

# 基于 HMM 和 ANN 混合模型的语音情感识别研究

林巧民<sup>1,2</sup>, 齐柱柱<sup>1</sup>

(1. 南京邮电大学 计算机学院, 江苏 南京 210023;  
2. 南京邮电大学 教育科学与技术学院, 江苏 南京 210003)

**摘要:**随着情感计算成为人工智能的一个重要方向,语音情感识别作为情感计算的一个重要部分,已经逐渐成为模式识别领域研究的热点之一。随着研究的不断深入,单独使用某一种模式识别时效果并不理想。为了提高识别率,提出了一种将隐马尔可夫模型(HMM)和径向基函数神经网络(RBF)相结合的方法。这种方法对不同情感状态分别设计 HMM 模型,经过维特比(Viterbi)算法得到最优状态序列,然后对得到的状态序列进行时间规整,以便生成等维的特征矢量,将其作为 RBF 模型的输入进行语音情感识别,最后的识别结果由 RBF 模型给出。实验结果表明,与孤立 HMM 相比,该方法在识别率上有较大的提高。

**关键词:**情感计算;人工智能;隐马尔可夫模型;神经网络;语音情感识别

中图分类号:TN912.34

文献标识码:A

文章编号:1673-629X(2018)10-0074-05

doi:10.3969/j.issn.1673-629X.2018.10.015

## Research on Speech Emotion Recognition Based on HMM and ANN Mixed Model

LIN Qiao-min<sup>1,2</sup>, QI Zhu-zhu<sup>1</sup>

(1. School of Computer Science, Nanjing University of Posts & Telecommunications, Nanjing 210023, China;  
2. School of Educational Science and Technology, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

**Abstract:** As emotion calculation becomes an important direction of artificial intelligence, speech emotion recognition, as an important part of emotional computing, has gradually become one of the hot spots in the field of pattern recognition. With the development of the research, the recognition effect is not very ideal when just used a single model to classify speech emotional status. In order to improve recognition rate, we propose a method in combination of Hidden Markov Model (HMM) and radial basis function neural network (RBF). This method designs HMM models for different emotional states, then gets the best sequence of emotional speech signal by Viterbi algorithm. Then, the feature parameters of the same state are structured as uniform dimension by the method of spatial orthogonal basis function expansion, which is used as the input of RBF for recognition of speech emotional states. Finally, the final results are given by RBF. The experiment shows that the proposed method has better recognition rate than isolated HMM.

**Key words:** emotion calculation; artificial intelligence; hidden Markov model; neural network; speech emotion recognition

## 0 引言

语音是人类沟通方式中最快和最自然的方法。研究人员认为语音是一种快速和有效的人机交互方法。然而,这要求机器应具有足够的智能来识别人类声音。自五十年代以来,已经对语音识别进行了大量研究,其中涉及了将人类语音转换为词序列的过程。尽管在语音识别方面的研究已经取得了重大进步,但仍然远远

没有实现人与机器之间的自然交互,这是因为机器不能理解说话者的情感状态。因此引入了语音情感识别<sup>[1]</sup>这一相对较新的领域,即定义为从他或她的语音中提取说话者的情感状态。语音情感识别可以从语音中提取有用的语义,并改进语音识别系统的性能<sup>[2]</sup>。

目前,大多数研究者都同意“调色板理论”<sup>[3]</sup>,其中指出任何情感都可以分解成主要情感和次要情感。

收稿日期:2017-12-05

修回日期:2018-04-05

网络出版时间:2018-05-28

基金项目:国家自然科学基金(61572260);江苏省重点研发计划(BE2015702)

作者简介:林巧民(1979-),男,博士,副教授,硕导,研究方向为信息网络,无线传感器网络关键技术、物联网隐私保护安全关键技术等;齐柱柱(1992-),女,硕士研究生,研究方向为云计算与物联网技术。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180525.1610.080.html>

在此将情感分为 5 种:高兴、惊奇、愤怒、悲伤和中性,并对其进行语音情感识别。在语音情感识别中,算法的优劣决定着识别率的高低。尽管目前研究者已取得大量成果,比如文献[4]仅使用隐马尔可夫模型对语音进行情感识别,文献[5]对传统的神经网络方法进行了分析,文献[6]使用深度神经网络和隐马尔可夫模型相混合的模型,相比单独使用统计模型得到了不错的识别率。因此单一使用某算法进行语音情感识别的效果并不理想。依据 HMM 对动态时间序列具有的极强的建模能力和较弱的分类决策能力,以及 ANN 具有的较强的并行处理能力和分类决策能力及不能处理语音动态变化的特征序列等特点,将 HMM 和 ANN 两种算法相结合,取长补短,以提高语音情感识别率。

1 情感特征参数提取

在对语音信号进行特征参数提取之前,首先要对语音信号进行预处理<sup>[7]</sup>,以去除语音信号中掺杂的背景噪音的影响,并且获得计算机能够识别的、较为理想的语音样本数据。语音信号预处理包括反混叠滤波、预加重、分帧加窗和端点检测等操作。

原始语音信号包含各种各样的信息,如语调、文字、情感、韵律等,那么可提取的情感特征参数也是多种多样的。首先要解决的一个关键问题是如何从这些情感特征参数中建立能反映个人情感特征的矢量<sup>[8]</sup>。因此要取得较好的语音情感识别效果,必须准确选取语音情感特征参数。一个重要的选择策略是:尽可能提取更易于提高语音情感识别率的情感特征参数,并减少语音信号中那些无用的冗余信息<sup>[9]</sup>。

1.1 基音频率

基因频率是人说话发浊音时声带振动的基本频率,简称基频,通常用  $F_0$  表示。基频的变化模式称为声调,包含了大量有用的语音情感激活度的信息。在国内外许多有关语音情感识别的研究中,基因频率是重要的参数之一,有助于研究语音情感的变化。

文中选取基频的静态特征  $F_0$  和动态特征作为特征参数,以此来模拟基频的瞬时特征和基频轮廓的变化情况,其中动态特征是利用基频的一阶和二阶差分求得。这样就得到基频的特征参数  $F_0, \frac{dF_0(i)}{di}$ ,  $\frac{d^2F_0(i)}{di^2}$ , 其中  $F_0(i)$  表示第  $i$  帧的基频。

1.2 短时能量

短时能量即音量高低,它是一帧采样点值的加权平方和。短时能量直接反映声音的音量大小,其中清音的能量较小,浊音的能量较高。一个人的情感不同时,其说话的音量也不同。例如在生气或者惊讶时,说

话的音量就较大,其短时能量也就越高。

文中选取短时能量的静态特征  $E(i)$  和动态特征作为特征参数,以此来模拟能量的瞬时特征和能量轮廓的变化,其中动态特征是利用能量的一阶和二阶差分求得。由此得到短时能量参数:  $E(i), \frac{dE(i)}{di}, \frac{d^2E(i)}{di^2}$ , 其中  $E(i)$  表示第  $i$  帧的短时能量。

1.3 振幅

语音信号的振幅特征也是语音情感特征参数的一种,愤怒或惊奇时人们音量变大,语音信号振幅较大,然而当悲伤或者平静时,语音信号具有较小振幅,因此振幅也常被用作语音情感识别中的特征参数。选取发音起始点间的平均振幅的最大值作为最大振幅,同时提取平均振幅和最大振幅做参数。

1.4 LPCC 系数

在语音情感识别中,线性预测倒谱系数(LPCC)常被用作情感特征参数,由线性预测系数(LPC)推导出。LPCC 系数的最大优点就是能较彻底地消除语音产生过程中的激励信息,并且能较好地反映声道响应。LPCC 系数能很好地模拟人的声道模型,十几个 LPCC 系数就能良好地描述语音信号的共振峰特性,同时求取 LPCC 系数时计算量小,易于实现,因此在语音情感识别中能获得良好的识别效果<sup>[10]</sup>。文中选取 10 阶 LPCC 系数作为情感特征参数,表示为  $C_{i1}, C_{i2}, \dots, C_{i12}$ , 其中  $i$  表示帧数,  $k = 1, 2, \dots, 12$ 。

2 语音情感识别模型

HMM 模型<sup>[10]</sup>的最大优势是有极强的建模能力,尤其对动态时间序列,在语音情感识别中已经取得了相当不错的效果,并大大提高了语音情感识别性能。然而,HMM 模型的分类能力弱、模式识别性能差,存在先验假设问题,需要先验统计知识等,先验假设也就是假设语音信号当前的状态只与前一个状态有关<sup>[11]</sup>。

HMM 模型中常用的 Baum-Welch 训练算法是基于最大似然准则,其分类决策能力较弱,而且仅根据累积概率最大值判断,忽略了其他状态的累积概率和每个模型之间的相似特征,降低了 HMM 情感识别能力。

ANN 模型<sup>[12]</sup>正好相反,具有极强的分类决策能力,良好的自适应和自学习能力,较强的鲁棒性和容错性,不需要预先假设,广泛应用于语音情感识别。但 ANN 模型动态特性描述能力较弱,只能解决不涉及时间序列处理的静态模式分类问题。ANN 模型是可以训练的,可不断积累学习经验以便提高性能,同时又因具有高度的并发性而能进行快速分类判别。

因此,将有较强动态时序建模能力的 HMM 和有

较强分类决策能力的 ANN 两种方法进行有机结合<sup>[13]</sup>,充分发挥两者各自的优势,进一步提高语音情感识别的准确率。该方法识别流程如图 1 所示。

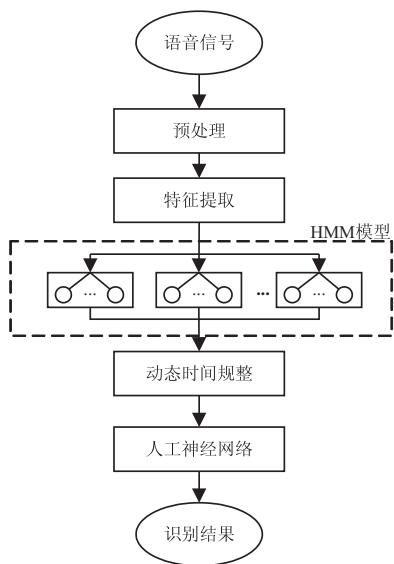


图 1 语音识别系统流程

## 2.1 混合模型原理

针对 HMM 和 ANN 各自的优缺点,将 HMM 模型的较强动态时序建模能力和 ANN 模型的较强分类决策能力相混合形成新的模型,HMM 模型的输出作为 ANN 模型的输入,对语音信号进行识别和分类,完成语音情感识别。

待识别语音样本通过 HMM 模型的 Viterbi 算法解码产生所有状态的累积概率  $x = \{x_1, x_2, \dots, x_L\} = \{\alpha_T^1(1), \dots, \alpha_T^1(N), \dots, \alpha_T^K(1), \dots, \alpha_T^K(N)\}$  作为神经网络分类器的输入特征,其中  $k$  ( $k = \{1, 2, \dots, K\}$ ) 为所要识别的语音情感类别数。神经网络模型由输入层、隐藏层和输出层组成,输入层包括  $L$  个神经单元,与 HMM 模型中各类情感的状态累积概率相对应;隐藏层是高斯核函数,可以对输入进行部分响应,将输入空间分成若干小范围,以实现分类和函数近似;输出层包含  $K$  个神经元,每个神经元对应一种要识别的情感。

将 HMM 模型与 ANN 模型融合在一起,这里选择的 ANN 模型是径向基函数神经网络(RBF),可以充分利用全部情感状态的累积概率,并对信号细节分量加以提取。RBF 神经网络是由输入层、输出层和隐含层组成的网络结构,其中输入层节点是线性神经元,输出层节点是线性求和单元,隐含层节点常采用高斯核函数,可以对输入产生局部响应,将输入空间划分为若干小的局部区间,以达到分类和函数逼近的目的。RBF 网络结构简单,参数训练易于实现,且不易陷入局部极小的麻烦。

综合两类方法各自的优点,研究 HMM 和 RBF 相结合的问题。

## 2.2 训练和识别

HMM/RBF 混合模型分为训练和识别两个阶段<sup>[14]</sup>。训练阶段分为 HMM 的训练和 RBF 的训练。在 HMM 训练中,首先对语音样本进行归一化处理,然后利用训练样本建立不同情感的 HMM 模型参数  $(A, B, \pi)$ 。HMM 模型参数训练采用 Baum-Welch 训练算法,本质上是求使似然概率  $P(O/\lambda)$  最大的一个局部最优解。 $\lambda$  是模型参数,  $\lambda = \{A, B, \pi\}$ ,  $O$  是指一个给定的观察值序列  $O = \{o_1, o_2, \dots, o_T\}$ ,  $T$  为观测序列长度。它是用已标记情感类别的情感语音数据对 HMM 模型进行训练,使其似然概率  $P(O/\lambda)$  趋于局部最大,重复这个过程,逐步改进模型参数,直到  $P(O/\hat{\lambda})$  收敛,其中  $\hat{\lambda}$  表示 HMM 模型训练过程中得到的新模型。在训练中并不是训练越多越好,训练过多反而会导致模型参数精度变差,因此选择判定模型收敛的方法是前后两次的输出概率的差值小于一定阈值或模型参数几乎不变为止。

RBF 模型的训练采用 BP 算法。经过 HMM 模型的 Viterbi 算法解码输出全部情感状态的累积概率,然后利用 RBF 模型进行非线性映射。HMM/RBF 混合模型的训练算法如下:

(1) 用 Baum-Welch 算法训练 HMM 模型,为每个情感状态分别建立一个 HMM 模型,获得训练好的 HMM 参数库。

(2) 输入待识别语音样本  $x_i$  ( $1 \leq i \leq M$ ),  $i$  是语音在语音库中的序号,  $M$  为其容量。用 HMM 模型对语音信号数据进行时间序列处理,采用 Viterbi 算法解码得到相应 HMM 参数输出的状态累积概率  $V = [\beta_{T(1)}, \dots, \beta_{T(j)}, \dots, \beta_{T(N)}]$  ( $1 \leq j \leq N$ ), 这表示状态  $s_j$  的累积概率。

(3) 用公式  $u_i = \frac{v_i - v_{\min}}{v_{\max} - v_{\min}}$  对  $V$  进行归一化,作为 RBF 神经网络的输入。因要识别五种情感,所以输出层有五个,对应矢量  $\mathbf{R}, \mathbf{R} = [r_1, r_2, \dots, r_i, \dots, r_5]$ , 其中只有  $r_i = 1$ , 代表识别出的情感,其他为 0。

(4) RBF 采用 BP 学习算法对 RBF 进行训练,直到满足网络的收敛精度要求为止。神经网络训练算法使用 BP 学习算法,并且代价函数为修正的互熵函数。假设输出层有  $N$  个节点,每个节点的输出为  $Y_n$ , 对应的期望输出为  $T_n$ , 修正的互熵函数可以表达为:

$$E = - \sum_{n=1}^N [T_n \log(Y_n) + (1 - T_n) \log(1 - Y_n)] \quad (1)$$

当期望输出为 1 时,互熵函数中的第二项为 0,可以加快网络的训练速度。

如图 2 所示,系统识别过程为:首先待识别样本经



过预处理和特征提取操作后,经过 HMM 模型的 Viterbi 算法<sup>[15]</sup>解码产生全部状态的累积概率保存在矢量  $V$  中,且不用 HMM 模型识别;然后对所得矢量进行时

间规整,可使用空间正交基函数展开的方法<sup>[16]</sup>,最终生成等维的特征矢量,将其作为 RBF 模型的输入进行非线性映射,获取识别结果。

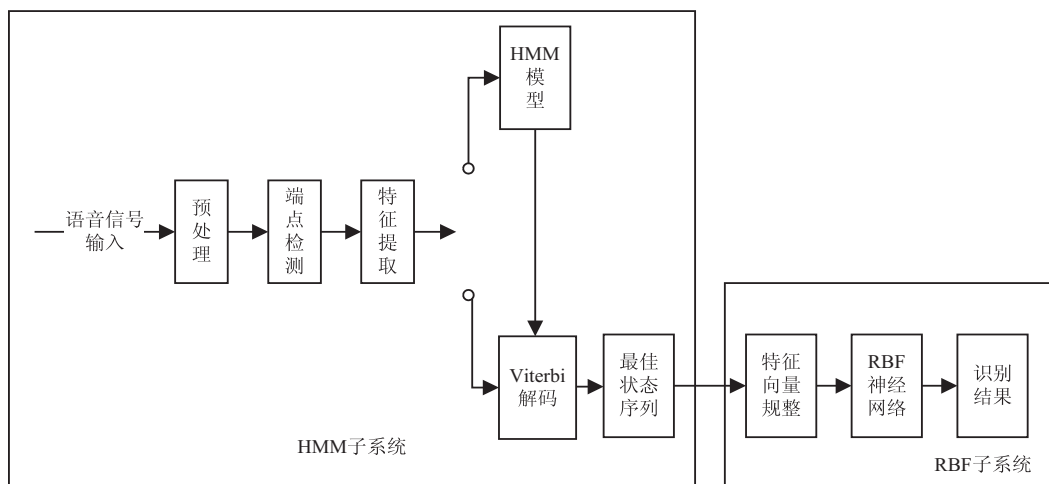


图 2 HMM/RBF 混合模型

### 2.3 Viterbi 算法

Viterbi 算法用于生成最佳状态序列,评估给定观察符号序列与给定 HMM 模型之间的最佳匹配的可能性,然后实现最优状态序列。指  $P(S, O/\lambda)$  最大时确定的状态序列,即 HMM 输出一个观察值序列  $O = o_1, o_2, \dots, o_T$  时,使输出概率最大的状态序列  $S = s_1 s_2 \dots s_T$  就是最佳。算法描述如下:

(1) 初始化:  $\alpha'_0(1) = 1, \alpha'_0(j) = 0, (j \neq 1)$ ;

(2) 递推公式:  $\alpha'_t(j) = \max_i \alpha'_{t-1}(i) a_{ij} b_{ij}(o_t) (t = 1, 2, \dots, T; i, j = 1, 2, \dots, N)$ ;

(3) 最后结果:  $P_{\max}(S, \frac{O}{\lambda}) = \alpha'_T(N)$ 。

每一次使  $\alpha'_t(j)$  最大的状态  $i$  所组成的状态序列,就是解码产生的最佳状态序列。

在使用 Viterbi 算法求取最佳状态序列时,由于使用递归计算的方法,概率值的连续乘法运算很容易导致下溢现象。为了解决该问题,通常使用两种方法:第一种是增加比例因子,用于求和运算;第二种是对概率值取对数后再进行计算,用于乘积运算。

### 2.4 状态归一化

由于人工神经网络的输入必须是等维数据<sup>[17]</sup>,首先要对所得的最佳状态序列进行归一化处理。为了获得等维的特征向量,采用正交多项式展开的方法对状态进行归一化。HMM 模型对应的 Markov 链由 5 个状态组成,表示为  $i = 1, 2, 3, 4, 5$ 。假设  $m$  是状态  $i$  中的特征向量数量,则向量集表示为:  $\{\vec{x}_1^i, \vec{x}_2^i, \dots, \vec{x}_j^i, \dots, \vec{x}_m^i\}$ 。

其中,  $\vec{x}_j^i = [x_{j1}^i, x_{j2}^i, \dots, x_{jL}^i]$ ,  $L$  表示特征向量的长度,把每个特征向量按行排列可得到如下矩阵:

$$C = \begin{bmatrix} x_{11}^i & x_{12}^i & \dots & x_{1L-1}^i & x_{1L}^i \\ x_{21}^i & x_{22}^i & \dots & x_{2L-1}^i & x_{2L}^i \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1}^i & x_{m2}^i & \dots & x_{mL-1}^i & x_{mL}^i \end{bmatrix} \quad (2)$$

矩阵  $C$  中每一列可看作为  $m$  次多项式系数,公式如下:

$$f(x) = x_{1n}^i + x_{2n}^i + \dots + x_{mn}^i, n = 1, 2, \dots, L \quad (3)$$

该多项式在 0 到 1 空间用正交基函数展开:

$$C_n = \frac{2n+1}{2} \int_{-1}^1 f(x) P_n(x) dx \quad (4)$$

其中,  $P_n(x)$  为勒让德多项式;  $C_n$  为展开系数。

为了简化计算,仅选择 6 个勒让德多项式作为正交基。尽管  $m$  是变量,但是每个  $m$  阶多项式可被扩展为 6 个系数。因此对于状态  $i$  而言,  $L$  列的向量通过勒让德多项式展开的系数共有  $6L$  个,  $L$  是常量。

## 3 实验结果

系统中采用的语音样本来自 CASIA 汉语情感语料库<sup>[17]</sup>,由四个专业发音人对相同的文本赋予不同的情感来阅读。挑选出愤怒(angry)、高兴(happy)、中性(neutral)、悲伤(sad)、惊奇(surprise)五种情感共 300 句语音状态作为实验对象。采用多次十折交叉验证的方法,将语音样本分为十份,轮流将其中 9 份做训练 1 份做测试,10 次结果的均值作为对算法精度的估计。实验中语音信号的采样频率为 16 kHz,量化精度为 16 bit,信噪比约为 35 dB,帧移为 5 ms,帧长为 16 ms。

表 1 和表 2 分别显示了单独使用 HMM 模型以及 HMM/RBF 混合模型在 5 类不同情感状态下的情感识别率。从表 1 看出,悲伤的识别率最高为 82.2%,其

平均识别率达到 77.86%,由此可知,采用单独 HMM 模型的识别效果一般。从表 2 看出,高兴和愤怒的识别率有明显提高,其平均识别率达到了 89.5%。

表 1 基于 HMM 的语音情感识别结果

情感	HMM 模型的识别率(平均 77.82)/%				
	高兴	愤怒	悲伤	惊奇	中性
高兴	79.1	9.6	0	7.2	4.1
愤怒	10.2	82.3	0	7.5	0
悲伤	0	0	74.1	0	25.9
惊奇	5.2	16.5	0	78.3	0
中性	14.3	0	8.7	1.7	75.3

表 2 基于 HMM 和 ANN 混合模型的语音情感识别结果

情感	HMM/RBF 混合模型的识别率(平均 89.5)/%				
	高兴	愤怒	悲伤	惊奇	中性
高兴	91.6	5.4	0	3	0
愤怒	5.6	92.3	0	2.1	0
悲伤	0	0	85.6	0	14.4
惊奇	3.5	7.2	0	89.3	0
中性	6.1	0	5.2	0	88.7

由图 3 可明显看出,混合模型在愤怒、高兴、悲伤、惊奇和中性 5 种不同情感识别效果上较单独的 HMM 模型有较为明显的提高。

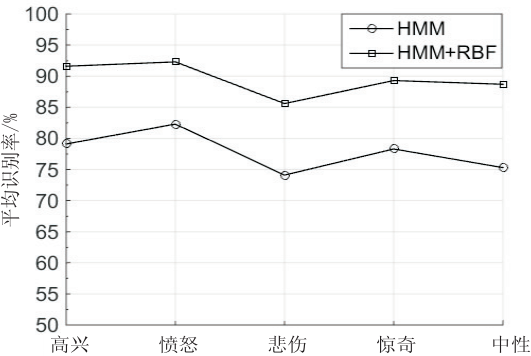


图 3 单独的 HMM 以及 HMM/ANN 混合模型的语音情感识别模型比较

4 结束语

目前,多种网络模型相结合是解决语音识别中的问题的有效途径和思路,因此提出了基于 HMM/RBF 的语音情感识别模型,并介绍了该模型在语音情感识别中的使用方法。实验结果表明,该模型比单一的模式识别在语音情感识别中有更好的识别效果。同时也有许多可以改进的地方,如在特征参数选择提取上,HMM 模型训练算法等方面,有待进一步的深入研究。

参考文献:

[1] 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.

[2] AYADI M E,KAMEL M S,KARRAY F. Survey on speech emotion recognition:features,classification schemes, and databases[J]. Pattern Recognition,2011,44(3):572-587.

[3] COWIE R, CORNELIUS R R. Describing the emotional states that are expressed in speech[J]. Speech Communication,2003,40(1-2):5-32.

[4] GUPTA A,TECH M,SHARMA N. Speech recognition using hidden Markov model[J]. Journal of Computing Technologies,2013,2(3):1-6.

[5] 蔡伟建. 人工神经网络理论在语音识别技术中的应用[C]//第八届全国信息获取与处理学术会议论文集. 出版地不详:中国仪器仪表学会,2010.

[6] LI Longfei, ZHAO Yong, JIANG Dongmei, et al. Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition[C]//Humaine association conference on affective computing and intelligent interaction. Geneva,Switzerland:IEEE,2013:312-317.

[7] 刘琦,尹国祥. 基于 Matlab 的语音信号预处理技术研究[J]. 电子技术与软件工程,2014(1):62-63.

[8] 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.

[9] 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.

[10] ANILA R,REVATHY A. Emotion recognition using continuous density HMM[C]//International conference on communications and signal processing. Melmaruvathur,India:IEEE,2015.

[11] 石锐,郑晓平,何庆华. 基于 HMM-ANN 的咳嗽音识别[J]. 世界科技研究与发展,2012,34(5):751-753.

[12] FU Liqin,MAO Xia,CHEN Lijiang. Relative speech emotion recognition based artificial neural network[C]//IEEE Pacific-Asia workshop on computational intelligence and industrial application. Wuhan,China:IEEE,2008:140-144.

[13] 曹鹏霞. 基于 HMM 和人工神经网络混合模型的汉语语音情感识别[D]. 长沙:湖南师范大学,2015.

[14] GUO Xianhai. Study of emotion recognition based on electrocardiogram and RBF neural network[J]. Procedia Engineering,2011,15:2408-2412.

[15] MAO Xia,CHEN Lijiang,FU Liqin. Multi-level speech emotion recognition based on HMM and ANN[C]//WRI world congress on computer science and information engineering. Los Angeles,CA,USA:IEEE,2009:225-229.

[16] MAO Xia,ZHANG Bing,LUO Yi. Speech emotion recognition based on a hybrid of HMM/ANN[C]//Proceedings of 7th WSEAS international conference on applied informatics and communications. Vouliagmeni, Athens, Greece: World Scientific and Engineering Academy and Society,2007:367-370.

[17] 张昕然. 跨库语音情感识别若干关键技术研究[D]. 南京:东南大学,2016.