基于集成学习的语音情感识别算法研究

李田港[1],叶 硕[1],叶光明[2],褚 钰[1]
(1.武汉邮电科学研究院,湖北 武汉 430000;
2.武汉烽火众智数字技术有限责任公司,湖北 武汉 430000)

摘 要:语音情感识别是语音识别的热门方向,心理学将情感识别分为离散型和连续型,离散型情感识别常用的声学特征为韵律学特征、基于谱的相关特征、音质特征,识别方法通常有KNN、SVM、HMM等。提出一种基于距离加权的改进KNN算法,引入类平均距离作为加权依据,并设计一种基于集成学习的加权投票算法,将改进KNN、SVM、BPNN分类方法进行集成,提高语音情感识别率。实验表明,改进后的KNN算法相比传统KNN,识别率在不同语种的语料库上均有提升,最大提升为9.6%,且表现结果稳定,准确率与SVM、BPNN大致相当,可用于集成学习;对比单一识别算法,所设计的集成学习算法具有较高可靠性,在生气、高兴、悲伤、惊慌及中性情感上均达到较好的识别效果,实现了离散型语音情感的识别。

关键词:语音识别;情感识别;SVM;W-KNN;BPNN;集成学习
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2020)06-0082-05
doi:10.3969/j.issn.1673-629X.2020.06.016

Research on Speech Emotion Recognition Algorithm Based on Ensemble Learning

LI Tian-gang[1], YE Shuo[1], YE Guang-ming[2], CHU Yu[1]
(1. Wuhan Research Institute of Posts and Telecommunications, Wuhan 430000, China;
2. Wuhan Fiberhome Wisdom Digital Technology Co., Ltd., Wuhan 430000, China)

Abstract: Speech emotion recognition is a popular direction of speech recognition. Psychology divides emotional recognition into discrete and continuous. For discrete emotion recognition, the acoustic features commonly used are prosodic features, spectral related features, and acoustic quality features. The recognition methods usually include KNN, SVM, HMM, etc. We propose an improved KNN algorithm based on distance weighting, introduce the class average distance as the weighting basis, and design a weighted voting algorithm based on ensemble learning, which integrates the improved KNN, SVM and BPNN classification methods to improve the speech emotion recognition rate. Experiments show that compared with traditional KNN, the recognition rate of the improved KNN algorithm is improved on corpus of different languages, with the maximum rate of 9.6% and stable performance result. The accuracy rate is roughly the same as SVM and BPNN, which can be used for ensemble learning. Compared with the single recognition algorithm, the designed ensemble learning algorithm has higher reliability, achieves better recognition effect in anger, happiness, sadness, panic and neutral emotion, and realizes the recognition of discrete speech emotion.

Key words: speech recognition; emotion recognition; SVM; W-KNN; BPNN; ensemble learning

0 引 言

语言是人类最常用的信息交换方式,一段语音中通常包含说话人的三部分内容:声音特征信息、语言内容信息、语音情感信息。同样的语言,使用不同的情感表达,效果往往不同。语音情感识别(speech emotion recognition,SER)即是对语音中包含的情感信息进行研究。

如何提取语音中具有判别特性的情感特征,一直是该识别任务的难点。心理学将情感分为了离散型和连续型[1],离散型语音情感是指语音中只包含一种情绪或只有一种突出情绪,针对离散语音,常用特征可以分为低级描述(common low-level descriptors,LLDs)和高级描述的水平统计函数(high-level statistical functions,HSFs)[2-3]。

低级描述包括:基音频率(fundamental frequency)、能量(energy)、过零率(zero-crossing)、抖

收稿日期:2019-07-18 **修回日期**:2019-11-20
基金项目:2018年度湖北省技术创新专项重大项目(2018AAA063)
作者简介:李田港(1995-),男,硕士研究生,研究方向为机器学习、语音识别。

动(jitter)、梅尔滤波特征(Mel-filterbank features)、共振峰位置/带宽(formant locations/bandwidths)、谐波噪声比(harmonics-to-noise ratio)。

高级描述包括:均值(mean)、方差(variance)、最小值(min)、最大值(max)、范围(range)、高阶矩(higher order moments)、线性回归系数(linear regression coefficients)等。

常用的分类方法包括:支持向量机(support vector machine, SVM)、K 最近邻(K-nearest neighbor, KNN)、反向传播神经网络(back propagation neural networks, BPNN)、隐马尔可夫模型(hidden Markov model, HMM)等。近年来神经网络发展迅猛,相关技术也被应用于语音情感识别的研究。

文中着重研究离散型的情感语音识别,提取语音多个与情感分类相关的特征,使用多种算法对语音情感进行识别,同时提出一种基于距离加权的 KNN 改进算法和基于集成学习的加权投票算法,用于更好地实现语音情感识别。

1　情感特征提取

目前与语音情感有关的声学特征大致可以分为三类:韵律学特征、基于谱的相关特征、音质特征。

韵律学特征注重音高、快慢以及轻重等方面的变化[4]。韵律能够帮助人们理解说话人语音信息中的重点,使话语自然流畅,人们能够更快地获取有效信息,包含着丰富的情绪特征。文中提取短时能量、基音频率、过零率、浊音帧差分基音等作为语音情感识别的主要特征。

基于谱的相关特征一般为 LPCC(linear predictor cepstral coefficient)与 MFCC(Mel-frequency cepstral coefficient)。谱特征被认为是声道形状变化和发声运动之间相关性的体现[5]。研究者发现语音中的情感内容对频谱能量在各个频谱区间的分布有着明显的影响[6]。由于人听到的声音高低和频率大小不成线性正比关系,MFCC 特征基于人耳听觉特性,在语音情感分类中具有良好的鲁棒性和正确度,因此文中提取的谱相关特征为 MFCC 及其一阶差分。

声音质量是人们赋予语音的一种主观评价指标,用于衡量语音是否纯净、清晰、容易辨识等[7]。用于衡量声音质量的声学特征一般有:共振峰频率及其带宽、频率微扰和振幅微扰、声门参数等,其中共振峰频率的分布特性决定语音的音色[8]。文中提取的声音质量特征为第一、二、三共振峰频率特征。

对语音信号进行预处理、分帧、加窗,以帧为最小单位对语音信号特征进行提取,并计算特征的相关统计量。具体特征及其统计量如表 1 所示。

表 1　语音情感特征提取

语音特征	基于特征的相关统计量				
	最大值	最小值	均值	方差	其他
短时能量	√	√	√	√	一阶抖动、线性回归系数及均方差
基音频率	√	√	√	√	一、二阶抖动
浊音帧差分基音	√	√	√	√	动态范围
MFCC	√	√	√	√	
共振峰频率	√	√	√	√	一阶抖动
一,二共振峰比	√	√	√		

2　情感识别算法

2.1　基于类平均距离的改进加权 KNN 算法

KNN 算法的思想是对于一个待分类样本,找出其在特征空间中最邻近的 K 个样本,将这 K 个样本中出现个数最多的类标签作为该待分类样本的分类标签,因此分类结果只与近邻的几个样本有关。

这种惰性学习方式依赖训练样本,每个训练样本都要保存,且对于每个待测样本都要计算一遍其与所有训练样本的距离,空间、计算开销较大,而各个近邻又对分类结果影响相同,这与实际的分类情况有所不同;此外,对每个待测样本分类,只用到了近邻样本的信息,不能充分利用样本的分布信息。

基于距离的加权 KNN 是对传统 KNN 算法的一种改进,其思想是对每个近邻样本,根据其与待测样本距离的远近,赋予不同的权重,距离越近权重越大。

文中提出的算法是对加权 KNN 做进一步改进,引入类平均距离[9]作为加权依据。类平均距离可以反映出同一类相邻样本之间的距离更接近,不同类的类平均距离有较大差异。改进算法的实现如下:

算法 1:类平均距离计算。

输入:语音样本训练集 D,总样本数 n,分类数 m,分类样本数 n_h,$h = 1,2,\cdots,m$;x_{li} 为第 l 类语音第 i 个样本的特征向量,$l_1,l_2,\cdots,l_m$ 为分类样本的标签,$i = 1,2,\cdots,n_h$;语音样本特征维数 K,x_{li}^j 表示第 i 个样本的第 j 维特征,$j = 1,2,\cdots,K$;

输出:m 维数组 array[ ]:存储类标签;m 维数组 array_ad[ ]:存储类平均距离。

1:for $h = 1,2,\cdots,m$ do

2:初始化 n 维数组 t[ ],t_ad[ ]和变量 counter 为 0;$l = l_h$;

3:for $i = 1,2,\cdots,n_h$ do

4:for $j = 1,2,\cdots,K$ do

5:将样本与同类其他样本的第 j 维特征值进行比较;找出在第 j 维上特征差值最小的样本;

6:若 t[ ]中无该样本,则存入 t[ ];

7:end for

8:for all t[ ] do

9:计算 t[ ]每个样本与 x_{li} 的欧氏距离;

10:end for

11:去掉最大值;求距离均值并存入 t_ad[ ]

12:counter 加 1;

13:end for

14:计算数组 t_ad[ ]前 q 个元素的均值,存入 array_d[h];将类标签 l 存入数组 array[h]

15:end for

在算法 1 中,首先计算一个类中每个样本与其同类最近邻样本的平均距离,再对该类所有样本与其同类最近邻样本的平均距离求平均值,以此平均值作为该类的类平均距离;然后求出待分类样本的 K 近邻样本,计算每个近邻样本类平均距离和该近邻样本与待分类样本之间距离的差值之比;将差值之比最大的 t 个近邻样本删除,认为这 t 个近邻样本与待测样本的距离和分别对应的类平均距离偏差大,与待分类样本属于同一类的可能性小, t 一般设置为 1。

算法 2:基于类平均距离的改进加权 KNN 分类器。

输入:语音样本训练集 D ,总样本数 n ,分类数 m ,分类样本数 n_h , $h = 1,2,\cdots,m$;最近邻参数 K 和整数阈值 t ;待分类样本以及数组 array[ ]和 array_d[ ];

输出:待分类样本的分类标签。

1:for all D do

2:计算每个训练样本与待分类样本的欧氏距离;找出 K 个距离最小的样本,存入 K 维数组 N [ ];

3:将对应的欧氏距离存入 K 维数组 d [ ];将对应的类标签存入 K 维数组 label[ ];

4:end for

5:for $i = 1,2,\cdots,K$ do

6:计算 $N[i]$ 与待分类样本距离 $d[i]$ 和类平均距离 array_ad[label[i]]的差值之比的绝对值:ratio[i] = |(d[i]-array_ad[label[i]])/array_ad[label[i]]|;

7:end for

8:将 ratio[ ]中最大的 t 个元素置 0;

9:for $i = 1,2,\cdots,K$

10:计算 weight[label[i]] = 1+ b ×ratio[i];

11:end for

12:将数组 weight 中最大元素对应的分类标签作为待分类样本的分类

算法 2 对余下的每个近邻样本,根据其差值之比计算权重,最后对待分类样本根据余下的近邻样本和权重进行加权多数投票,确定待分类样本的类别。

2.2 集成学习算法

集成学习是一种将多个学习结果进行择优整合[10],从而获得比单个学习器更好的学习效果的机器学习方法。为提升语音情感分类正确度,文中还考虑支持向量机(SVM)、反向传播神经网络(BPNN)、隐马尔可夫模型(HMM)在语音情感识别任务中的表现。

SVM 是建立在统计学习理论和结构风险最小化准则基础上的一种算法。将原始的数据样本从线性不可分的低维空间映射到一个高维的特征空间,并寻找一个满足分类要求的最优分类超平面,以此完成分类任务[11]。该方法在保证分类精度的同时能够使超平面两侧的空白区域最大化,对解决小样本分类任务具有较好的表现。

BPNN 是最常用的神经网络学习方式,误差的逆传播可有效计算权重参数的梯度。文中搭建一个三层神经网络,分为输入层、隐藏层、输出层。输入为语音信号多维特征向量组成的训练样本,隐藏层神经元数为 10 个,隐藏层和输出层使用对数 Sigmoid 激励函数,用于解决高维非线性可分问题,每个输出神经元分别输出 5 类情感的预测概率。使用均方误差作为损失函数。设置训练次数为 150,网络性能目标为 10^{-6}。

HMM 是一种用参数表示的、基于语音信号的时间序列结构建立的、用于描述其随机过程统计特性的概率模型[12-13]。语音的情感在短时间内具有一致性,将语音部分分割成极小的时间片段,那么该片段的特性近似稳定,总过程可视为从某一相对稳定特性到另一相对稳定特性的转移。因此对不同情感构建 HMM,可以完成对语音情感的分类。文中拟采用从上述的五种分类算法中选出三种分类效果较好的算法,对其分类结果以加权投票的方式进行集成学习。

3 实验测试

3.1 数据集

本实验采用三种不同语种公开数据集:德国柏林德语语料库[14]、中国中科院汉语语料库以及 EmoV-DB 英语情感语料库。德国柏林德语语料库中包含 7 种情感,共 535 句情感语音信号,音频采样频率为 16 kHz,16 bit 量化。从中选择了 anger、happiness、neutral、sadness 和 fear 共 5 种情感,每种情感随机选取 50 条语音组建德语数据集。

汉语情感语料库,采样频率为 16 kHz,16 bit 量化,共 4 个发音人,每人的语音有 6 种情感。从中选择 anger、happiness、neutral、sadness 和 fear 共 5 种情感,每种情感随机选取 50 条语音组建汉语数据集。

Emov-DB,其采样频率为 16 kHz,16 bit 量化,共 4 个发音人,每个人的语音有 5 种情感 neutral,sleepiness,anger,disgust 和 amused。每种情感随机选择 50 条语音组建英语数据集。

文中所述 5 种算法分别在 3 个语音样本训练集进行折数为 5 折的交叉验证:随机将每个训练集均分为 5 份大小一致的互斥子集,每个子集都保持情感种类分布的一致性,每次实验从中选择 4 个子集的并集作为训练集,余下的 1 个子集作为验证集。为了保证数

据分布的均衡性,对数据集的样本按不同类别分别进行均分再合并。最终可以获得 5 组训练/验证集,进行 5 次训练,得到 5 个验证结果的均值。

3.2　实验结果

实验首先对比不同半径下,改进的 W-KNN 算法和传统 KNN 算法各自的表现,对三个数据集采用 5 折交叉验证求均值,在 K 取值为 1-20 的情况下,5 种情感分类结果如图 1 所示,从左至右依次为德语、汉语、英语数据集,实折线为文中改进后算法的识别率。可以看出,改进后的 KNN 算法优于传统 KNN 算法,在三个数据集上最大提升幅度分别为 5.6%、9.6%、5.2%。

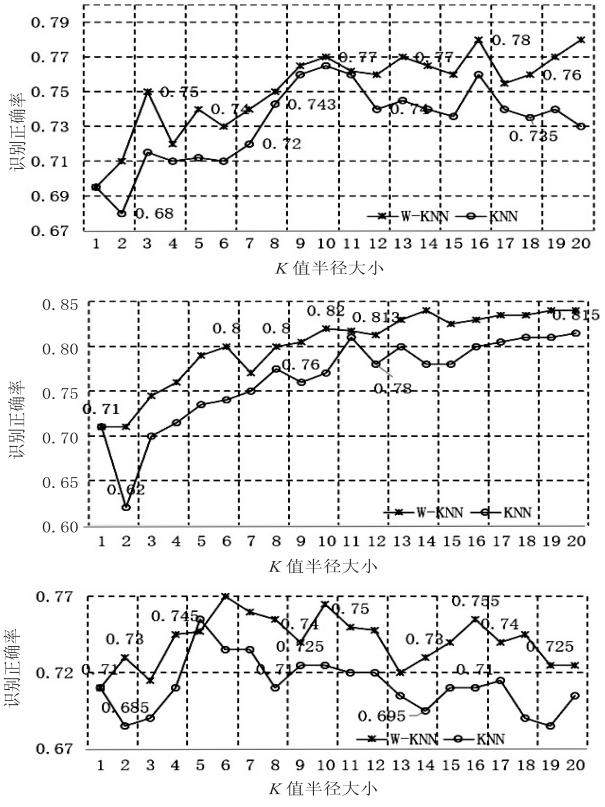


图 1　不同 K 值下的 KNN 算法与 W-KNN 算法的识别正确率

图 2 是 KNN 算法和改进 KNN 算法在综合考虑计算成本和计算速度后,分别在三个数据集上对五种情感的识别率。从左至右依次为德语、汉语、英语数据集,星折线为文中改进后算法的识别率,情绪依次为生气、高兴、中性、悲伤、惊恐。可以看出,德语数据集上,改进算法对生气、高兴、中性和悲伤情感的识别效果优于传统算法;汉语数据集上,改进算法对生气、高兴、中性和惊恐情感的识别效果优于传统算法;英语数据集上,改进算法对高兴、中性和悲伤情感的识别效果优于传统算法。不同情感的识别正确率有一定差距,分析原因,认为语音情感识别的正确率与语种有一定关系,不同语种对不同情感的表现力不相同。

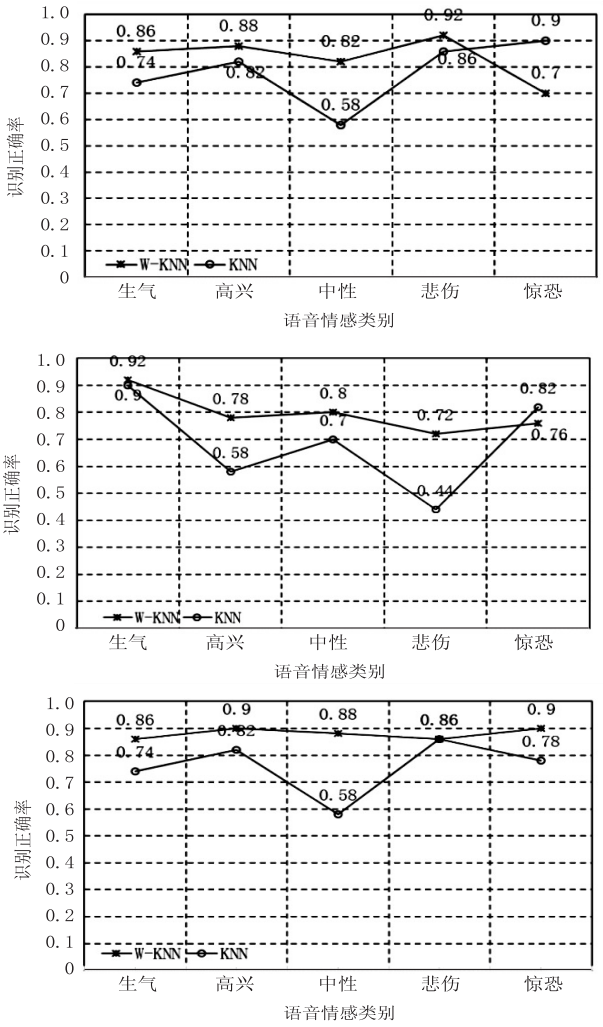
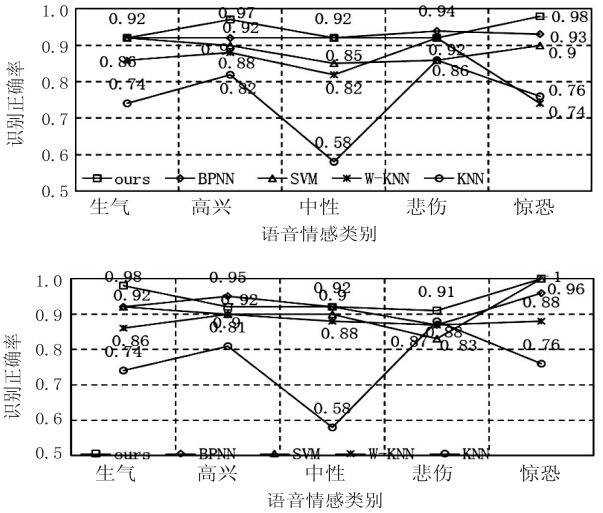


图 2　KNN 与 W-KNN 对不同情感的识别正确率

进一步,文中使用 SVM,HMM 和 BP 神经网络在三个数据集上进行验证,实验发现,基于 HMM 的分类方法,对 HMM 的状态数 N 的选择具有较大依赖,但 N 的增大,将增加时间开销,还会造成系统冗余,为此,只选取 W-KNN、SVM 和 BP 神经网络结果进行集成学习,各个算法结果如图 3 所示,从上到下依次为德语、汉语、英语数据集上的实验结果。



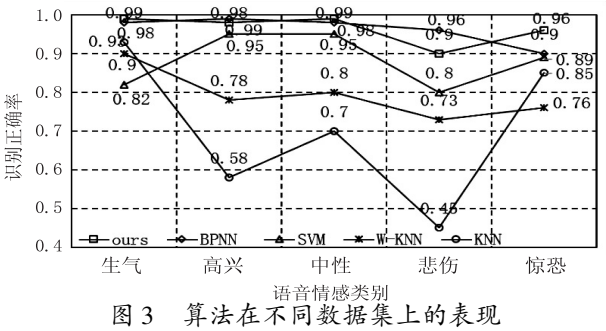


图 3 算法在不同数据集上的表现

结果表明,改进后的 W-KNN 的效果基本与其余集成所用分类方法持平,不存在明显短板,在集成中可以有效贡献准确率。在德语语料库中,文中算法对高兴和惊恐情感的识别均优于其余算法,在生气和中性情感上与 BP 神经网络算法持平;汉语语料库中,文中算法对生气、悲伤和惊恐情感的识别均优于其余算法;英语语料库中,文中算法对生气、中性和厌恶情感的识别均优于其余算法。图 4 是上述 5 种算法对语音情感整体识别正确率的折线图,从图中可以看到,所提出的基于集成学习的加权投票算法具有最好的识别效果。德语和汉语语料库中,文中算法均优于其余算法;英语语料库中,文中算法与 BP 神经网络算法识别效果持平。这一点与所选基分类器的分类表现差异较大有关,BP 神经网络在英语语料库上的表现很大程度上影响了其他算法在集成学习中所占权重。

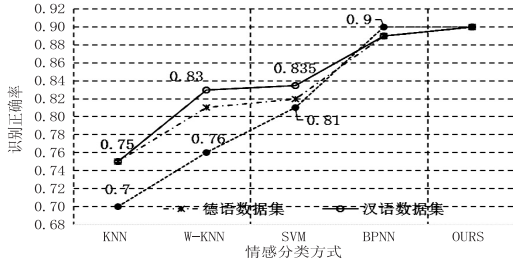


图 4 五种算法情感识别率

4 结束语

语音情感识别作为语音识别的一环,具有重要的研究价值。通过研究发现,在离散语音情感识别中,语音信号的不同特征对最后的分类结果有不同程度的影响,如何寻找更有效的情绪特征,是当前研究者面临的一大问题。

此外,由于人类情感具有模糊的时间边界,情感识别是一项具有挑战性的任务。每个人的情感表达方式不同,一个话语可能包含不止一种情感[15],维度型情感语音的识别依然存在许多难点需要攻克,针对此类问题可以使用长短时记忆网络做进一步的研究。

参考文献:

[1] 张雪英,张 婷,孙 颖,等. 情感语音数据库优化及 PAD 情感模型量化标注[J]. 太原理工大学学报,2017,48(3):469-474.

[2] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans: IEEE, 2017: 2227-2231.

[3] HSIAOP W, CHEN C P. Effective attention mechanism in dynamic models for speech emotion recognition[C]//International conference on acoustics, speech and signal processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 2526-2530.

[4] 刘振焘,徐建平,吴 敏,等. 语音情感特征提取及其降维方法综述[J]. 计算机学报,2018,41(12):2833-2851.

[5] 宋 静,张雪英,孙 颖,等. 基于 PAD 情绪模型的情感语音识别[J]. 微电子学与计算机,2016,33(9):128-131.

[6] 陶华伟. 基于谱图特征的语音情感识别若干问题的研究[D]. 南京:东南大学,2017.

[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.

[8] 高 慧,苏广川,陈善广. 不同情绪状态下汉语语音的声学特征分析[J]. 航天医学与医学工程,2005,18(5):350-354.

[9] 严晓明. 基于类别平均距离的加权 KNN 分类算法[J]. 计算机系统应用,2014,23(2):128-132.

[10] SHIHP Y, CHEN C P, WU C H. Speech emotion recognition with ensemble learning methods[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 2756-2760.

[11] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):1-10.

[12] SHAHIN I. Emotion recognition based on third-order circular suprasegmental hidden Markov model[C]//2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). Jordan: IEEE, 2019: 800-805.

[13] BUYUK O, ARSLAN L M. HMM-based text-dependent speaker recognition with handset-channel recognition[C]//Signal processing and communications applications conference. Diyarbakir, Turkey: IEEE, 2010: 383-386.

[14] XIAO Z, WU D, ZHANG X, et al. A cross-corpus recognition of emotional speech[C]//International symposium on computational intelligence and design (ISCID). Hangzhou: IEEE, 2016: 42-46.

[15] TZIRAKIS P, ZHANG J, SCHULLER B W. End-to-end speech emotion recognition using deep neural networks[C]//IEEE international conference on acoustics, speech and signal processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5089-5093.