

基于 MFCC 和 CHMM 技术的语音情感分析 及其在教育中的应用研究

张永皋¹, 马青玉², 孙青¹

(1. 南京师范大学 教育科学学院, 江苏 南京 210097
2. 南京师范大学 物理科学与技术学院, 江苏 南京 210097)

[摘要] 语音情感识别作为一个新的研究热点, 因其能解决教育中情感缺失的问题, 而越来越受到研究者的重视. 选取符合人类听觉系统感知的梅尔频率倒谱系数(MFCC)与各态历经型的连续隐马尔可夫模型(CHMM)进行语音情感特征的分析, 并对大量的语音信号进行情感识别实验, 识别正确率达到 86.7%, 为教育中的情感补偿提供了切实可行的依据.

[关键词] MFCC, CHMM, 情感识别

[中图分类号] TN 912.3 [文献标识码] A [文章编号] 1672-1292(2009)02-0089-04

Investigation on Speech Emotion Analyses and Its Application in Education Based on MFCC and CHMM Techniques

Zhang Yonggao¹, Ma Qingyu², Sun Qing¹

(1. School of Educational Science, Nanjing Normal University, Nanjing 210097, China
2. School of Physics and Technology, Nanjing Normal University, Nanjing 210097, China)

Abstract As a new research hotspot, speech emotion recognition attracts more and attentions from researchers for its solution of the emotion loss in distance education. Based on the perception of the human auditory system, the Mel Frequency Cepstral Coefficients (MFCC) and the Continuous Hidden Markov Model (CHMM) techniques are used in this paper for speech characteristic analyses and emotion identification. The recognition accuracy of 87.6% is achieved in the experiment for large numbers of speech signals and the potential application for emotion compensation is provided in education.

Key words MFCC, CHMM, speech emotion recognition

情感计算^[1]是指与情感有关的、由情感引发或者能够影响情感的因素的计算. 相应的智能系统能够通过各种传感器获取由人的情感引起的生理及行为特征信号, 创建相应的情感模型, 实现对人类情感的理解, 并对此做出合适的反馈. 目前情感计算采用图像、声音、姿态和动作等多种信息进行综合的分析, 已取得了良好的效果. 由于信息量和计算量巨大, 情感计算的速度较慢, 不能进行实时处理.

语言是人类最重要的信息交流工具, 它自然方便、准确高效, 不仅是社会交际的工具, 也是人类思维的工具. 在思维过程中, 人们所用的语言是语言的内部形态, 而语音则是这种内部形态的外部实现. 对语音进行情感计算, 来实现对情感的感知, 具有重要的意义.

在教育过程中, 语音同样具有不可替代的作用. 教师主要通过语音传递教学内容, 实现知识的传输. 教师的语速、语调、情感、姿态等这种“非言语行为”必不可少^[2]. 人们在接收语音所包含信息的同时, 也接收了包含在语音中的语调及情感信息这类很重要的“非言语行为”.

随着计算机及网络技术的发展, 借助多媒体技术、网络技术来实现教育最优化的网络教育在资源共享、强交互性等方面有很大的优势, 但网络教育存在着明显的忽视情感的教育倾向, 最突出的表现就是网络教学中师生交互的情感缺失. 在人机交互的学习环境中, 可以开发出语音情感识别系统, 对学生的情感信号进行

收稿日期: 2008-12-16

基金项目: 教育部留学回国人员科研启动基金(2008102SBJ0154)、南京师范大学留学回国人员基金(2008102XLH0070)和南京师范大学教务处精品实验资助项目.

通讯联系人: 马青玉, 博士, 副教授, 研究方向: 声信号处理与生物医学电子技术应用. E-mail: maqingyu@njnu.edu.cn

捕捉和识别,并判定学生的学习状态和对所学知识的接受情况,来有效地解决情感交流匮乏的问题。

本文针对远程教育对情感识别需要实时性的要求,通过对语音情感的特征进行分析,实现了对情感的识别,达到了良好的效果。语音情感分析的结果可以被应用到远程教育教学中,在一定程度上弥补情感缺失的缺点。

1 语音情感分析及识别方法选择

不同人的语音具有不同的特征,通过语音就能够感知说话者不同的情感,因此通过语音特征分析可以有效地进行语音的情感分析。如何让计算机能够从语音中自动识别出讲话者的情感状态,近年来越来越多的研究者进行了语音情感方面的研究,实现了对语音信息的情感识别^[3-4]。

语音的特征可以从两个方面考虑:从语音信号的产生过程来看,对发音器官的声学模型分析,可以得到语音的韵律特征;从人类的听觉感知方面,可以构建听觉模型来获取情感信息。语音的韵律特征是说话者情感状态的一个重要指示。研究表明,时间构造、振幅构造、基频构造和共振峰构造等这些韵律特征都能反映出情感的变化^[5-6],可以通过获取语音信号的基频、共振峰等参数的平均值、差分、最大值、最小值等作为情感识别的特征。另一方面,人的听觉系统是一个特殊的非线性系统,它响应不同频率信号的灵敏度是不同的。在 1000 Hz 以下,人类听觉的感知能力与频率成线性关系;而在 1000 Hz 以上,感知能力则与频率成对数关系。为了模拟人耳对不同频率感知特性,人们提出了 Me 频率倒谱系数(MFCC),其与线性频率的转换关系是 $f_{me1} = 2595 \log_{10}(1 + f/700)$,其中 f 为频率。在语音信号处理中 MFCC 参数以帧为单位来计算,首先要通过快速傅里叶变换(FFT)得到该帧信号的功率谱 $S(n)$,然后将 $S(n)$ 通过 Me 频率滤波器组得到 Me 频率下的功率谱,并通过对数能量的处理,得到对数功率谱 $S(m)$ 。Me 滤波器组是在语音的频谱范围(大约为 20 ~ 20 kHz)内设置若干个带通滤波器 $H_m(k)$ ($0 \leq m < M$), M 为滤波器的个数,最后将对数功率谱 $S(m)$ 经过离散余弦变换,得到式(1)所表示的 Me 频率倒谱系数 $c(n)$:

$$c(n) = \sum_{m=1}^{M-1} S(m) \cos\left\{\frac{\pi n(m+1/2)}{M}\right\} \quad (0 \leq m < M). \quad (1)$$

在实际应用中,标准的 MFCC 参数只反映了语音参数的静态特性,而人耳对语音的动态特性更为敏感,通常用 MFCC 的差分倒谱参数来描述这种动态特性:

$$d(n) = \sum_{i=-k}^k i \cdot c(n+i) / \sqrt{\sum_{i=-k}^k i^2}. \quad (2)$$

式中, c 表示每一帧语音参数; k 为常数,表示计算某一帧的前后 k 帧参数的差分,此处 k 取值为 2。通过式(2)可以得到一阶 MFCC 差分参数,用同样公式对一阶差分参数进行计算可得二阶 MFCC 差分参数。

本文选择与语言的双重随机过程相吻合的连续隐含马尔科夫模型(CHMM)来进行情感识别。CHMM 一方面用隐含的状态对应于声学层各相对稳定的发音单位,并通过状态转移和状态驻留来描述发音的变化;另一方面它引入了概率统计模型,使用概率密度函数计算语音参数对 CHMM 模型的输出概率,通过搜索最佳状态序列,以最大后验概率为准则找到识别结果。人的语言过程可以看作是一个双重随机过程,语音信号本身是一个可观测的时变序列,是由大脑根据语法知识和语言需要(不可观测的状态)发出的音素的参数流,能描述语音信号的整体非平稳性和局部平稳性。本文结合高斯混合模型(GMM),用 CHMM 来对语音情感进行识别,避免了对特征参数进行矢量量化引入的量化误差。在各态历经型结构 CHMM 的参数重估 Baum-Welch 算法^[7]中,状态转移概率 a_{ij} 的重估过程为其它所有状态转移到该状态的概率之和,可以表示为:

$$a_{ij} = \frac{\text{状态 } S_i \text{ 向状态 } S_j \text{ 转移的概率}}{\text{状态 } S_i \text{ 的概率}} = \frac{\sum_{t=1}^T \xi(i,j)}{\sum_{t=1}^T \gamma_t(i)}. \quad (3)$$

式中, $\xi(i,j)$ 为利用前向概率和后向概率计算得到的过渡概率; $\gamma_t(i)$ 在 t 时刻为状态 i 的概率。Bjorn Schuller^[8]等人通过实验证明,随着采用状态数的增加,通过训练得到模型的识别率逐渐提高,但同时运算也增大。在远程教学过程中实现语音情感识别需要达到实时识别的要求。本文选择 12 维特征数的 MFCC

以及一阶差分作为训练和识别的情感特征。

2 实验研究

2.1 情感语音库

本文采用柏林语言情感库 (Berlin emotional speech database)^[9] 作为实验用语音库, 该情感语音库共包含 5 男 5 女录制的 494 个语句, 共区分为 6 种情感: 生气 (anger)、厌烦 (boredom)、恐惧 (fear)、高兴 (happiness)、悲伤 (sadness) 和中性 (neutral)。语音为采样率 16 kHz 的单声道波形文件。在 6 种情感中, 本文选择生气、厌烦、悲伤和高兴 4 种情感共 260 句, 其中每种随机抽取 30 个语句作来识别, 其它作为训练语句。

2.2 实验过程

在实验研究中, 首先对语言进行特征分析, 获得 MFCC 的不同频段的分布情况。图 1 显示了同一人不同情感表达的同一句话获得的 MFCC 分布图。悲伤语句 MFCC 在 12 阶的频率中分布较平稳, 其它情感的分布较离散, 其中生气在低音处语音包括最多, 说明人在生气时声音低沉, 相反在高兴时声音欢快。

将每种情感语句提取 MFCC 特征后, 在 CHMM 模型中进行训练。CHMM 计算流程如图 2 所示。把提取特征参数按情感依次输入每个 CHMM 模型, 通过 Baum-Welch 算法进行训练, 通过设定最大迭代次数和输出概率的最小相对变化来结束每个模型的训练, 最后得到每种情感的 CHMM 模型。

在获得待识别语音的特征向量后, 将该特征向量对先前计算得到的每类 CHMM 模型计算相应的概率, 最后进行判断决策, 概率最大的即识别为该语音的情感状态。

2.3 实验结果

(1) 提取 MFCC 作为特征参数, 进行语音识别, 结果如表 1 所示。

可以看到, 使用 MFCC 作为特征参数, 只使用从左到右型的 CHMM 能够达到 82.5% 的识别率, 说明只使用模拟听觉系统的特征来进行情感识别能够达到较好的效果。

(2) 以 MFCC 和一阶差分作为特征参数, 实验结果如表 2 所示。

实验结果表明, 使用 MFCC 及表示各帧之间关系的一阶差分作为特征向量能够达到 83.3% 的识别正确率, 而各态历经型 CHMM 的识别结果达到了 86.7%, 识别正确率有了显著提高。

表 1 使用 MFCC 作为特征参数的识别结果

Table 1 Identification results with the feature parameter of MFCC

从左到右型	生气	厌烦	悲伤	高兴	统计
生气	83	0	0	17	83
厌烦	0	80	10	10	80
悲伤	0	3	97	0	97
高兴	20	0	0	80	70

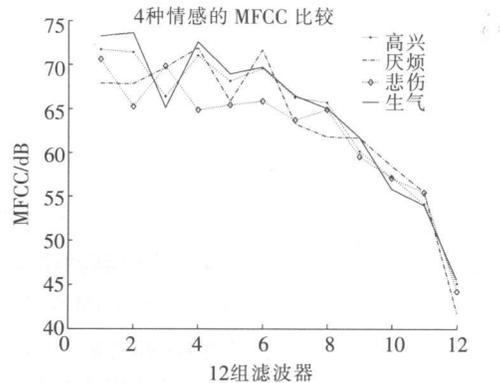


图 1 同一人用不同情感表达同一句话的 MFCC

Fig.1 MFCC result extracted from the same words in different emotion

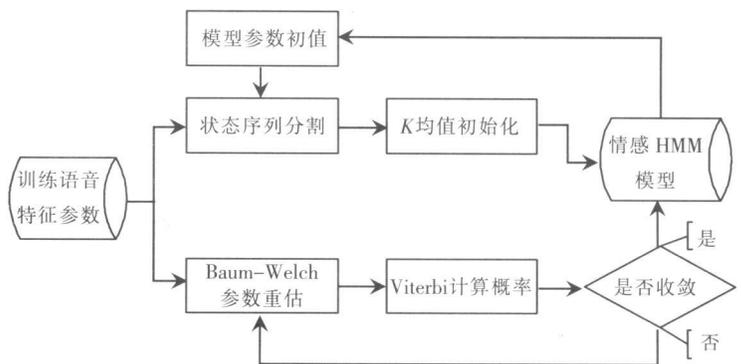


图 2 CHMM 的计算流程图

Fig.2 Calculation flow chart of CHMM

表 2 使用 MFCC 及一阶差分作为特征参数的识别结果

Table 2 Identification results with the feature parameters of MFCC and its first-order difference

从左到右型 / 各态历经型	生气	厌烦	悲伤	高兴	统计
生气	83 / 100	0 / 0	0 / 0	17 / 0	83 / 100
厌烦	0 / 0	83 / 90	13 / 3	4 / 7	83 / 90
悲伤	0 / 0	7 / 7	93 / 93	0 / 0	93 / 93
高兴	30 / 36	0 / 0	0 / 0	70 / 64	70 / 64

(3) 时间测量.

在 MFCC 计算过程中, 在每帧为 256 时计算 FFT 需要 1 024 次乘法计算, 最后的复杂度大约为 $\sum_{k=1}^M 1 024^k$ 其中 M 为滤波器个数. 在 Viterb 算法的递推过程中, 如果 T 为语音信号的帧的数量, N 为 CHMM 的状态数, 需要的时间复杂度为 $\sum_{t=2}^T \sum_{j=1}^N \sum_{i=1}^N \left(1 + \sum_{i=1}^N (1 + 2) \right)$. 假定该语音的分帧为 300 帧, CHMM 的状态数为 16 需要计算次数的量级为 10^8 , 而一台普通的 P4 3.0 的机器每秒种计算次数达到 10^9 , 每计算一次情感所需要的时间大约为 0.1 s 在语音识别实验中, 由于使用 MATLAB 编程和语音的长短问题, 利用本算法进行一次语音情感识别时间大约为 0.4~0.6 s 证明本算法基本能符合实时情感分析的要求.

3 结 语

针对教育中的情感缺失问题, 提出了基于 MFCC 和 CHMM 技术的语音情感分析技术. 在以往的研究中, Albino Nogueiras^[10] 使用 HMM 对语音情感分析达到了高于 80% 的识别率, Björn Schuller^[6] 等人也达到了 81.3% 的识别正确率, 但是他们都是将基频、能量等多维特征考虑进去, 这样需要很多的计算量. 本文选择了符合人类听觉的 MFCC 及其一阶差分作为情感特征, 对语音情感的特征进行分析. 在 CHMM 的状态转移方面, 使用各态历经型实验得到的结果除高兴有些降低外, 总识别率达到 86.7%, 其中生气达到了 100% 的识别, 说明了对 MFCC 及 CHMM 的分析的有效性. 本文只选择一种语音特征, 忽略了基频等其它韵律特征, 缩减了特征提取所需的时间, 保证较快的识别速度. 在以后的研究当中, 可以开发语音情感识别模块, 嵌入到教学系统中, 在教育过程中发挥情感分析和情感补偿的作用.

[参考文献] (References)

- [1] 张迎辉, 林学閤. 情感可以计算——情感计算综述[J]. 计算机科学, 2008, 35(5): 5-8
Zhang Yinghui, Lin Xueyan. Affect is computable——a survey on affective computing[J]. Computer Science, 2008, 35(5): 5-8 (in Chinese)
- [2] 隋云翔, 刘平. 非言语沟通与课堂教学[J]. 现代中小学教育, 1991(4): 50-52
Sui Yunxiang, Liu Ping. Non-verbal communication in classroom teaching[J]. Modern Primary and Secondary Education 1991(4): 50-52 (in Chinese)
- [3] 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: 自然科学版, 2006, 46(1): 86-89
Jiang Danning, Cai Lianhong. Speech emotion recognition using acoustic features[J]. Journal of Tsinghua University Science and Technology Edition, 2006, 46(1): 86-89 (in Chinese)
- [4] 詹永照, 曹鹏. 语音情感特征提取和识别的研究与实现[J]. 江苏大学学报: 自然科学版, 2005, 26(1): 72-75
Zhan Yongzhao, Cao Peng. Research and implementation of emotional feature extraction and recognition in speech signal[J]. Journal of Jiangsu University Natural Science Edition, 2005, 26(1): 72-75. (in Chinese)
- [5] 赵力, 将春辉, 邹采荣, 等. 语音信号中的情感特征分析和识别的研究[J]. 电子学报, 2004, 32(4): 606-609
Zhao Li, Jiang Chunhui, Zou Cairong, et al. A study on emotional feature analysis and recognition in speech[J]. Acta Electronica Sinica, 2004, 32(4): 606-609. (in Chinese)
- [6] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition [C] // IEEE International Conference on Acoustics, Speech, & Signal Processing Hongkong, China, 2003: 401-404.
- [7] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2003
Zhao Li. Voice Signal Processing[M]. Beijing: China Machine Press, 2003. (in Chinese)
- [8] Ayadi M, Kamel M, Karray F. Speech emotion recognition using Gaussian mixture vector autoregressive models [C] // IEEE International Conference on Acoustics, Speech, & Signal Processing Honolulu, Hawaii, USA, 2007: IV-957-IV-960.
- [9] de Cheveign A, Kawahara H. YIN, a fundamental frequency estimator for speech and music [J]. Journal Acoustical Society of America, 2002, 111(4): 1917-1930.
- [10] Nogueiras A, Moreno A, Bonafonte A, et al. Speech emotion recognition using hidden Markov models [C] // Proceedings of Eurospeech, Aalborg, Denmark, 2001: 2 679-2 682.

[责任编辑: 严海琳]