实用语音情感的特征分析与识别的研究

黄程韦*[①]　赵　艳[①]　金　赟[①②]　于寅骅[①]　赵　力[①]
[①](东南大学水声信号处理教育部重点实验室　南京　210096)
[②](徐州师范大学物理与电子工程学院　徐州　221116)

摘　要：该文针对语音情感识别在实际中的应用，研究了烦躁等实用语音情感的分析与识别。通过计算机游戏诱发的方式采集了高自然度的语音情感数据，提取了 74 种情感特征，分析了韵律特征、音质特征与情感维度之间的关系，对烦躁等实用语音情感的声学特征进行了评价与选择，提出了针对实际应用环境的可拒判的实用语音情感识别方法。实验结果表明，文中采用的语音情感特征，能较好识别烦躁等实用语音情感，平均识别率达到 75%以上。可拒判的实用语音情感识别方法，对模糊的和未知的情感类别的分类进行了合理的决策，在语音情感的实际应用中具有重要的意义。

关键词：语音识别；实用语音情感；韵律特征；音质特征；拒判方法
中图分类号：TP391.42　　**文献标识码**：A　　**文章编号**：1009-5896(2011)01-0112-05
DOI: 10.3724/SP.J.1146.2009.00886

A Study on Feature Analysis and Recognition of Practical Speech Emotion

Huang Cheng-wei[①]　Zhao Yan[①]　Jin Yun[①②]　Yu Yin-hua[①]　Zhao Li[①]
[①](*Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education*, *Southeast University*, *Nanjing* 210096, *China*)
[②](*School of Physics and Electronics Engineering*, *Xuzhou Normal University*, *Xuzhou* 221116, *China*)

Abstract: Practical speech emotions as impatience and happiness are studied especially for evaluation of emotional well-being in real world applications. Induced natural speech emotion data is collected with a computer game, 74 emotion features are extracted, prosody features and voice quality features are analyzed according to dimensional emotion model, evaluation and selection of acoustic features are carried out for practical emotions in this paper, a method of practical speech emotion classification with rejection decision is proposed for real world occasions. The experiment results show, the speech features analyzed in this paper are suitable for classification of practical speech emotions like impatience and happiness, average recognition rate is above 75%, and the method of emotion classification with rejection decision is necessary for the proper recognition decision of ambiguous or unknown emotion samples, especially for the real world challenges.

Key words: Speech recognition; Practical speech emotion; Prosody features; Voice quality features; Rejection decision

1　引言

人类的情绪能力在人们的工作和生活的各个方面起到了不可或缺的重要作用，近年来与情绪相关的研究已成为国际上多个学科的研究热点[1–3]。情绪状态的自动评估具有重要的实际意义，特别是在航空航天等军事应用领域中，长时间的、枯燥的、高强度的任务会使相关人员面临严酷的生理以及心理考验，引发一些负面的情绪。目前国内外对情感识别的研究，主要集中在几类基本情感的识别上[4–11]，尚不能满足实际应用中的需求。本文针对实际应用中的需求，重点研究了语音通话中“烦躁”情感的自动识别。在航空航天等应用领域，长时间的飞行任务中，由于枯燥的重复性作业、狭小的机舱空间、以及高度紧张的精神状态，都容易引发机组人员的烦躁情绪。烦躁情绪出现后，如果不妥善的处理，对人员的工作能力会造成重大的影响，甚至引起人为的疏忽导致事故。因此，对烦躁情感的自动识别研究具有重要的实际意义。

实际应用当中对语音情感识别技术提出了诸多

2009-06-16 收到，2010-10-19 改回
国家自然科学基金(60472058，60975017，51075068)和江苏省自然科学基金(BK2008291)资助课题
*通信作者：黄程韦　Chengwei.Huang@yahoo.com.cn

挑战。以往基于表演语料的识别系统，在实际条件下，系统的情感模型与真实的情感数据不能符合得很好，导致了识别正确率的显著下降。在本文中，我们将通过心理学实验的方法来采集实用语音情感的诱发数据，尽可能地使训练数据接近真实的情感数据。在实际环境中出现的情感具有模糊性和多样性，在实用语音情感的识别中，有必要考虑可拒判的识别方法。传统的识别方法，是将出现的样本硬性地划分为已知类别中的某一类，在实际中存在较多模糊不清的情感样本时，分类的可信度就较差，误判的概率就较高。因此本文采用可拒判的实用语音情感识别方法，对于不确定的或未知的情感样本，分类器给出拒绝判断的识别结果，即不属于需要检测的实用语音情感类别中的任何一类。

2　实用语音情感的诱发

根据 Scherer 的观点[5]，人类声音中蕴含的情感信息，受到无意识的心理状态变化的影响，以及社会文化导致的有意识的说话习惯的控制。然而在目前的语音情感数据的采集中，广泛使用的是表演的方式，在实际的语音通话和自然交谈中，说话人的情感对语音产生的影响，常常是不受说话人控制的，通常也不服务于有意识的交流目的[11]，而是反映了说话人潜在的心理状态的变化。相反，演员能通过刻意的控制声音的变化来表演所需要的情感。为了能更好地研究实际环境中的情感语音，有必要采集除表演语音以外的，较高自然度的情感数据，在本文中，通过计算机游戏诱发情感的方法[11,12]来采集实用语音情感数据。

在实验心理学中，计算机游戏通过画面和音乐的视觉、听觉刺激，能提供一个互动的、具有较强感染力的人机交互环境，能够有效地诱发出被试人员的正面与负面的情感。特别是在游戏接连胜利时，被试人员由于在游戏虚拟场景中的成功与满足，被诱发出喜悦的情感；在游戏连续失败时，被试人员在虚拟场景中受到挫折，容易引发包括烦躁在内的负面情感。在进行较长时间的实验过程中，重复性的游戏操作和失败，能顺利地诱发烦躁情感。对于语句文本的设计，考虑到烦躁等实用语音情感识别的一个主要应用领域为长期的航空、航天和航海任务所引发的负面情绪的评估，20 句无情感倾向性的工作用语短句选自国际海事组织(IMO)发布的《标准航海通信用语》(SMCP)。

3　情感语音的特征分析

3.1 特征提取

情感特征的优劣对情感最终识别效果的好坏有非常重要的影响，如何提取和选择能反映情感变化的语音特征，是目前语音情感识别领域最重要的问题之一[13,14]。近年来，Johnstone 等人[11,12]的研究证明语音信号中的音质特征，不仅与情感的“效价维”关系密切，而且也能够部分反映 3 维维度模型中的“控制维”的信息。用于识别和建模的特征向量一般有两种构造方法，全局统计特征和动态特征。由于动态特征对音位信息的依赖性太强，不利于建立与文本无关的情感识别系统，因此在本文中使用了 74 个全局统计特征，在下面列出，其中前 36 个特征为韵律特征，后 38 个特征为音质特征。

特征 1–10：短时能量及其差分的均值、最大值、最小值、中值、方差；

特征 11–25：基音及其一阶、二阶差分的均值、最大值、最小值、中值、方差；特征 26：基音范围；

特征 27–36：发音帧数、不发音帧数、不发音帧数和发音帧数之比、发音帧数和总帧数之比、发音区域数、不发音区域数、发音区域数和不发音区域数之比、发音区域数和总区域数之比、最长发音区域数、最长不发音区域数；

特征 37–66：第 1、第 2、第 3 共振峰及其一阶差分的均值、最大值、最小值、中值、方差；

特征 67–69：250 Hz 以下谱能量百分比、650 Hz 以下谱能量百分比、4 kHz 以上谱能量百分比。

特征 70–74：谐波噪声比(HNR)的均值、最大值、最小值、中值、方差。

其中谐波噪声比用来做为反映情感变化的音质特征[15]。负面与正面的情绪往往在愉悦度上具有较大的差异，因而与情感的愉悦度关系密切的音质特征对识别实用语音情感具有重要的价值。

3.2 基于情感维度空间模型的特征分析

根据文献[11,12,14]的研究结果，韵律特征主要和激活度的相关性较大，音质特征与愉悦度的相关性较大。我们使用 PCA 方法分别进行愉悦度和激活度上的特征空间分析，截取 PCA 的前两个维度构成 2 维特征空间，图 1 为韵律特征构成的 2 维 PCA 空间，图 2 为音质特征构成的 2 维 PCA 空间。

可以看到，在仅使用韵律特征时，平静和其余两种情感能较好的区分开来。然而烦躁和喜悦两种情感，在激活度上差别相对较小，在仅使用韵律特征时，两种情感的样本分布区域重叠的较多。这与韵律特征主要和激活维对应的理论是一致的。使用音质特征后，烦躁和喜悦的样本之间能够获得较好的区分，音质特征的使用对区分烦躁和喜悦两种愉悦度上距离大的情感是有效的，说明音质特征与愉悦度的相关性较大。综合使用 74 个韵律特征和音质特征，如图 3 所示，烦躁、喜悦和平静 3 种情感的

样本分布得到了较好的区分。

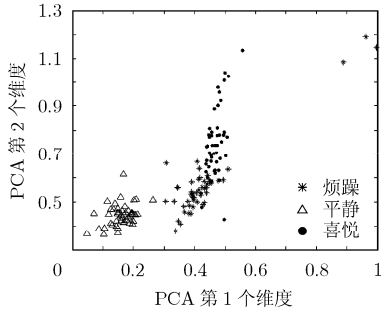

图 1 韵律特征空间中的样本分布

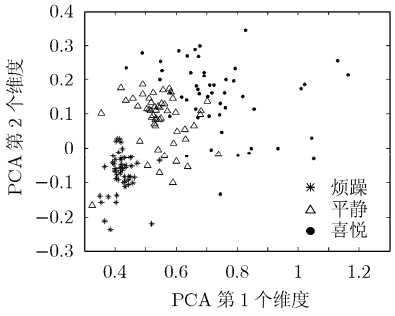

图 2 音质特征空间中的样本分布

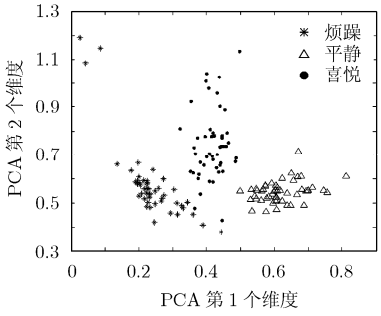

图 3 韵律与音质特征空间中的样本分布

3.3 实用语音情感的特征评价与特征选择

情感特征的选择一直是语音情感识别中最受重视的问题之一，对一个特征的优劣的评价，我们考虑两个方面：特征的均值，以及特征的方差。综合考虑这两个方面的因素，采用 fisher 准则进行特征评价[16,17]。对烦躁、喜悦、平静 3 种情感选择出的前 10 个最佳特征如表 1 所示。

表 1 前 10 个最佳特征

重要程度排序	特征
1	250 Hz以下谱能量的百分比
2	基音一阶差分的均值
3	基音的均值
4	第1共振峰的中值
5	第1共振峰的最小值
6	短时能量的方差
7	第2共振峰的均值
8	发音帧数和总帧数之比
9	谐波噪声比均值
10	650 Hz以下谱能量的百分比

4 可拒判的实用语音情感识别方法

情感样本在特征空间里的分布，可以用多个高斯函数的叠加来描述。理论上来说，只要混合足够的高斯分量，高斯混合模型(GMM)能够拟合任意的概率密度分布函数。本文采用 GMM 对烦躁、喜悦和平静 3 种情感进行建模，每种情感对应一个 GMM 模型，通过最大后验概率准则判决。x_i 表示第 i 条语句样本，λ_j 表示情感类别，最大后验概率可以表示为

$$p(\lambda_j \mid x_i) = \frac{p(x_i \mid \lambda_j) P(\lambda_j)}{P(x_i)} \tag{1}$$

其中 $p(x_i \mid \lambda_j)$ 通过每个情感的 GMM 模型得到。对于给定的语句样本，特征矢量出现的概率是一个常量，假设每种情感等概率地出现，C 为情感类别数。

$$P(\lambda_j) = \frac{1}{C}, \quad 1 \le i \le C \tag{2}$$

那么，待识别的样本可以判决为

$$j^* = \arg\max_j p(x_i \mid \lambda_j) \tag{3}$$

其中 $j*$ 表示样本所属的类别。

针对实际环境下情感的模糊与不确定性，实用语音情感种类的多样性，有必要研究可以拒判的识别实用语音情感方法。下面将采用一种基于似然概率模糊熵的拒判方法，采用模糊熵来对样本与情感类别之间的符合程度进行度量，从而实现对未知类别样本的拒判。待识别的样本到达时，分别通过 C 种情感的 GMM 模型，得到 C 个 GMM 似然概率密度值，以 GMM 似然概率密度值映射到 0 到 1 之间作为第 i 个样本归属于第 j 个情感类别的隶属度 $\mu_j(x_i)$：

$$\mu_j(x_i) = \frac{\arctan\left(p(x_i \mid \lambda_j)/10\right)}{\pi/2} \tag{4}$$

其中采用的投影函数为

$$y = \frac{\arctan(x/10)}{\pi/2} \tag{5}$$

对于第 j 个情感类别的所有可能的样本构成的模糊集 $E_j = \{x_1, x_2, \cdots, x_n\}$，其隶属度分别为 $\mu_j(x_1)$, $\mu_j(x_2), \cdots, \mu_j(x_n)$，令其模糊熵为 $e(\mu_j(x_i))$，类似于随机熵的证明，可以得到模糊熵的表达式为[18]

$$e(\mu_j(x_i)) = -K \ln \mu_j(x_i) \tag{6}$$

其中 K 是大于 0 的数。将式(4)代入式(6)得，第 i 个样本归属于第 j 个情感类别的模糊熵为

$$e(\mu_j(x_i)) = -K\left(\ln \arctan\left(p(x_i \mid \lambda_j)/10\right) - \ln(\pi/2)\right) \tag{7}$$

对第 i 个待识别样本的 C 个似然概率值构成的判决集合的平均模糊熵评价为

$$S(x_i)=\frac{1}{C}\sum_{j=1}^{C}\mu_{\mathrm{j}}(x_i)e(\mu_j(x_i)) \tag{8}$$

将式(7)代入式(8)有

$$S(x_i)=-\frac{2K}{\pi C}\sum_{j=1}^{C}\arctan\left(p(x_i|\lambda_j)/10\right)\cdot\left(\ln\arctan\left(p(x_i|\lambda_j)/10\right)-\ln(\pi/2)\right) \tag{9}$$

对烦躁、喜悦和平静 3 种情感类别的 GMM 模型，可以得到 3 个 GMM 似然概率密度值，分别代表样本与 3 个情感类别的符合程度。似然概率密度值构成的判决集合的模糊熵越高表示样本属于烦躁、喜悦和平静 3 种情感的不确定程度越大，当模糊熵超过一定阈值 Th 时则发生拒判，常数 K 取 $\pi/2$。

$$S(x_i)>\mathrm{Th} \tag{10}$$

将式(9)代入式(10)，即

$$\frac{1}{C}\sum_{j=1}^{C}\arctan\left(p(x_i|\lambda_j)/10\right)\cdot\left(\ln(\pi/2)-\ln\arctan\left(p(x_i|\lambda_j)/10\right)\right)>\mathrm{Th} \tag{11}$$

其中 Th 为实验中确定的模糊熵阈值。阈值的选取既要保证待识别的目标情感类别得到正确的识别，又要兼顾未知的样本不确定的情感得到拒判。

5　实验测试结果

进行与说话人无关文本无关的情感识别测试，每种情感随机抽取 400 条，分为两组，一组 300 条样本，进行 GMM 情感模型的训练，3 种情感共计 900 条，另一组 100 条样本，用于测试识别率，3 种情感共计 300 条。在诱发语音库的原始语音中，通过听辨实验被剔除的情感语句共有 479 条，这些语句被认为是情感隶属度较低的数据，选取其中隶属度最低的 100 条，作为不确定的未知情感类别样本，用于拒判测试。分别采用 PCA 方法的前 10 个特征维度和最佳特征组选择方法的前 10 个最佳特征，使用可拒判的实用语音情感识别方法，对烦躁、喜悦和平静 3 种情感的识别率进行测试。模糊熵阈值的设置关系到样本的拒判，设定得过低，则对不确定样本的拒判效果不明显。设定得过高，则拒判的过多，会使得系统平均识别率降低。当部分样本离已知的情感模型距离较远时需要拒判，同时拒判也会使得某些测试样本不能得到正确识别。所以应该在保证烦躁、喜悦、平静等 3 个类别能够获得满意的识别率的前提下，调节模糊熵阈值。当平均识别率发生明显的下降时，此时的阈值为上限，实验中模糊熵阈值设为 0.1。

本实验中的训练样本数与测试样本数比例为 3:1，为了获得更充分的实验测试数据，将训练样本集中的 900 条样本随机等分成 3 份后，与测试样本集中的 300 条样本轮换，进行轮换测试。平均测试识别结果如表 2 和表 3 所示。

表 2 PCA 方法识别结果

测试样本	识别结果(%)			
	烦躁	喜悦	平静	拒判
烦躁	**74.5**	6.25	12.25	7.0
喜悦	8.5	**71.75**	15.25	4.5
平静	7.75	5.5	**83.5**	3.25
不确定样本	4.75	15.25	21.5	**58.5**

表 3 最佳特征组识别结果

测试样本	识别结果(%)			
	烦躁	喜悦	平静	拒判
烦躁	**75.25**	8.75	10.0	6.0
喜悦	7.5	**70.25**	16.5	5.75
平静	7.5	6.75	**81.5**	4.25
不确定样本	12.75	15.25	11.75	**60.25**

根据识别测试结果，烦躁、喜悦和平静 3 种情感，在本实验中容易发生混淆的是烦躁和平静，喜悦和平静。在情感的维度空间模型中喜悦与烦躁位于愉悦度的两端，差别较大，而平静位于它们之间，因此相对来说喜悦容易与平静混淆，烦躁容易与平静混淆。从特征空间中的样本分布情况来看，平静类别的样本分布明显要比烦躁和喜悦的样本分布更为集中，平静情感的样本具有较高的一致性，因此其识别率较烦躁和喜悦高。

6　结论

为进行烦躁、喜悦和平静等实用语音情感的识别，本文提取了 74 个语音情感特征，平均识别率达到 75%以上，证实了本文中的情感特征用于识别烦躁等实用语音情感是有效的。通过 PCA 方法进行了基于情感维度空间的特征分析，结果显示韵律特征与激活度相关性较大，音质特征与愉悦度的相关性较大。通过 fisher 判别准则，对情感特征进行了评价，结果显示表 1 中的 10 个特征能较好区分烦躁等实用语音情感。基于似然概率模糊熵的可拒判的实用语音情感识别方法，能对模糊和未知的情感类别的分类进行合理的决策，可拒判识别方法在语音情感的实际应用中是必要的。

参 考 文 献

[1] Spellman B A and Willingham D T. Current Directions in Cognitive Science. Boston: Allyn & Bacon, 2007: 1–3.

[2] Picard R W. Affective Computing. Cambridge: MIT Press, 1997, Chapter 6.

[3] Vinciarelli A, Pantic M, Bourlard H, and Pentland A. Social signal processing: survey of an emerging domain. *Image Vision Computing*, 2009, 27(12): 1743–1759.

[4] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, and Taylor J G. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 2001, 18(1): 32–80.

[5] Scherer K R. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 2003, 40(1/2): 227–256.

[6] Zeng Z, Pantic M, Roisman G I, and Huang T. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(1): 39–58.

[7] Casale S, Russo A, Scebba G, and Serrano S. Speech emotion classification using machine learning algorithms. 2008 IEEE International Conference on Semantic Computing. Santa Clara, CA, USA, Aug. 4–7, 2008: 158–165.

[8] Zhao Yan, Zhao Li, Zou Cai-rong, and Yu Yin-hua. Speech emotion recognition using modified quadratic discriminatioin function. *Journal of Electronics* (*China*), 2008, 25(6): 840–844.

[9] 韩文静, 李海峰, 韩纪庆. 基于长短时特征融合的语音情感识别方法. 清华大学学报(自然科学版), 2008, 48(S1): 708–714.
Han Wen-jing, Li Hai-feng, and Han Ji-qing. Speech emotion recognition with combined short and long term features. *Journal of Tsinghua University* (*Science and Technology*), 2008, 48(S1): 708–714.

[10] Pao Tsang-long, Chen Yu-te, and Yeh Jun-heng. Emotion recognition and evaluation from mandarin speech signals. *International Journal of Innovative Computing, Information and Control*, 2008, 4(7): 1695–1709.

[11] Johnstone T. Emotional speech elicited using computer games. Fourth International Conference on Spoken Language, Philadelphia, PA, USA, 1996, Vol. 3: 1985–1988.

[12] Johnstone T, Van Reekum C M, hird K, and Kirsner K, *et al*.. Affective speech elicited with a computer game. *Emotion*, 2005, 5(4): 513–518.

[13] 王治平，赵力，邹采荣. 基于基音参数规整及统计分布模型距离的语音情感识别. 声学学报，2006, 31(1): 28–34.
Wang Zhi-ping, Zhao Li, and Zou Cai-rong. Emotion speech recognition based on modified parameter and distance of statistical model of pitch. *Acta Acustica*, 2006, 31(1): 28–34.

[14] Tato R S, Kompe R, and Pardo J M. Emotional space improves emotion recognition. ICSLP, Denver, Colorado, USA, 2002: 2029–2032.

[15] Borchert M and Dusterhoft A. Emotions in speech – experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. Proceeding of NLP-KE'05, Wuhan, China, 2005: 147–151.

[16] Xiao Zhong-zhe, Dellandrea E, and Dou Wei-bei, *et al*.. Features extraction and selection for emotional speech classification. IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, 2005: 411–416.

[17] Ho T and Basu M. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 24(3): 289–300.

[18] 王治平. 情感语音信号特征分析与识别. [博士论文], 东南大学, 2004.
Wang Zhi-ping. Feature analysis and emotion recognition in emotional speech.[D.Ph. dissertation], Southeast University, 2004.

黄程韦： 男，1984 年生，博士生，研究方向为语音情感识别等.
赵 艳： 女，1978 年生，博士生，研究方向为语音信号处理.
金 赟： 男，1979 年生，博士生，研究方向为耳语音信号处理等.
于寅骅： 男，1984 年生，硕士生，研究方向为语音情感识别等.
赵 力： 男，1958 年生，教授，博士生导师，研究方向为语音与音频信号处理、图像与视频信号处理、情感信息处理等.