DOI: 10.13382/j.jemi.2017.03.018

基于MFCC的语音情感特征提取研究*

李 虹[1] 徐小力[1] 吴国新[1] 丁春艳[2] 赵学梅[3]
(1.北京信息科技大学机电学院 北京 100192;2.中央民族大学 北京 100081;3.丽江东巴文化研究院 丽江 674100)

摘 要:为了研究"世界记忆遗产"东巴经典古籍的音频分类,以通过语音情感特征提取的方法分类鉴别东巴音频类别,并实现对东巴经典语音的情感状态识别,同时提高人机交互性能,提出采用Mel频率倒谱系数(MFCC)实现语音情感特征的提取。通过引入MFCC的一阶差分、二阶差分描述语音特征的动态特征,并整合短时能量特征,最终形成MFCC和短时能量相叠加的语音信号特征参数,达到提取反映语音情感特征的目的。实验验证表明,该语音信号特征提取方法能够更明显地区分出包含在语音中的情感信息,为语音情感特征的识别研究及东巴古籍音频分类鉴别提供理论基础。

关键词:东巴古籍;语音情感特征;特征参数;MFCC;短时能量
中图分类号:TP391.1 文献标识码:A **国家标准学科分类代码**:520.2020

Research on speech emotion feature extraction based on MFCC

Li Hong[1] Xu Xiaoli[1] Wu Guoxin[1] Ding Chunyan[2] Zhao Xuemei[3]
(1. Mechanical & Electrical Institute, Beijing Information Science & Technology University, Beijing 100192, China;
2. Minzu University of China, Beijing 100081, China; 3. Lijiang Dongba Culture Institute, Lijiang 674100, China)

Abstract: To research the audio classification of "memory of the world heritage" Dongba classic books, the way of speech emotion feature extraction is adopted to identify the Dongba audio categories and realize the emotional state recognition of Dongba audio. And to improve the performance of human computer interaction, the way of speech emotion feature extraction based on Mel frequency cepstrum coefficient (MFCC) is adopted. The first order difference and the second order difference are introduced to describe the dynamic characteristics of speech features. The speech signal characteristic parameters based on MFCC and the short-time energy are finally formed to extract the characteristics of speech emotion feature. The experiment research shows that the characteristics of the speech signal can be more clearly differentiate emotional information contained in the speech, and lay the foundation for recognizing speech emotion and the Dongba audio classification.

Keywords: Dongba classic books; speech emotion feature; characteristic parameters; MFCC; short-time energy

1 引 言

一直以来,人类的情感研究都是语言学、生物学、神经学等领域的重要研究方向,其在人工智能、人机交互等发展中担当着极其重要的角色[1]。随着电子技术的不断发展,人机交互越来越受到研究者的重视。如今,用户已经无法满足于仅仅通过键盘等冰冷机械的交互方式来完成与计算机间的对话,自然、和谐的人机沟通向着理解用户情绪及意图、对不同环境给与不同反馈的方向大步发展。而作为人类交流主要途径的语音信号因其易于捕获的特性,使得语音情感识别成为人类情感研究中最重要的方式之一。

对于东巴古籍诵读音频而言,因其音调、韵律特征明显,情感表达丰富且场景易于分辨,故而采用语音特征提取的方法对音频中包含的语音情感进行识别,从

收稿日期:2017-01 Received Date: 2017-01
*基金项目:国家社科基金重大项目(12&ZD234)、现代测控技术教育部重点实验室开放课题(KF20161123205)、北京市重点实验室开放课题(KF20161123208)资助项目

而达到区分东巴古籍音频类别的目的。由于特征提取的结果很大程度上影响这语音情感判定的准确性。所以,语音情感特征提取方法的研究具有十分重要的研究意义。

2　语音情感特征分析

语音信号是人类情感交流的必要手段,其特征主要是指它的声学特征、语音信号时频特性及语音信号的统计特性等[2]。目前来说,世界上已有的语音特征提取方法较多,研究者对于特征提取方法对语音情感识别的有效性还未研究定论。大体上能够将语音特征简单分为 3 个类别:韵律特征、谱特征、其他特征。

1)韵律特征

韵律特征即为声音的物理属性,其主要表征现象就是人的发声器官所产生出的声波信号。在语音情感识别中,其应用广泛的原因是韵律特征能够传递语句中较多的语音情感信息。经典的韵律特征主要包括如下几点[3]。

(1)基音频率特征。该特征属于语音的韵律特征基本属性之一,主要包含基音频率特征信息的包络以及基音频率的线性预测系数等。

(2)共振峰特征。该特征属于韵律特征的细节描述,主要包括一阶共振峰、二阶共振峰、共振峰带宽等特征信息。

(3)能量特征。该特征属于韵律特征的强度标识,往往包含有 4 阶 Legendre 参数、shimmer 参数等特征信息。

(4)时间特征。包含说话部分和不说话部分比值,语速等。

2)谱特征

较之于韵律特征的连续性,谱特征通常由语音信号的短时表示。由于语音信号的产生过程是多个发音器官共同作用的结果,发音器官的物理特性使发音器官难以在短时发生较大变化,在 5 ~ 50 ms 语音信号可以被认作是平稳的,即其谱特性变化不大。

比较经典的谱特征有短时傅里叶变换,线性预测系数(linear predictor coefficients,LPC),Mel 频率倒谱系数(Mel frequency cepstrum coefficients,MFCC),感知线性预测倒谱系数(perceptual linear predictive cepstral coefficients,PLP),线谱对参数(line spectrum pair,LSP),短时连贯性(short time coherence,SMC)。

3)其他特征

除此之外,在语音情感识别中也往往使用其它的语音特征提取方法,比如 Teager 能量算子(teager energy operator,TEO)、经验模态分解(empirical mode decomposition, EMD)、分形维(fractal dimension)、深度学习等。

3　语音情感特征提取方法

3.1　情感特征及语音参数关系

针对语音的语种和研究应用环境的差异,语音情感的研究方法和重点也有所不同。对于情感的分类,大部分学者认为的主要情感包括生气(anger)、高兴(happiness)、悲伤(sadness)和厌恶(disgust)。不同情感在实际情况中对应的是不同的语音声道特征和激励源的统计特征,Murray 和 Amott 总结了情感和语音参数之间的关系如表 1 所示[4]。

表 1　情感和语音参数之间的关系(Murray & Amott 1993)

Table 1　The relationship between emotion and speech parameters(Murray & Amott 1993)

规律	生气	高兴	悲伤	恐惧	厌恶
语速	略快	快或慢	略慢	很快	非常快
平均基音	非常高	很高	略低	非常高	非常低
基音范围	很宽	很宽	略窄	很宽	略宽
强度	高	高	低	正常	低
声音质量	有呼吸声胸腔声	有呼吸声共鸣音调	有共鸣声	不规则声音	嘟囔声胸腔声
基音变化	重音处突变	光滑向上弯曲	向下弯曲	正常	宽最终向下弯曲
清晰度	清晰	正常	含糊	精确	正常

根据宗教功能及文化属性,其东巴祭仪可分为祭祀神灵类、驱除鬼怪类及丧葬超度类。其中,各类祭祀吟唱方式、声调及包含的情感特征均有不同,故而提出采用语音特征提取的方法对音频中包含的语音情感进行识别,能够大致区分东巴古籍音频的类别。

3.2　语音情感特征分析

3.2.1　语速和能量特征

1)语速

语音情感与语速有关,平均发话时长由发音的音节数与持续时间的比值确定。分析计算对比同一个人不同

情绪下的读每个音节所占时长,得到结果如表2所示。MATLAB建模分析如图1所示。

表2　语速分析

Table 2　The speed analysis

语速	生气	高兴	悲伤	平静
录音员1	0.312 9	0.240 5	0.348 3	0.225 0
录音员2	0.365 5	0.518 6	0.577 6	0.293 8
录音员3	0.252 3	0.237 4	0.277 1	0.254 8
录音员4	0.208 6	0.251 0	0.363 7	0.316 7

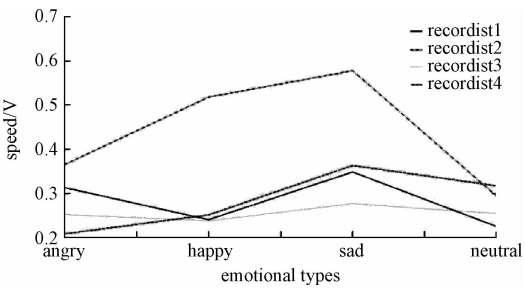


图1　不同录音员平均语速对比分析

Fig. 1　The analysis of the average speed between different recordists

研究分析可知,高兴和生气时,其语速较快;悲伤时,语速较慢;平静时的语速因人而异,差别较大。因此用语速单纯判定语音情感特征的效果不理想,必须利用语速与语音能量等语音特征相结合的方法来判定语音情感特征的区别。

(2)短时能量

短时能量定义:设语音波形时域信号为 $x(n)$,加窗函数 $\omega(n)$ 分帧处理后得到的第 i 帧语音信号为 $y_i(n)$,则 $y_i(n)$ 满足[5]:

$$y_i(n)=\omega(n)\cdot x((i-1)\cdot inc+n),1\leqslant n\leqslant L,1\leqslant i\leqslant fn,$$

式中:L 为帧长,inc 为帧移长度,f_n 为分帧后总帧数。

计算第 i 帧语音信号 $y_i(n)$ 的短时能量公式为:

$$E(i)=\sum_{n=0}^{L-1}y_i^2(n),1\leqslant i\leqslant f_n$$

分析对比同一个人在不同语音段下以不同情绪朗读语音的短时能量,得到结果如表3所示。MATLAB建模如图2所示。

表3　短时能量分析

Table 3　The analysis of short-time energy

短时能量	生气	高兴	悲伤	平静
语音1	0.562 2	0.178 4	0.180 7	0.171 7
语音2	0.867 8	0.326 6	0.170 7	0.177 6
语音3	0.541 9	0.213 6	0.154 9	0.176 9
语音4	1.434 0	0.428 0	0.251 5	0.287 1

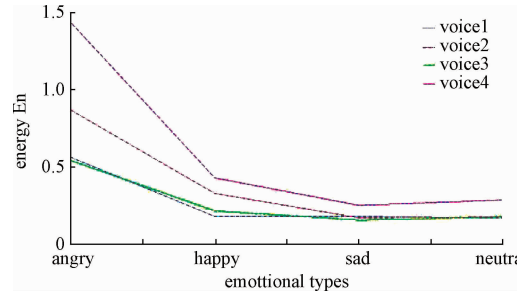


图2　短时能量对比分析

Fig. 2　The analysis of short-tine energy

研究分析可知,生气和高兴时的语音信号能量较高,平静时语音信号的能量较为居中,悲伤时的语音信号能量最低。短时能量能够在一定程度上反映基本情绪变化。它在语音端点检测方面有着较理想的应用效果。若希望进一步区别语音情感特征的差异性,尤其是在基于人耳听觉特性的细微差异方面进行情感特征提取与分析,由 Davis 和 Mermelstein 提出的梅尔频率倒谱系数 MFCC 往往具有优越性。

3.2.2　MFCC 频率倒谱系数

MFCC 基于人的听觉机理分析语音的频率以获得好的语音特性,1 Mel 为 1 000 Hz 的音调感知程度的 1/1 000。具体定义为[6]:

$$f_{\text{mel}}=2595\log_{10}(1+\frac{f_{\text{Hz}}}{700})$$

或

$$f_{\text{mel}}=1125\ln(1+\frac{f_{\text{Hz}}}{700})$$

式中:f_{Hz} 为实际线性频率,f_{mel} 为 Mel 频标。

研究 Mel 频率与线性频率对应关系及 Mel 滤波器组的频率相应曲线可知,Mel 频率和线性频率存在非线性对应关系,滤波器在低频区域分布较为密集,如图3所示。

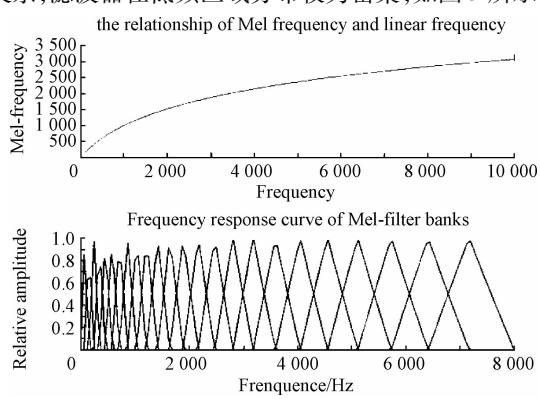


图3　Mel 频率与线性频率对应关系及 Mel 滤波器组频率响应曲线

Fig. 3　Correspondence between Mel frequency and linear frequency and frequency response curve of Mel filter bank

Mel 滤波器倒谱参数特征在语音特征提取中占有重要的地位,其计算简单、区分能力较为突出[7]。MFCC 的特征参数提取原理如图 4 所示。

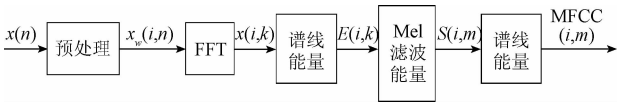


图 4　MFCC 特征参数提取原理

Fig. 4　Schematic diagram of MFCC feature extraction

具体过程如下。

1)原语音信号经过预加重、分帧、加窗处理后得到单帧的短时信号 $x(n)$[8]。预加重是为了补偿高频分量的损失,提升高频分量;分帧的目的是能够将较短的单帧作为稳态信号处理,使帧间参数平稳过度[9];加窗的目的是减少频域的泄漏[10]。

2)快速傅里叶变换(FFT)。对单帧信号进行变换,得到频域数据[11]:

$$X(i,k) = FFT[x_i(m)]$$

3)计算每帧谱线能量[12]:$E(i,k) = [X(i,k)]^2$

4)将每帧谱线能量谱通过 Mel 滤波器,计算在该 Mel 滤波器中的能量[13]。频域中,相当于把单帧能量谱 $E(i,k)$ 与 Mel 滤波器的频域响应 $H_m(k)$ 相乘并相加[14]:

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k)H_m(k) \quad 0 \leqslant m < M$$

式中:i 为第 i 帧,k 为频域中第 k 条谱线。

5)将 Mel 滤波器的能量取对数后计算 DCT[15]:

$$MFCC(i,n) = \sum_{m=0}^{M-1} \log[S(i,m)\cos(\frac{\pi n(2m-1)}{2M})]$$

式中:m 指第 m 个 Mel 滤波器(共有 M 个),i 指第 i 帧,n 是 DCT 后的谱线。

4　分析方法

4.1　研究方法

研究分析语音信号 MFCC 特征分量的相对重要性,由于 MFCC 表征语音信号的静态特性,从而引入 MFCC 的一阶差分、MFCC 的二阶差分表述其动态特性,构成新的 MFCC 特征分量;引入短时能量特征,利用矩阵变换将其与 MFCC 倒谱向量进行行列匹配,得到新的 MFCC 与短时能量相叠加的特征参数表征语音信号。

4.2　研究结果分析

由专业人员录制的同一语句的不同情感表述的音频以分析情感特征参数变化。选用其中 4 种表现力较强的情感,即生气、高兴、难过、平静,其语音信号波形如图 5 所示。

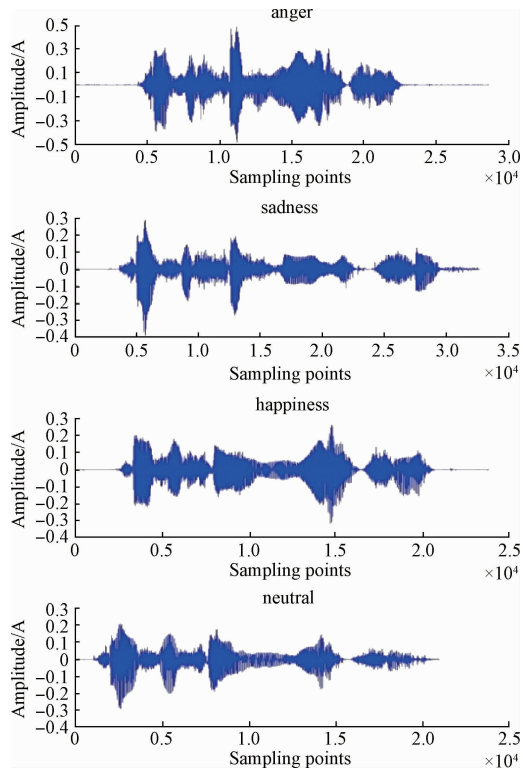
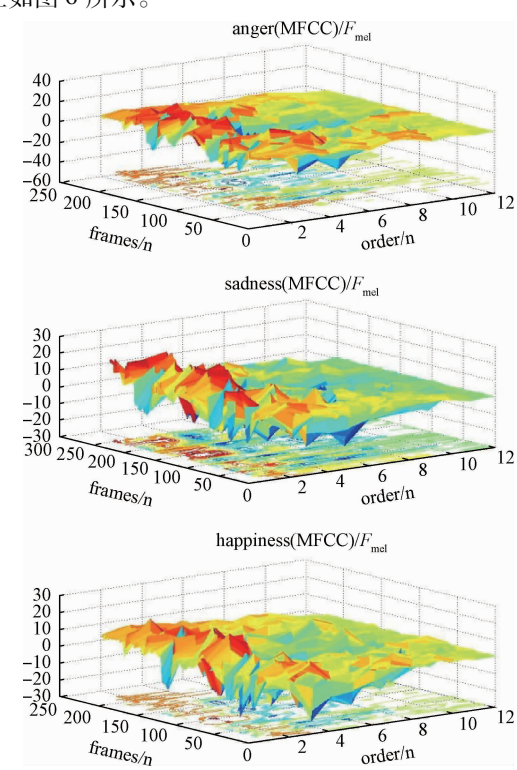


图 5　语音信号

Fig. 5　The waveforms of the speech signals

实验研究中,选取预加重系数为 0.95,帧长 $wlen = 256$,帧移 $inc = 128$,汉明窗,Mel 滤波器组取 24 个。针对以上 4 中语音进行了 MFCC 的语音情感特征提取。信号特征如图 6 所示。



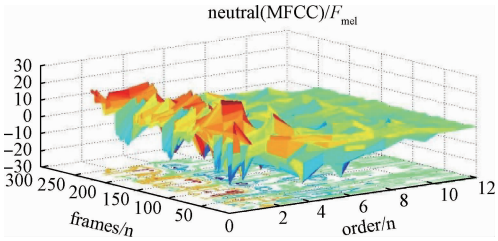


图6　语音情感特征提取信号图(MFCC)

Fig. 6　The signal features of speech emotion feature extraction (MFCC)

增加帧能量特性,利用矩阵变换将其与MFCC矩阵进行行列匹配,得到新的MFCC与短时能量相叠加的特征参数,信号特征如图7所示。

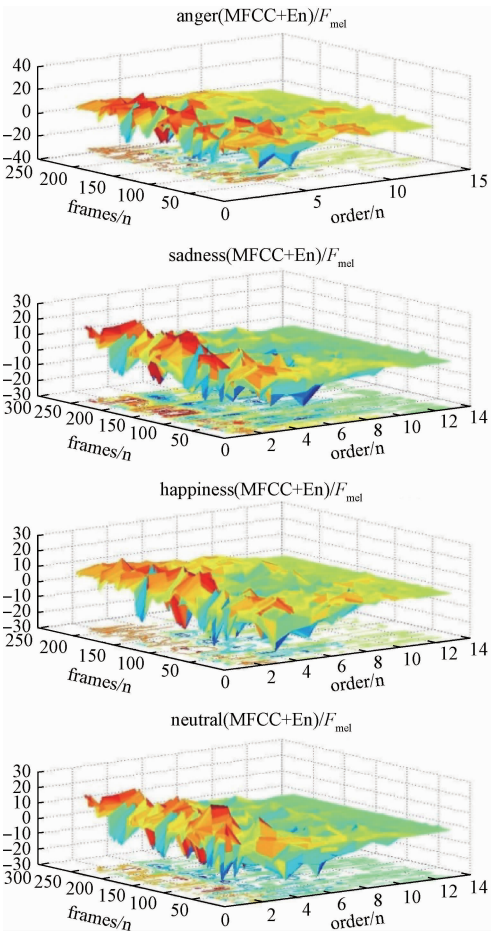


图7　语音情感特征提取信号图(MFCC+En)

Fig. 7　The signal features of speech emotion feature extraction (MFCC+En)

由图7分析可知,不同的情感的语音特征在阶数/帧数的维度上分布往往存在较大差异,若根据不同情感分析相应的大量语句,即可以得到不同情感的特征参数,从而建立较为精准分类模型,无论是基于神经网络识别还是其他方法的识别,其都能够提供较好的语音信号特征基础。另外,MFCC特征主要体现在低阶部分,故而舍弃高阶部分,并加入其他语音特征,以更好的表征语音信号的情感特征。

5　结　论

实验研究分析,不同情感状态的MFCC参数具有较大差异,MFCC特征参数能够较好的表述语音信号特征。在此基础上,加入短时能量特征,并引入MFCC的一阶差分、二阶差分以表征其动态特性,能够更好的反映不同情感状态的语音信号特征。此语音信号特征提取方法能够更明显地区分出包含在语音中的情感信息,为语音情感特征的识别研究及东巴古籍音频分类鉴别提供理论基础。

为更好的得到语音信号特征参数以建立语音情感特征类库,可进一步进行较为广泛的基音频率特性,共振峰特性等语音特征的研究工作,为语音情感识别及合成提供理论研究基础。

参考文献

[1] 孙颖. 语音情感识别与合成的研究[D]. 太原: 太原理工大学, 2011
SUN Y. Research on speech emotion recognition andsynthesis [D]. Taiyuan: Taiyuan University of Technology, 2011.

[2] 郭鹏娟. 语音情感特征提取方法和情感识别研究[D]. 西安: 西北工业大学. 2007.
GUO P J. Speech emotion feature extraction andemotion recognition [D]. Xi'an: Northwestern Polytechnical University,2007.

[3] 孙亚新. 语音情感识别中的特征提取与识别算法研究[D]. 广州: 华南理工大学. 2015.
SUN Y X. Research on feature extraction and recognition algorithm for speech emotion recognition [D]. South China University of Technology, 2015.

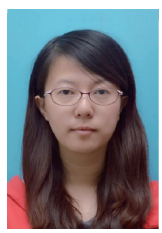
[4] KAWANAMI H, HIROSE K. Consideration on the prosodic features of utterances with attitudes and emotions[R]. Technical Report of IEICE, 1997(11): SP97-67.

[5] 宋知用. MATLAB在语音信号分析与合成中的应用[M]. 北京: 北京航空航天大学出版社,2013.
SONG ZH Y. The Application of MATLAB in the Analysis and Synthesis of Speech Signal[M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2013.

[6] 章熙春, 曹燕, 张军, 等. 语音MFCC特征计算的改进方法[J]. 数据采集与处理, 2005, 20(2):

161-165.
ZHANG X CH, CAO Y, ZHANG J, et al. An improved method for computing the MFCC features of speech[J]. Data Acquisition and Processing, 2005, 20(2): 161-165.
[7] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.
HAN W J, LI H F, RUAN H B, et al. Review on speech emotion recognition [J]. Journal of Software, 2014, 25(1): 37-50.
[8] 陈立江, 毛峡, ISHIZUKA M. 基于 Fisher 准则与 SVM 的分层语音情感识别[J]. 模式识别与人工智能, 2012, 25(4): 604-609.
CHEN L J, MAO X, ISHZUKA M. Multi-level speech emotion recognition based on fisher criterion and SVM[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(4): 604-609.
[9] 林奕琳, 韦岗, 杨康才. 语音情感识别的研究进展[J]. 电路与系统学报, 2007, 12(1): 90-98.
LIN Y L, WEI G, YANG K C. A survey of emotion recognition in speech [J]. Journal of Circuits and Systems, Journal of Circuits and Systems, 2007, 12(1): 90-98.
[10] 黄晨晨, 巩微, 伏文龙, 等. 基于深度信念网络的语音情感识别的研究[J]. 计算机研究与发展, 2014, 51(Sl): 75-80.
HUANG CH CH, GONG W, FU W L, et al. Research of speech emotion recognition based on DBNs[J]. Journal of Computer Research and Development, 2014, 51(Sl): 75-80.
[11] 黄程韦,赵艳, 金赟, 等. 实用语音的情感特征分析与识别[J]. 电子与信息学报, 2011, 33(1): 112-116.
HUANG CH W, ZHAO Y, JIN Y, et al. A study on feature analysis and recognition of practical speech emotion [J]. Journal of Electronics & Information Technology, 2011, 33(1): 112-116.
[12] 赵力, 黄程韦. 实用语音情感识别中的若干关键技术[J]. 数据采集与处理, 2014, 29(2): 157-170.
ZHAO L, HUANG CH W. Key technologies in practical speech emotion recognition [J]. Journal of Data Acquisition and Processing, 2014, 29(2): 157-170.
[13] 罗宪华, 徐海明. 基于特定人的语音情感识别系统构建[J]. 中国人民公安大学学报: 自然科学版, 2015(4): 72-75.
LUO X H, XU H M. Construction of speech emotion recognition system based on specific person[J]. Journal of People's Public Security University of China: Science and Technology, 2015 (4): 72-75.
[14] LAN S K, SHI Y B. An improved algorithm for mfcc parameters in speaker recognition system[J]. Journal of Luoyang Institute of Science and Technology: Natural Science Edition, 2013, 23(4): 23-24.
[15] ZHANG J, FAN M, FENG W Q, et al. Improvement of speaker feature extraction algorithm based on MFCC parameters[J]. Audio Engineering, 2009, 33(9): 61-63.

作者简介



李虹,1989 年出生,现为北京信息科技大学在读研究生,主要研究方向为机械电子工程。
E-mail:331057469@ qq. com

Li Hong was born in 1989, M. Sc. candidate in Beijing Information Science and Technology University. The main research direction is mechatronic engineering.