基于 MFCC 的语音情感识别

韩　一[1],王国胤[1,2],杨　勇[1,2]
(1. 重庆邮电大学 计算机学院,重庆 400065;2. 西南交通大学 计算机学院,成都 610031)

摘　要:情感语音中携带着丰富的信息,在人机交互领域有着广阔的应用。Mel 频率是基于人耳听觉特性提出来的,它与 Hz 频率成非线性对应关系。Mel 频率倒谱系数(MFCC)则是利用它们之间的这种关系,计算得到的 Hz 频谱特征,MFCC 已经广泛地应用在语音识别领域。由于 Mel 频率与 Hz 频率之间非线性的对应关系,使得 MFCC 随着频率的提高,其计算精度随之下降。因此,在应用中常常只使用低频 MFCC,而丢弃中高频 MFCC。针对该问题进行了研究,修正了 Hz-Mel 非线性对应关系,提升了中高频系数的计算精度,并将其作为低频 MFCC 的补充,应用到语音情感识别中。实验证明,改进之后的算法与经典算法比较,在不同的特征组合上识别率都有不同程度的提高,从而证明了 Mid MFCC 特征计算方法的有效性。

关键词:MFCC;语音情感识别;情感计算
中图分类号:TP391.42　　**文献标识码**:A　　**文章编号**:1673-825X(2008)05-0597-06

Speech emotion recognition based on MFCC

HAN Yi[1], WANG Guo-yin[1,2], YANG Yong[1,2]
(1. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China;
2. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, P. R. China)

Abstract: Emotion speech carries rich information, which is widely used in the human-computer interaction (HCI). Mel-frequency is proposed based on the human auditory characteristics, and it is nonlinearly corresponded with Hz-frequency. Mel-frequency cepstral coefficients (MFCC) is one kind of Hz spectral characteristics; MFCC is calculated based on the nonlinear relationship between Mel-frequency and Hz-frequency and has a wide application in the speech recognition area. But because of such nonlinear relationship, the accuracy of MFCC reduces as the frequency increases. Hence, low MFCCs are usually used and high MFCCs are discarded in applications. This paper analyses this problem and proposes an improved algorithm by amending the nonlinear relationship to improve the accuracy of high MFCCs which are the complementary features to low MFCCs for emotion speech recognition. The experiment result proves that the recognition rate of improved algorithm increases compared to the classical algorithm, and the proposed Mid MFCC is effective.

Key words: Mel-frequency cepstral coefficients (MFCC); emotion speech recognition; affective computation

0　引　言

语言是人类交流的重要工具,人类的话语中不仅包含了文字符号信息,而且还包含了感情信息。对语音情感信息处理,在信号处理和人工智能领域中具有重要意义。目前常用的方法是对情感语音信号进行全局分析,提取基音频率(pitch)、振幅能量(energy)、语速(rate)、共振峰(formant)等参数,分析这些特征参数的时序和分布特点,找到不同情感语音的韵律规律,作为情感识别的依据。从信号分析的角度来看,语音信号是由众多不同频率的信号重叠在一起组成的,分析信号的频谱特性同样有助于情感识别的研究。Mel 倒谱频率系数(Mel-frequency cepstral coefficients, MFCC)是基于人耳的听觉特性提出的[1],它采用一种非线性的频率单位(Mel 频率)来模拟人的听觉系统。近年来,国内外

收稿日期:2008-02-22　修订日期:2008-07-20
基金项目:新世纪优秀人才支持计划和重庆市自然科学基金(CSTC2007BB2445);重庆市计算机网络与通信技术重点实验室开放课题基金"情感识别的关键技术研究"

学者做了很多相关研究工作,将 MFCC 应用到语音情感识别中。蒋丹宁[2]研究了声学特征参数的统计特征和时序特征,将 MFCC 作为频谱特征应用到情感识别中,通过特征融合之后的识别率达到了 92.9%。T. L. Pao[3]将 MFCC 和 LPCC 混合在一起作为识别参数,使用加权的 D-KNN 作为分类器,对 3 400 句样本组成的情感语音库进行了测试,最高的识别率达到 79.55%。Thao Nguyen[4]在 SUSAS 情感语音库的基础上,研究了基音周期、能量、语速、共振峰、通带带宽、MFCC 等特征参数,最后选取前 2 阶 MFCC 的统计特征,与其他韵律特征参数一起组成情感识别特征向量,结合支持向量机(SVM)和决策树(Decision Tree)的方法,得到平均识别率为 71%。

MFCC 在低频区域具有很好的频率分辨率,对噪音的鲁棒性也很好,但中、高频系数精度就不尽如人意。所以,大多只使用低阶 MFCC 作为识别参数,或者与韵律特征进行混合识别,而舍弃中高阶 MFCC。本文针对 MFCC 的不足,提出了改进算法,以 MFCC 的经典算法为基础,对 Hz-Mel 的非线性关系进行了修正,引入了 2 个新的系数:MidMFCC 和 IMFCC。它们分别在中、高频区域具有很好的计算精度,可作为低阶 MFCC 的补充,对全频域的频谱特征进行计算。实验结果表明,改进之后的算法提高了语音情感识别率。

1 信号预处理

1.1 预加重

语音信号的平均功率谱受声门激励和口鼻辐射影响,高频端大约在 800 Hz 以上按 6 dB/倍频程跌落,即 6 dB/oct(2 倍频)或 20 dB/dec(10 倍频)。求语音信号频谱时,频率越高,相应的成分越小。为此,要在预处理中进行预加重处理(Pre-emphasis)。预加重的目的是使信号的频谱变得平坦,保持从低频到高频的整个频带中,能用同样的信噪比求频谱,以便于频谱分析或声道参数分析。预加重一般是采用一阶的数字滤波器 μ : $H(Z) = 1 - \mu z^{-1}$,μ 值接近于 1,或者采用公式 $y(n) = x(n) - \alpha x(n-1)$,其中,$x(n)$ 为原始信号序列;$y(n)$ 为预加重后序列;α 为预加重系数。

1.2 窗函数

语音信号是一种典型的非平稳信号,处理中一般使用窗函数截取其中一段来进行分析,截取出来的那部分信号被认为是短时平稳的。加窗处理的另一个作用就是消除由无限序列截断导致的 Gibbs 效应。常见的窗函数有:

1)矩形窗(Rectangular Window)

$$w(n) = \begin{cases} 1 & (0 \leqslant n \leqslant N-1) \\ 0 & \text{其他} \end{cases} \tag{1}$$

2)汉明窗(Hamming Window)

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & (0 \leqslant n \leqslant N-1) \\ 0 & \text{其他} \end{cases} \tag{2}$$

3)哈宁窗(Hann Window)

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left(\dfrac{2\pi n}{N-1}\right) & (0 \leqslant n \leqslant N-1) \\ 0 & \text{其他} \end{cases} \tag{3}$$

汉明窗和哈宁窗都属于广义升余弦函数,通过分析他们的频率响应幅度特征,可以发现,矩形窗的谱平滑性能好,但是旁瓣太高,容易造成频谱泄露,损失高频成分;哈宁窗衰减太快,低通特性不平滑;汉明窗由于其平滑的低通特性和最低的旁瓣高度而得到广泛的应用。它们的频率响应如图 1 所示。

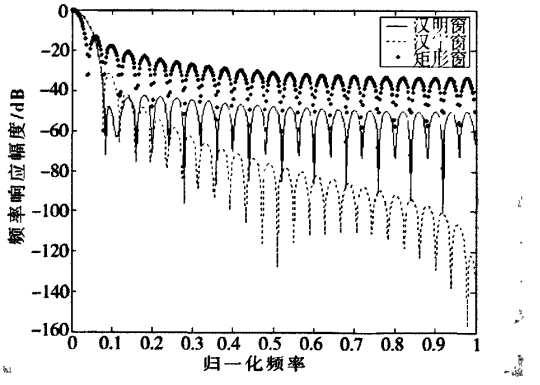

图 1 窗函数频率响应图

Fig. 1 Frequency response of window function

2 Mel 倒谱系数特征表示

2.1 Mel 倒谱系数提取

人耳对不同频率的语音具有不同的感知能力,是一种非线性的关系。结合人耳的生理结构,运用对数关系来模拟人耳对不同频率语音的感知特性,Davies 和 Mermelstein 于 1980 年提出了 Mel 频率的概念[1]。其意义为 1 Mel 为 1 000 Hz 的音调感知程度的 1/1 000。Hz 频率 f_{Hz} 与 Mel 频率 f_{Mel} 之间的转换关系如公式(4)。

$$f_{\text{Mel}} = 1\ 127 \times \ln\left(1 + \frac{f_{\text{Hz}}}{700}\right) \tag{4}$$

Mel 倒谱频率系数(MFCC)是在上述 Mel 频率概念基础上提出的,其计算流程如图 2 所示。

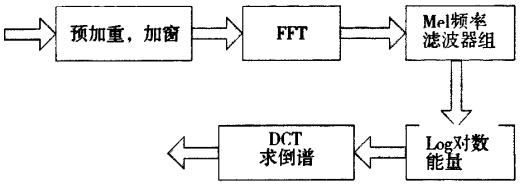

图 2　MFCC 计算流图

Fig. 2　Calculation flow of MFCC

1)将原始语音信号预加重,分帧加窗之后得到一帧语音信号。

2)对一帧语音信号进行快速傅立叶变换(fast fourier transform,FFT),得到信号的离散功率谱 $X(k)$。

3)定义一个由 M 个三角型带通滤波器组成的滤波器组,每个滤波器的中心频率为 $f(m)$,$m = 1, 2, \cdots, M$,相邻滤波器交叉重叠在一起,且其中心频率在 Mel 频率轴上为等间距分布,滤波器组在频域上覆盖从 0 Hz 到 Nyquist 频率,即采样率的二分之一。三角滤波器的中心频率 $f(m)$ 和频率响应 $H(k)$ 分别为

$$f(m) = \frac{N}{F_s} B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right) \tag{5}$$

$$H_i(k) = \begin{cases} 0, k < f(i-1) \cup k > f(i+1) \\ \dfrac{2(k - f(i-1))}{(f(i+1) - f(i-1))(f(i) - f(i-1))}, \\ \quad (f(i-1) \leqslant k \leqslant f(i)) \\ \dfrac{2(f(i+1) - k)}{(f(i+1) - f(i-1))(f(i+1) - f(i))}, \\ \quad (f(i) \leqslant k \leqslant f(i+1)) \end{cases} \tag{6}$$

(5),(6)式中:f_1 和 f_h 分别是滤波器组覆盖范围的低通频率和高通频率;F_s 是信号采样频率,单位都是 Hz;M 是滤波器组中滤波器的个数;N 是进行 FFT 变换时的点数;$B^{-1}()$是公式(4)的反函数。

$$B^{-1}(b) = 700(e^{b/1127} - 1) \tag{7}$$

4)通过步骤 3),每个滤波器产生输出频谱能量,取对数之后便得到一组如下系数 $S(m)$ 为

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), m = 1, 2, \cdots, M \tag{8}$$

再经过离散余弦变换(DCT)将 $S(m)$ 转换到时域,就是 MFCC。MFCC 系数 $c(i)$ 的计算过程为

$$c(m) = \sum_{n=0}^{M-1} S(m)\cos\left(\frac{\pi m(n+0.5)}{M}\right), 1 \leqslant m \leqslant M \tag{9}$$

MFCC 的 Hz-Mel 尺度对应的曲线和滤波器组分布如图 3 所示。

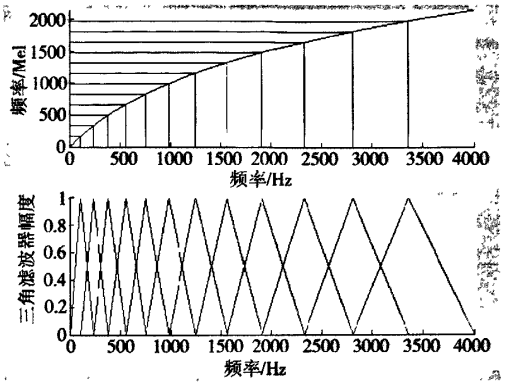

图 3　MFCC 尺度对应曲线和滤波器幅度分布图

Fig. 3　MFCC scale curve and distribution of fliter rany

2.2　Mel 倒谱系数改进——IMFCC 和 MidMFCC

MFCC 用数学的方法来模拟人耳的听觉特性,使用一串在低频区域交叉重叠排列的三角型滤波器,捕获语音的频谱信息。如图 3 所示,由于 Hz-Mel 频率非线性的对应关系,滤波器在低频区域的分布比较密集,而在中、高频区域使用的滤波器数量较少,单个滤波器在该区域的通带带宽较大。导致了低频 MFCC 具有很好的计算精度,频率分辨率高;而中高频 MFCC 计算精度不足,频谱信息被弱化,例如第 1 阶滤波器的带宽约为 100 Hz,而第 12 阶滤波器则覆盖了 2 600 ~ 4 000 Hz 这个范围,带宽达到 1 400 Hz。因此,语音情感识别中,通常只使用低阶 MFCC 作为识别参数,而中高阶 MFCC 被舍弃。能否依照 MFCC 计算原理,选用合适的方法来提高中、高频的计算精度,以提高识别率,成为本文的研究重点。

从图 2 中不难发现,倒谱系数的计算精度问题是由 Hz-Mel 频率之间非线性对应关系导致的。改变这种度量尺度,使得在低频区域分布的滤波器个数减少,而在高频区域的滤波器个数增加,即随着频率的提升,相邻滤波器之间的距离逐步缩小,这就是 Inverted MFCC(IMFCC)[5] 的思路。这种方法与常用的 MFCC 恰恰相反,虽然低频区域系数的计算精

度降低了,但是在高频系数精度有了明显的提升。IMFCC 的 Hz-Mel 尺度对应关系为

$$f_{\mathrm{IMFCC}} = 2146.1 - 1127 \times \ln\left(1 + \frac{4000 - f_{\mathrm{Hz}}}{700}\right) \tag{10}$$

(10)式中:4 000 是带通滤波器组的高通截止频率;2146.1 是 4 000 Hz 对应的 Mel 尺度频率。IMFCC 的频率尺度对应和滤波器分布如图 4 所示。

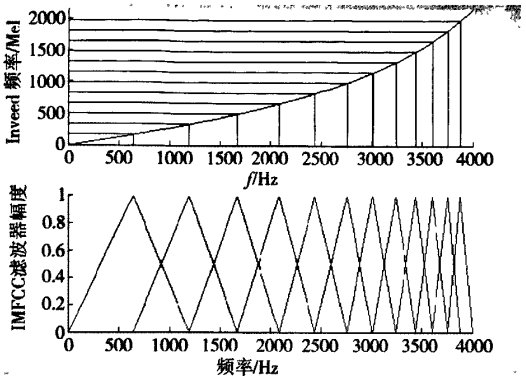

图 4 IMFCC 尺度对应和滤波器分布图

Fig. 4 IMFCC scale curve and distribution of filter

至此,MFCC,IMFCC 分别解决了低、高频段计算精度问题,但是中频区域的精度依然比较粗糙,单个滤波器的通带带宽在 800 Hz 左右。解决中频域的精度问题,就需要合适的、针对中频部分的 Mel-Hz 频率非线性关系。处理这种非线性关系时,仍然采用对数函数的形式,要求函数曲线在两端变化比较平缓,而中间变化剧烈,使得滤波器在低、高频区域分布稀疏,而在中频区域的分布密集,从而保证在中频段系数的计算精度。根据上述思想,本文提出一种旨在提高中频部分计算精度的 Mel 倒谱系数,称为 Middle MFCC(MidMFCC)。

参考 MFCC 和 IMFCC 的 Hz 频率与 Mel 频率之间的转换关系,MidFCC 非线性对应关系的函数曲线应当起始于点(0,0),终止于点(4 000,2 146.1),并且关于点(2 000,1 073.05)对称。即 MidMFCC 的对应关系在 0 ~ 2 000 Hz 这个区域类似于 IMFCC 的高频部分,在 2 000 ~ 4 000 Hz 这个区域类似于 MFCC 的低频部分,根据新的 Mel-Hz 频率转换关系设计对数关系式为

$$y = k \times \ln\left(1 + \frac{x}{p}\right) \tag{11}$$

对公式(11)取合适的 k,p 值,使得(11)式满足以下 3 个条件:

① MidMFCC 对应关系的低频部分要与 MFCC 对应关系的低频部分很好拟合;

② 当 $x = 2\,000$ 时,$y \approx 1\,073.05$,这样能使 MidMFCC 对应关系曲线起始于点(0,0),终止于点(4 000,2 146.1),并且关于点(2 000,1 073.05)对称;

③ 曲线在低频区域的斜率之和与高频部分的斜率之和的比值,接近与(4)式低频区域斜率之和与高频区域斜率之和的比值。

通过计算,具体参数:$k \approx 527$,$p \approx 300$,由此得到了(12)式。

$$y = 527 \times \ln\left(1 + \frac{x}{300}\right) \tag{12}$$

它的曲线和 Hz-Mel 曲线对比如图 5 所示。

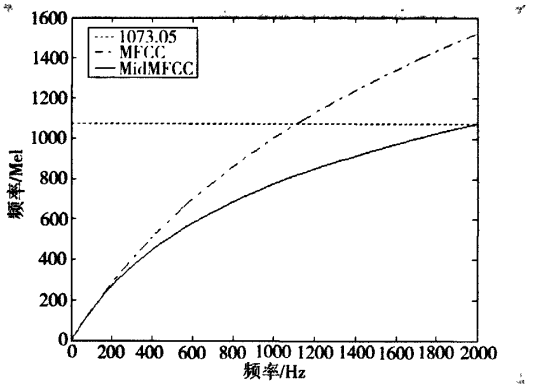

图 5 曲线对比图

Fig. 5 Curve comparison

在公式(12)的基础上,得到了 MidMFCC 的 Hz-Mel 尺度对应关系,如公式(13),这是一个分段函数。

$$f_{\mathrm{MidMFCC}} = \begin{cases} 1073.05 - 527 \times \ln\left(1 + \dfrac{2\,000 - f_{\mathrm{Hz}}}{300}\right), 0 < f_{\mathrm{Hz}} \leqslant 2\,000 \\ 1073.05 + 527 \times \ln\left(1 + \dfrac{f_{\mathrm{Hz}} - 2\,000}{300}\right), 2\,000 < f_{\mathrm{Hz}} \leqslant 4\,000 \end{cases} \tag{13}$$

MidMFCC 的频率对应关系和滤波器分布情况如图 6 所示。

基于 MFCC 原理提出的 MidMFCC 和 IMFCC 特征参数,解决了传统 MFCC 在语音识别应用中,中高阶参数计算精度不足,作为识别用特征参数可信度不高的问题。计算整个频域频谱特征,将中阶 MidMFCC 和高阶 IMFCC 作为低阶 MFCC 的补充,能够提高语音情感识别的效果。

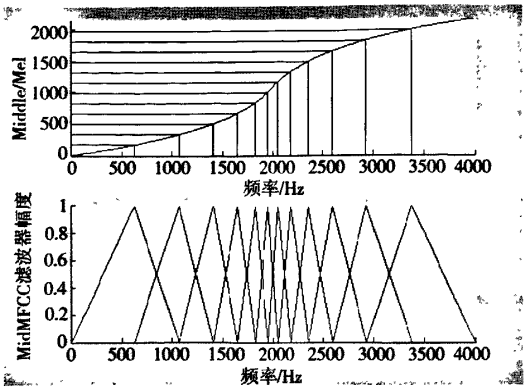

图6　MidMFCC尺度对应和滤波器分布图

Fig. 6　MidMFCC scale curve and distribution of filter

3　实验结果

3.1　语音库

本实验使用的语音样本在室内封闭安静环境下采集得到,提供语音资料的自愿者共14名,其中男性11名,女性3名。语音的采样频率为16 000 Hz,采样精度为16 bit,单声道。样本内容包括18条常用短语,用6种情感状态来表述,分别为高兴、悲伤、生气、恐惧、惊讶和正常,通过剪辑和自愿者主观听觉判断,最后得到1 433句符合要求的情感语音样本。

3.2　实验条件说明

众多MFCC应用文献均选用通带50~4 000 Hz作为计算频域,出于以下3个方面考虑:

① 普通人在正常情况下很难发出50 Hz以下及4 000 Hz以上的声波;

② 50 Hz为工作电流频率,不计算50 Hz的信号可以避免电噪声干扰;

③ 4 000 Hz以上的频谱能量微乎其微,不予以考虑。

实验中选用12阶滤波器计算50~4 000 Hz的通带,帧长512个采样点,帧间重叠128个采样点。因此一条语音样本可以划分成n帧,每一帧计算出12阶MFCC,最后得到12组MFCC特征序列($12\times n$的倒谱系数矩阵)。特征序列可以用序列的统计特征来表述,以达到将二维的系数矩阵降维成一维的特征向量,再用SVM分类器进行识别。

3.3　预处理对比实验

首先,对第2节介绍的预处理方法进行验证,针对是否预加重、使用矩形窗还是汉明窗截取,进行对比实验。选用中位值、方差、最大值3个统计特征来描述每组特征序列,因此一条样本由一个36维的特征向量来表示。SVM作为分类器,使用5交叉验证的方法,计算5次的平均识别率,作为实验的最终结果,识别率如表1所示。

表1　预处理对比实验结果

Tab. 1　Result of pre-emphasis

	无预加重	预加重
矩形窗	87.2%	88.9%
汉明窗	91.1%	92.3%

从表1中,不难发现,通过预加重(加重因子为0.98)和汉明窗截取之后的识别率,要明显优于其他3种组合,这个结果与本文之前的预处理分析相符合。

3.4　Mel倒谱系数验证实验

基于上面的结论,改进算法验证实验使用预加重和汉明窗。此实验的目的是验证改进算法较之经典算法,识别率有所提高,一组使用经典算法的12阶滤波器组,得到12阶MFCC;另外一组使用改进算法的20阶混合滤波器组,分别得到1-6阶MFCC,3-10阶MidMFCC,和7-12阶IMFCC。选用20阶混合滤波器组,一方面保证了低、中、高频段的计算精度,另一方面覆盖整个频域,不遗漏频谱信息。混合滤波器组分布如图7所示。

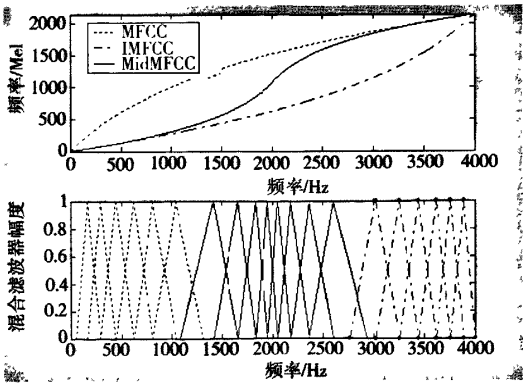

图7　混合型滤波器组分布图

Fig. 7　Distribution of mixed filter group

这样,就得到了需要进行对比验证的2组系数序列:一组为传统的12阶MFCC序列,另一组是混合倒谱系数序列。分别对每个特征序列计算2个、3个、4个统计特征(最大值、平均值、中位值和变化率),依旧采用SVM分类器,5交叉验证的方法,将5次实验的平均识别率作为最终结果,如表2所示。

表2 混合MFCC和纯MFCC实验结果对比

Tab. 2 Results of mixed MFCC and pure MFCC

	混合MFCC/%	传统MFCC/%	提高率/%
2个统计特征	92.9	92.0	+0.9
3个统计特征	93.4	92.3	+1.1
4个统计特征	97.2	95.6	+1.6

从结果可以看出,混合倒谱系数与传统MFCC相比,识别率分别提高了1%左右,特别在使用4个统计特征时,即使MFCC识别率高达95.6%,混合倒谱系数的平均识别率仍然能提高1.6%,达到97.2%。但是混合MFCC也有不足之处,为了满足全频域覆盖和计算精度这2个要求,所使用的混合滤波器个数比MFCC的要多,增加了计算量。

4 结 论

本文参考了MFCC特征参数在工程方面的应用,低阶MFCC计算精度高,具有很好的频率分辨率,是鲁棒的特征识别参数,但中高阶参数的计算精度不高,频谱信息在计算过程中被弱化。针对这个问题我们进行了分析,提出了改进方法,分别提高了中频和高频区域的计算精度,得到了MidMFCC和IMFCC,并将其作为低阶MFCC的补充,应用到语音情感识别中,取得了不错的实验结果。当然,提高识别率是要付出计算代价的,由于要实现全频域覆盖,混合20阶MFCC与传统12阶MFCC方法相比较,三角滤波器个数增加了8个,增加了计算消耗。如果追求识别结果,考虑到实际计算量对目前计算机硬件的要求,这部分计算负担是可以忽略的。下一步工作,我们将继续优化改进的算法,使其在提高计算精度的同时也能够使用尽可能少的滤波器,减少计算时间消耗。

参考文献:

[1] DAVIS S B, MERMELSTEIN P. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Speech Acoustic Processing, 1980, 28: 357-366.

[2] 蒋丹宁,蔡莲红. 基于语音声学特征的情感信息识别[J]. 清华大学学报. 2006, 46(1): 86-89.

[3] PAOT L, CHEN Y T, YEH J H, et al. Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification [EB/OL]. (2005-03-10) [2008-02-10] http://www.actapress.com/PaperInfo.aspx? PaperID=27854reasor=500.

[4] YEN T N, BASS I, Li M K, et al. Investigation of Combining SVM and Decision Tree for Emotion Classification. [EB/OL]. (2005-10-20) [2008-02-10] http://portal.acm.org/citation.cfm? id=1106780.1107199cou=dl=ACM.

[5] CHAKROBORTY S, ROY A, MAJUMDAR S, et al. Capturing Complementary Information via Reversed Filter Bank and Parallel Implementation with MFCC for Improved Text-Independent Speaker Identification [EB/OL]. (2007-04-12) [2008-02-10] http://portal.acm.org/citation.cfm? id=1260199.1260281.

作者简介:

韩一(1982-),男,重庆人,硕士研究生。主要研究方向为语音情感识别。E-mail:tanze_cq001@163.com。

王国胤(1970-),男,重庆人,教授,博士生导师。主要研究方向为人工智能。E-mail:wanggy@cqupt.edu.cn。

杨勇(1976-),男,云南人,副教授,博士研究生。主要研究方向为情感计算,模式识别。Email:yangyong@cqupt.edu.cn

(责任编辑:刘 勇)