基于情感特征分类的语音情感识别研究*

周晓凤，肖南峰，文 翰
（华南理工大学 计算机科学与工程学院，广州 510006）

摘 要：针对语音信号的实时性和不确定性，提出证据信任度信息熵和动态先验权重的方法，对传统 D-S 证据理论的基本概率分配函数进行改进；针对情感特征在语音情感识别中对不同的情感状态具有不同的识别效果，提出对语音情感特征进行分类。利用各类情感特征的识别结果，应用改进的 D-S 证据理论进行决策级数据融合，实现基于多类情感特征的语音情感识别，以达到细粒度的语音情感识别。最后通过算例验证了改进算法的迅速收敛和抗干扰性，对比实验结果证明了分类情感特征语音情感识别方法的有效性和稳定性。

关键词：语音情感识别；情感特征分类；改进 D-S 证据理论；证据信任度信息熵；动态先验权重；数据融合
中图分类号：TP389.1 **文献标志码**：A **文章编号**：1001-3695(2012)10-3648-03
doi:10.3969/j.issn.1001-3695.2012.10.012

Research of speech emotion recognition based on emotion features classification

ZHOU Xiao-feng, XIAO Nan-feng, WEN Han
(*School of Computer Science & Engineering, South China University of Technology, Guangzhou* 510006, *China*)

Abstract: Because the speech signals were highly real-time uncertainty, this paper proposed evidences' trust entropy and dynamic prior weights to improve the basic probability function of traditional D-S theory. As the emotion recognition result was not the same by emotion features in different emotions, it presented a classification method of emotion features. In order to realize the fine-grain speech emotion recognition, it used the recognition data of different classification and the improved D-S theory to realize the emotion recognition based on multi-classification emotion features. The improved D-S theory is proved to be effective by simulation. And comparing simulation results show that the multi-classification emotion features are effective and stability.

Key words: speech emotion recognition; emotion features classification; improved D-S theory; evidences' trust entropy; dynamic prior weights; data fusion

随着计算机技术的蓬勃发展，人机交流也变得越来越普遍，情感理解能力日渐成为衡量一个机器是否具有智能的关键因素。目前已经出现能够根据人类特定的命令去执行特定动作的产品，但是要实现人机的自然交互，机器必须能够更准确地理解和区分人类的情感。作为人类日常情感交流的主要方式，人类话语中所携带的情感信息越来越受到人们的重视。本文在建立语音情感数据库的基础上，提取语音信号的基音、共振峰、LPCC 和 MFCC 等情感特征组成语音情感特征集，并将情感特征集分为两类，针对现有 D-S 证据理论融合技术在基本概率赋值方法上的缺陷，提出改进方法。在此基础上，对恐惧、愤怒、悲哀、高兴、惊奇和平静六种情感进行识别。

1 情感识别语音语料库的建立

目前汉语语音情感识别尚未有统一的公开数据库可供研究，因此普遍采用两种方式获得情感语音资料：a）通过具有丰富情感表达能力的人，利用录音软件采集其在各种模拟情感状态下的语音数据作为识别用的语料；b）是通过剪辑影视剧作品中的相关情节得到识别用的语料。方法 a）由于主观影响比较大，对情感状态的理解和表达因人而异，所以本文采用方法 b）。考虑语音的三音子因素，选择合适且尽量无噪声的语音片段，其采样频率为 44 100 Hz，采样位数为 16 位。本文采用的语音情感数据库由恐惧、愤怒、悲哀、高兴、惊奇和平静六种基本情绪组成。每种情绪剪辑 300 句相关语音片段，并将语音数据的顺序打乱，邀请剪辑语音数据以外的 8 人收听语音数据并对其进行分类，最后根据分类结果并结合语音文件的质量，每种情绪选取 200 句语音数据，共 1 200 句语音情感数据。随机抽取 720 句用于训练，480 句用于识别。

2 特征提取及分类

2.1 Mel 倒谱系数的提取[1]

MFCC 自 1980 年由 Davis 等人提出以来，在语音识别领域应用广泛，是语音情感识别比较常用的短时光谱特征。它通过构造人的听觉模型，以语音通过该模型的输出为声学特征，直接通过 DFT（离散傅里叶变换）进行变换。提取 MFCC 的步骤如图 1 所示。由于 MFCC 特征反映的是语音信号的静态特征，而其一阶及二阶差分反映的是语音信号的动态特征，因此本文还提取了 MFCC 的一阶和二阶差分。

$$\mathrm{MFCC}_n = \sum_{k=1}^{M} \ln x'(k) \cos[\pi(k-0.5)n/M] \quad n=1,2,\cdots,L \qquad (1)$$

其中：$x'(k)$ 为第 k 个滤波器的输入功率谱。

收稿日期：2012-03-06；**修回日期**：2012-04-08 **基金项目**：国家自然科学基金资助项目（61171141）

作者简介：周晓凤（1987-），女，河北石家庄人，硕士研究生，主要研究方向为家庭服务机器人、语音情感识别（wo.jiushi@163.com）；肖南峰（1962-），男，江西南昌人，教授，博导，博士，主要研究方向为仿人机器人；文翰（1977-），男，湖南益阳人，博士研究生，主要研究方向为机器学习.

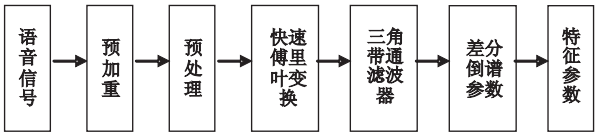

图1　MFCC提取流程

2.2　LPCC 参数的提取

LPCC 是在 LPC(线性预测系数)的基础上得到的。线性预测分析认为,语音信号样点之间存在相关性,可以用过去的若干个样本点或它们的线性组合预测现在或将来的样本值。它可以提供一个很好的声道模型,而且只要十几个倒谱特征参数就能很好地反映出语音信号的共振峰特性。一般倒谱分析阶数取 16 左右就能较好地表征语音的特征参数,本文对 LPC 的阶数取 14、15、16、17 时,分别进行识别实验,LPC 阶数为 14 时,效果最佳,因此本文 LPC 的阶数取 14。

2.3　其他特征参数的提取

短时能量和短时过零率特征能够用来很好地区分清音段和浊音段。清音的短时能量明显小于浊音的短时能量,而清音的过零率却高于浊音的过零率。基音频率[2]反映的是整个语音情感信号的语调轨迹,它较好地体现了人的情感变化,根据各情感状态间语调的差异,即从听觉上就可以判断基音频率的变化。

共振峰是声道的共振频率,是反映声道参数的一个重要参数。共振峰会随着声道的状态而改变。不同的情感发音会使舌头、牙齿和嘴唇等发声器官的状态发生变化,从而导致声道状态的变化,共振峰也会随之改变。在发声特征中,共振峰可以作为判断不同情感语音的一个重要特征[1]。

通过上述分析,本文将获取的语音情感特征,根据其在语音情感识别过程中所起的不同作用进行分类,具体特征及其分类如表 1 所示。本文为每一种情感的各类情感特征分别单独进行 HMM 训练[3]。对各类特征矩阵利用改进的 D-S 证据理论进行数据融合。

表 1　情感特征分类

听觉模型	MFCC 及其统计特征(一阶差分、二阶差分及其最大值、最小值、均值、中值、方差) 基音及其统计特征(一阶差分、二阶差分及其最大值、最小值、中值和基频微扰[4]) 短时能量及其统计特征(一阶差分及其最大值、最小值、中值) 短时过零率及其统计特征(一阶差分及其最大值、最小值、中值)
发声模型	LPCC 及其统计特征(一阶差分、二阶差分及其最大值、最小值、均值、中值、方差) 第一、二、三共振峰及其统计特征(一阶差分及其最大值、最小值、中值)

此外,为了避免不同类型的特征取值范围的不同所导致的特征参数取值影响 HMM 的训练和识别,本文根据式(2)~(4)对基音、共振峰和短时特征进行了归一化处理[5],从而降低它们因取值范围不同带来的差异。其中,所有归一化后的特征具有零均值和单位方差的特点。

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_{ik}\quad k=1,2,\cdots,l \tag{2}$$

$$\delta^2=\frac{1}{N}\sum_{i=1}^{N}(x_{ik}-x_k)^2 \tag{3}$$

$$x_{ik}=\frac{x_{ik}-x_k}{\delta_k} \tag{4}$$

3　改进 D-S 证据理论

证据理论由 Dempster 等人提出并发展,也称 D-S 理论。D-S 证据理论是建立在一个完备集合的识别框架 U 上的,其中 U 包含了某个问题的所有可能答案且元素之间是互不相交的基本命题,即识别结果只能取 U 中的某一个基本命题。D-S 证据理论通过基本概率分配函数(通常为 m 函数)来为该框架的各个基本命题分配信任度。D-S 证据理论一个关键和重要的问题就是如何为各个基本命题分配初始信任度,即初始概率密度。由于证据之间的信息有可能有交叉和冲突的情况,在利用传统 D-S 证据理论时,若存在高冲突的证据则会产生不合理的结果。本文在基本概率赋值方法上对该问题进行了改进。

若 A 为 U 中的一个元素,则 $A\in U$,则 m 满足

$$\begin{cases}m(\varnothing)=0\\0\leqslant m(A)\leqslant 1,\ \forall A\subseteq U\\\sum\limits_{A\subseteq U}m(A)=1\end{cases} \tag{5}$$

其中:$m(\varnothing)$ 表示 A 为空集时初始概率密度为 0,且 A 的幂集 2^A 中元素初始概率密度总和为 1。如果 $m(A)>0$,则称 A 为证据的焦元。

假设识别框架 U 中包含 n 个证据,且任意两个证据 i、j 的基本概率密度为 m_i、m_j,焦元分别为 A_i、A_j,由文献[6],证据 i 和 j 之间的距离为

$$d(m_i,m_j)=\sqrt{\frac{1}{2}(m_i-m_j)^{\mathrm{T}}D(m_i-m_j)} \tag{6}$$

其中:m_i、m_j 分别为由 m_i、m_j 在 2^U 中所有元素上的取值组成的行向量;$\boldsymbol{D}$ 为 $2^n\times 2^n$ 的矩阵,它的元素为 $\boldsymbol{D}(A,B)=\frac{|A\cap B|}{|A\cup B|}$,其中,$A,B\in 2^U$,$|A\cap B|$ 为 A 和 B 中命题交集的个数,同理,$|A\cup B|$ 为 A 和 B 中命题并集的个数。若 $d(m_1,m_2)$ 愈大,则 m_1 和 m_2 间的差别愈大,其中,若 $i=j$,则 $d(m_i,m_j)=0$。

由文献[7]定义证据 m_1 和 m_1 之间的相似系数为

$$s_{ij}=1-d_{ij} \tag{7}$$

表示为矩阵的形式:$\boldsymbol{S}=[s_1,s_2,\cdots,s_n]$,其中:

$$s_i=[1-d_{i1},1-d_{i2},\cdots,1-d_{in}]^{\mathrm{T}} \tag{8}$$

定义证据总体相似度为

$$\zeta_i=\sum_{k=1}^{n}s_{ik} \tag{9}$$

其中:ζ_i 反映了证据 i 与其他证据的相似程度,ζ_i 越大,说明证据 i 与其他证据之间的相似度越大,则与其他证据之间的差别越小,证据 i 所受的支持程度越大。

定义证据 i 总体相似度的信息熵为

$$e_i=-\sum_{k\in\{1,2,\cdots,n\}}\zeta_k\ln\zeta_k \tag{10}$$

使用折扣率[8] β_i 对证据总体相似度的信息熵进行调整:$\bar{E}=e_i/\max(e_i)\ (i=1,2,\cdots,n)$,对调整后的 $\bar{E}$ 进行归一化后作为证据的后验权重。

由式(8)~(10)可以看出,$\bar{E}$ 同时反映了证据的可信度和不确定性,能够从信息熵的角度反映证据 i 的支持程度,很好地体现了证据的动态信息。如果证据 i 与其他证据相似度比较大,则这些证据的相互支持程度就比较大,随着样本的增多,该证据 i 总体相似度的信息熵值不会明显的下降,则证据 i 在决策的过程中所占的比重就大,后验权重也越大;反之亦然。

根据证据的先验权重和后验权重计算证据 i 的复合权重:

$$w_i=\sqrt{\alpha_i\bar{e}_i}\qquad i=1,2,\cdots,n \tag{11}$$

$$\delta_i=w_i/\mathrm{sum}(w_i) \tag{12}$$

其中:δ_i 满足 $\sum\limits_{i=1}^{n}\delta_i=1$。$\delta_i$ 为证据 i 的归一化复合权重,它代表

了证据 i 的复合可信程度。

采用文献中的算例对本文方法进行了验证,并与文献中的方法进行了比较。算例的数据如表 2 所示,$m_i(A)$、$m_i(B)$ 和 $m_i(C)$ 分别表示各个传感器对识别目标 A、B 和 C 的支持程度,已知 5 个传感器的先验权重分别为 0.15、0.15、0.20、0.35、0.15。

表 2 5 个传感器的基本概率分配函数

传感器	目标 A	目标 B	目标 C
1	$m_1(A)=0$	$m_1(B)=0.90$	$m_1(C)=0.10$
2	$m_2(A)=0.5$	$m_2(B)=0.20$	$m_2(C)=0.30$
3	$m_3(A)=0.55$	$m_3(B)=0.10$	$m_3(C)=0.35$
4	$m_4(A)=0.55$	$m_4(B)=0.10$	$m_4(C)=0.35$
5	$m_5(A)=0.55$	$m_5(B)=0.10$	$m_5(C)=0.35$

由表 2 可知,只有传感器 1 对目标 B 的支持度较高,而其他传感器对目标 A 具有较高的支持度,可知传感器 1 获取的信息与实际有较大偏差,产生了冲突证据。按照正常推断目标应该为 A。传感器由两个增加到 5 个的过程中,本文算法与文献[7]算法的处理结果对比如表 3 所示。

表 3 实验对比结果

证据组合方法	m_1,m_2	time	m_1,m_2,m_3	time	m_1,m_2,m_3,m_4	time	m_1,m_2,m_3,m_4,m_5	time
文献[7]	$m(A)=0.1215$ $m(B)=0.8240$ $m(C)=0.0545$	0.007983	$m(A)=0.4185$ $m(B)=0.4784$ $m(C)=0.1031$	0.009135	$m(A)=0.7603$ $m(B)=0.1301$ $m(C)=0.1096$	0.009288	$m(A)=0.8771$ $m(B)=0.0427$ $m(C)=0.0802$	0.009764
本文	$m(A)=0.1519$ $m(B)=0.7503$ $m(C)=0.0978$	0.011876	$m(A)=0.4559$ $m(B)=0.4354$ $m(C)=0.1087$	0.011901	$m(A)=0.8197$ $m(B)=0.0707$ $m(C)=0.1096$	0.012523	$m(A)=0.9101$ $m(B)=0.0147$ $m(C)=0.0752$	0.012577

由表 3 可知,文献[7]考虑了证据的先验权重和后验权重,随着目标 A 的支持证据不断增加,$m(A)$ 也在不断增加,但融合结果收敛速度较慢。本文方法在增加至 3 个传感器的情况下就已识别目标 A,而文献[7]方法在 4 个传感器的情况下才识别目标 A,且本文方法的收敛速度明显快于文献[7]的方法,且对证据 A 的支持程度也比较大。本文不但考虑了证据的先验权重和后验权重,还引入了证据相似度的信息熵,考虑了信号动态变化的趋势,因此,本文方法收敛速度更快,融合效果更佳,稳定性更强。

表 3 中 time 列表示程序运行时间,实验在主频为 2.53 GHz、内存为 2 GB 的环境中进行测试,本文方法所需时间仅比文献[7]方法多 0.004 ~ 0.006 s,且随着传感器数目的增加,所需时间的增长速度慢于文献[7]方法。为便于观察,在传感器不断增加的过程中,将本文方法和文献[7]方法的识别效果绘制成图,如图 2 和 3 所示,横轴代表使用的传感器;图 2 纵轴代表使用的时间,图 3 纵轴代表对证据 A 的支持程度。

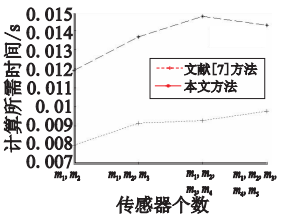

图2 使用时间对比

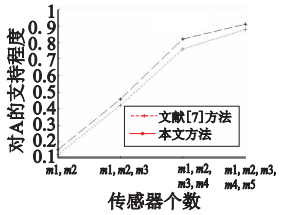

图3 对证据A的支持程度对比

4 语音情感识别实验

为了验证分类情感特征和改进的 D-S 证据理论在语音情感识别中的识别效果,对基于单类情感特征的语音情感识别方法进行对比实验。在情感特征提取后,需要对各类情感特征矩阵进行处理,以降低它的维度,实验采用 PCA[4] 方法对各类情感特征矩阵进行降维。由图 4 可知,当特征空间降到八维时,平均识别率达到最高。因此采用 PCA 降维时,将原始特征空间降到八维。

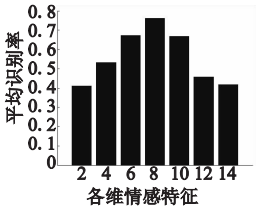

图4 平均识别率与维数的关系

在本文的语音情感识别中,识别框架 $U=\{\{恐惧\},\{愤怒\},\{悲伤\},\{高兴\},\{惊奇\},\{平静\}\}$,其中提取的各类情感特征相当于证据。各证据的先验权重是由各类情感特征的识别率决定的,考虑到各类情感特征对不同的情感状态具有不同的支持度,本文提出了动态先验权重的思想,即各类情感特征对不同的情感状态进行识别时具有不同的先验权重。动态先验权重的确定方法为:单独使用每类情感特征进行语音情感识别[3],将各类对情感状态 $i(i=1,2,\cdots,6)$ 取得的识别率进行归一化的结果作为情感特征 i 的先验权重。情感特征对各情感状态的识别率越大,则其所占的权重越大。其具体数值如表 4 所示。

表 4 MFCC 特征类和 LPCC 特征类的先验权重

特征类	angry	fear	happy	sad	surprise	peace
MFCC	0.3281	0.6410	0.3448	0.5103	0.3472	0.4737
LPCC	0.6719	0.3590	0.6552	0.4897	0.6528	0.5263

由表 4 可知,MFCC 特征类对恐惧这种情感状态的支持度明显高于 LPCC 特征类,而 LPCC 特征类对愤怒、高兴和惊奇这两种情感状态的支持度明显高于 MFCC 特征类,两类特征对悲伤和平静这两种情感状态的支持度相差不大。

由本文提出的改进 D-S 证据理论算出各情感状态的后验概率,结合先验概率,并对单类情感特征的语音情感识别概率进行处理,综合两类情感特征对每个情感状态的不同支持度,对语音情感进行识别,达到更稳定的效果。

图 5 显示了本文方法与基于单类情感特征语音情感识别方法的结果对比。由图 5 可知,在两类情感特征的识别结果相差很大时,本文方法能够有效提高支持度情感特征类的权重,以降低另一类情感特征的影响。在两类情感特征的识别率差别不是很大时,本文方法能够在一定程度上提高情感状态的识别率。

图5 利用单类情感特征的语音情感识别与本文方法的对比

5 结束语

本文提出了一种将语音情感特征进行分类,并将各类情感特征的支持度矩阵利用改进的 D-S 证据理论进行决策级融合的语音情感识别方法。在利用 D-S 证据理论进行数据融合时,由于情感特征对不同情感状态具有不同的支持度,动态先验权重被引入以降低情感特征对情感状态支持度 (下转第 3676 页)

(上接第 3650 页)的差异。通过与基于单类语音情感特征的语音情感识别的方法进行对比实验,结果表明,本文提出的方法在一定程度上提高了语音情感识别方法的识别率和稳定性。在下一步的研究中,还需进一步细化情感特征的分类,从多个侧面更好地体现语音信号的情感信息,提高语音情感识别的识别率。

参考文献:

[1] 刘幺和,宋庭新. 语音识别与控制应用技术[M]. 北京: 科学出版社, 2008:3-40.

[2] 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.

[3] 国辛纯, 郭继昌, 窦修全. 基于 HMM 的语音信号情感识别研究[J]. 电子测量技术,2006,29(5):69-70.

[4] 罗宪华, 杨大利, 徐明星. 面向非特定人的语音情感识别特征研究[J]. 北京信息科技大学学报:自然科学版,2011,26(2):72-76.

[5] THEO S, KONSTANTINOS K. 模式识别[M]. 北京: 电子工业出版社,2010:138-160.

[6] 孟光磊, 龚光红. 证据源权重的计算及其在证据融合中的应用[J]. 北京航空航天大学学报,2010,36(11):1365-1368.

[7] 李军,锁斌,李顺. 基于证据理论的多传感器加权融合改进方法[J]. 计算机测量与控制,2011,19(10):2592-2595.

[8] 祁瑞敏,王新,张国栋. 改进的信息融合算法在矿井提升机健康诊断中的应用[J]. 煤矿机械,2011,32(5):234-235.

[9] 陈建厦, 李翠华. 语音情感识别的研究进展[J]. 计算机工程,2005,31(13):35-37.

[10] 余建潮, 张瑞林. 基于 MFCC 和 LPCC 的说话人识别[J]. 计算机工程与设计,2009,30(5):1189-1191.