

转化方法：将频域信号通过 24 个三角滤波器，其中中心频率在 $1000Hz$ 以上和以下的各 12 个。滤波器的中心频率间隔是在 $1000Hz$ 以下为线性分布，$1000Hz$ 以上为等比分布。输出为：

$$Y_i = \sum_{k=f_i-1(d)}^{f_i(d)} \frac{k - f_{i-1}(d)}{f_i(d) - f_{i-1}(d)} E_k + \sum_{k=f_i(d)+1}^{f_{i+1}(d)} \frac{f_{i+1}(d) - k}{f_{i+1}(d) + f_i(d)} E_k \quad (10)$$

其中 E_k 为频谱上第 k 个频谱点能量，Y_i 为第 i 个滤波器输出，$f_i(d)$ 为第 i 个滤波器中心频率。

③ 用离散余弦变换将滤波器的输出变换到倒频谱：

$$C_k = \sum_{j=1}^{24} log(y_j) cos(k(j\frac{1}{2})\frac{\pi}{24}) k = 1,2,\cdots P \quad (11)$$

其中 p 为阶数，在本文中取 P = 12，而 C_k 即为所求的 MFCC 参数。

④为了更好的描述语音信号的动态特性，加入一阶差分倒谱为：

$$\triangle c_1(m) = \sum_{n=-2}^{2} ic_{1-n}(m) \quad 1 \leqslant m \leqslant P \quad (12)$$

其中 l 表示第 l 帧，m 表示第 m 维。

⑤ 对 MFCC 参数和一阶差分 MFCC 参数做均值归一化，得到最终的强鲁棒性的 MFCC 参数。

5　实验结果与分析

5.1　实验中语音模板和测试模板的建立

实验中选用的语音数据来自于：共有 10 人（五男五女），共有 10 个不同的词语，每人每个词语录制 10 遍。分别在日常的工作环境中进行录音，并认为的在录音中加入不同类型的噪声，例如工厂噪声，办公室噪杂的噪声，交通噪声等。在测试时将模板库分为参与训练者的模板和未参与训练者模板两部分，其中参与训练者样本集为 3 男 3 女，未参与训练者样本集为 2 男 2 女。在参与训练者模板，从每人每个词语 10 遍的语音模板中任取 5 个作为训练模板，其余 5 个连同未参与训练者一起作为测试模板。模板匹配和模板识别均采用改进的 DTW 算法。

5.2　各种算法识别检测结果比较与分析

在仿真各种算法的过程中分别选用了三种端点检测方法和三种常用的特征提取方法：

① 基于语音信号能零率的端点检测和 LPCC 特征参数的提取；②基于带噪语音信号方差法端点检测和 MFCC 特征参数的提取；③本论文提出的基于基音能频值端点检测和差分 MFCC 特征参数的提取；为了说明端点检测算法的优越性，如图 4 分别给出了三种不同的端点检测算法对

同一词语在低信噪比下的检测仿真图。如表1－表3给出了在不同信噪比下各种端点检测的检测结果。

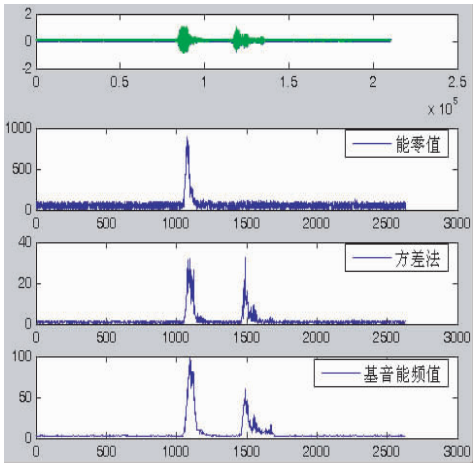

图4　同一词语信噪比在5db左右时三种端点检测算法仿真图

表1　噪声类型为办公噪声时同一词语不同信噪比下三种端点检测算法的帧范围

算法类型:10db 语音帧范围	5db 语音帧范围
能零率 1043－1160,1480－1655	1052－1147, /
方差法 1056－1158,1478－1673	1056－1143,1480－1669
基音法 1058－1156,1465－1681	1060－1148,1163－1682

表2　噪声类型为交通噪声时同一词语不同信噪比下三种端点检测算法的帧范围

算法类型:10db 下语音帧范围	5db 下语音帧范围
能零率 1036－1142, /	1045－1163, /
方差法 1057－1162,1487－1678	1068－1163,1484－1675
基音法 1058－1160,1470－1682	1062－1158,1471－1691

表3　噪声类型为工厂噪声时同一词语不同信噪比下三种端点检测算法的帧范围

算法类型:10db 下语音帧范围	5db 下语音帧范围
能零率 / ,	/ / , /
方差法 1068－1135,1492－1663	1072－1145,1489－1663
基音法 1062－1140,1469－1695	1068－1149,1478－1697

利用改进 DTW 算法分别对不同算法得到的结果进行识别，如表4可以发现各种算法所完成的识别时间也是不同的，在强噪声环境下，本论文所提出的算法更能较好的符合孤立词识别时实时性的要求。笔者同时做了利用改进 DTW 算法的孤立词识别率的统计，如表5，在办公噪声下且信噪比为 10db 下不同算法的精确度，显然基于基音能频值和差分 MFCC 的方法的识别率明显高于传统的识别方法。

表4　各种算法执行时间

算法名称	执行时间(单位:秒)
能零值＋LPCC 算法	3.3120
方差法＋MFCC	3.1250
基音能频值＋差分 MFCC	2.9540

表5　信噪比10db下各种识别方法的识别率

算法(SNR＝10db)	参与训练者	未参与训练者	平均识别率
能零值＋LPCC 算法	84.34%	72.51%	78.43%
方差法＋MFCC	95.23%	90.02%	92.63%
基音法＋差分 MFCC	97.69%	95.26%	96.48%

6　结论与展望

实验证明，基于基音频能值和梅尔参数的算法不仅可以有效的提高强噪声环境下语音信号的识别率，同时还具有计算方便，实时性高，鲁棒性好的特点，因此对于语音识别系统的开发和设计也给予了可参考的价值。但是语音信号识别的研究是一项极其复杂而艰巨的工作，它不仅依赖于人类对语音信号本身的认识和探索程度，还依赖于生理学、心理学、通信科学、计算机科学等相关领域的发展情况。因此对于如何提出在强声环境下，提高语音识别系统的识别率和鲁棒性的好的算 法将对于语音识别的发展有重大意义[3]。

参考文献

1　韩纪庆，张磊，郑铁然. 语音信号处理. 北京：清华大学出版社. 2004，9.

2　蔡莲红，黄德智，蔡锐. 现代语音技术基础与应用. 北京：清华大学出版社，2005.

3　杨行峻，迟惠生等. 语音信号数字处理. 北京：电子工业出版社，1995.

4　郭继云，王守觉，苑海涛. 一种基于频能比的端点检测算法. 计算机工程与应用，2004，(3).

5　J C Tunqua，B Mak. B Reaves，A robust algorithm for word boundary detection in the presence of noise. IEEE Tran on Speech and Audio Processing，1994，2(3)；406－412.