A Method of Speech Emotion Recognition Based on Complex Frame Segment Feature*

ZHANG Xia[1*], *YANG Yong*[2], *ZHAO Li*[2]

(1.*School of Mechanical, Electrical and Information Engineering, Putian University, Putian Fujian* 351100, *China*;
2.*School of Information Science and Engineering, Southeast University, Nanjing Jiangsu* 210096, *China*)

Abstract: A method of speech emotion recognition is proposed based on complex frame segment feature. Through combining several successive frames as a segmental unit witch is treated as an input vector for Gaussian Mixture Model(GMM). The inter-frame correlation information is effectively introduced into the process of speech emotion recognition. Furthermore, principal components analysis neural nerwork(PCANN) is adopted before GMM for the purpose of frame parameter compression, to improve the performance of output probability density function. Corresponding experiments are performed and the results show that the recognition rate of the proposed method is improved to some extend comparing with the traditional status output independent GMM, thus the effectiveness of introducing dynamic inter-frame correlation information into the process of speech emotion recognition is validated.

Key words: speech emotion recognition; Gaussian mixture model; principal components analysis neural network; complex frame segment feature

EEACC: 6130E **doi**: 10.3969/j.issn.1005-9490.2022.02.038

基于复数帧段特征的语音情感识别方法*

张 霞[1*], 杨 勇[2], 赵 力[2]

(1.莆田学院机电与信息工程学院, 福建 莆田 351100; 2.东南大学信息科学与工程学院, 江苏 南京 210096)

摘 要: 提出了一种基于复数帧段特征的语音情感识别方法, 采用相继的复数帧组成的特征参数矢量作为语音情感识别 GMM 的输入, 能有效地在语音情感识别 GMM 中引入帧间相关动态信息, 同时为了改善复数帧段输入 GMM 的输出概率密度函数性能, 在 GMM 的前端增加语音帧段参数压缩的主分量分析神经网络(PCANN)。语音情感识别实验证实了引入帧间相关动态信息方法的有效性, 新方法在识别率上较状态输出独立 GMM 方法有一定程度的提升。

关键词: 语音情感识别; 高斯混合模型; 主分量分析神经网络; 复数帧段特征

中图分类号: TN912.3 **文献标识码**: A **文章编号**: 1005-9490(2022)02-0479-04

在日常生活中, 语音是人类进行交流的重要媒介, 语音信号在传达语句含义信息的同时, 也传递了情感信息。同样一句话由于说话人表达的情感不同, 听话者感知时就会有较大的语义差别。要想进一步提高人机交互能力, 实现真正意义的人工智能, 就需要赋予计算机像人一样地观察、理解和生成各种情感特征的能力, 使计算机能够更加自动适应操作者[1]。过去的研究者在进行语音信号处理时, 把语音中这部分信息作为噪声通过规则化处理给去掉了。随着近年来对情感识别研究的深入, 研究者逐渐意识到这些情感信息的重要性, 开始进行专门研究分析, 并将研究成果应用到了各个领域, 获得了很好的经济和社会效益。

语音情感识别中最重要的是分类算法, 应用最广泛的模式分类器有: 隐马尔可夫模型(Hidden Markov Model, HMM)、高斯混合模型(Gaussian Mixture Model, GMM)、支持向量机(Support Vector Mechine, SVM)及人工神经网络(Artificial Neural Network, ANN)等[2]。作为初期计算性能较好的算法, HMM 以一阶 Markov 链为基础发展起来, 有不可见状态和可见状态两种常规状态, 是双重随机过程[3]。Nwe 等[4]通过 HMM 对六种情感进行判断、预测, 最终在缅甸

项目来源: 福建省中青年教师教育科研项目(JAT200535)
收稿日期: 2021-01-19 **修改日期**: 2021-03-17

语料库的识别率达到 78%。GMM[5] 是一种单状态的隐性马尔可夫模型,由于它结构简单所以被广泛用于各种语音信号分类中。GMM 作为统计模型能吸收不同语音信号的声学特性的变动[6],但由于该模型采用状态输出独立假设,影响了其描述语音信号时间上的帧间相关动态特性的能力。本文提出了一种采用相继的复数帧组成的特征参数矢量作为输入特征量的方法来弥补传统 GMM 语音帧间相关动态信息利用不足的问题。然而要很好地利用复数帧段输入 GMM 的关键是要解决当输入特征参数矢量的维数增加时,GMM 输出概率密度函数协方差矩阵的估计误差以及计算量增大的问题。对此,提出一种基于主分量分析神经网络(Principal Components Analysis Neural Network,PCANN)[7] 和 GMM 混合结构的语音情感识别方法,在 GMM 的前端增加了一个用于语音参数压缩的主分量分析神经网络,既改善了状态输出独立 GMM 的缺陷,又解决了上述问题。

1 高斯混合模型 GMM

一个具有 M 个成员的 GMM 的概率密度可由 M 个高斯概率密度函数的加权求和得到,由下式表示[8-9]:

$$p(\boldsymbol{x}|\lambda)=\sum_{i=1}^{M} w_i b_i(\boldsymbol{x}); \quad \sum_{i=1}^{M} a_i=1 \tag{1}$$

这里 $\boldsymbol{x}$ 是 D 维的输入随机向量;$b_i(\boldsymbol{x})(i=1,2,\cdots,M)$ 是第 i 个成员的高斯概率密度函数;$w_i(i=1,2,\cdots,M)$ 是 i 个成员权值系数。完整的 GMM 可表示为:$\lambda=\{w_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}(i=1,2,\cdots,M)$,其中 $\boldsymbol{\mu}_i$ 表示第 i 个成员的平均值向量,$\boldsymbol{\Sigma}_i$ 表示第 i 个成员的协方差矩阵。每个成员密度函数是一个 D 维的高斯分布函数,可由如下表示:

$$b_i(\boldsymbol{x})=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}}\times \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

对于一个长度为 T 的测试输入时间序列 $X=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$,它的 GMM 似然概率可以表示为:

$$P(X|\lambda_j)=\prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda_j) \tag{3}$$

或用对数域表示为:

$$L(X|\lambda_j)=\lg P(X|\lambda_j)=\sum_{t=1}^{T}\lg p(\boldsymbol{x}_t|\lambda_j) \tag{4}$$

假设有 N 个未知类别,分类时运用贝叶斯定理,在 N 个未知类别的模型中,得到似然概率最大的模型对应的类别即为识别结果:

$$j^*=\arg\max_{1\leqslant j\leqslant N} L(X|\lambda_j) \tag{5}$$

2 主分量分析神经网络的原理和算法

主分量分析(Principal Components Analysis,PCA)是一种机器学习算法[10]。主要是通过协方差矩阵将原来维数较高的具有一定相关性的数据,线性组合成维数较少的互不相关的数据[11-12]。利用复数帧段输入 GMM 的关键是要解决当输入特征参数矢量的维数增加时,输出概率密度函数协方差矩阵的估计误差以及计算量增大的问题,在 GMM 的前端增加了一个语音参数压缩的 PCANN。图 1 所示是能够提取前 m 个主分量的 PCANN 结构图[7]。

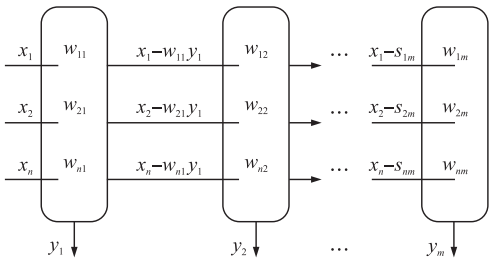


图 1 提取 m 个主分量的神经网络

图中 $s_{im}=\sum_{j=1}^{i} y_j w_{ij}, i\in(1,2,\cdots,n)$。设有 L 个样本的复数帧段输入语音矢量为:$\boldsymbol{X}_l=[x_{li},\cdots,x_{ln}]^{\mathrm{T}}, l\in(1,2,\cdots,L)$,输出语音的主分量特征矢量为:$\boldsymbol{Y}_l=[y_{li},\cdots,y_{lm}]^{\mathrm{T}}, l\in(1,2,\cdots,L)$,通常 $n\gg m$;权值矩阵为:$\boldsymbol{W}=[\boldsymbol{W}_1,\cdots,\boldsymbol{W}_m]$,其中,$\boldsymbol{W}_i=[w_{i1},\cdots,w_{in}]^{\mathrm{T}}$ 为权值向量。则在 PCANN 中利用 Sengar 算法[13],可得到如下第 i 个权值向量修改规则为:

$$\boldsymbol{W}_i[k+1]=\boldsymbol{W}_i[k]+\eta\left(y_{ik}\boldsymbol{X}_l-y_{ik}^2\boldsymbol{W}_i[k]-\sum_{j=1}^{i-1} y_{jk}^2\boldsymbol{W}_j[k]\right) \tag{6}$$

式中:η 为增益因子,η 的选取决定网络收敛的快慢。k 为迭代次数。可以证明,按照公式(6)进行权值迭代更新,网络收敛后,m 个输出的权值向量位于样本协方差矩阵的前 m 个最大特征值对应的特征矢量方向上。利用上述算法提取的多个主分量,在理论上已经能保证各权向量的正交性,但实际应用中发现算法收敛太慢,迭代次数太多。因此实验中我们在训练一定次数以后强制进行一次正交化,从而既可使训练时间大大减少,又能保证得到较好的识别效果。权值的正交化采用格兰姆-施密特规则,设第 $i+1$ 个权向量经去冗余法提取后为:

$$\boldsymbol{W}_{i+1}=\boldsymbol{W}_{i+1}-\sum_{j=1}^{i}\frac{\langle\boldsymbol{W}_{i+1},\boldsymbol{W}_j\rangle}{\|\boldsymbol{W}_j\|}\boldsymbol{W}_j \tag{7}$$

利用 $\|\boldsymbol{W}_j\|=1$,可得:

$$\boldsymbol{W}_{i+1}=\boldsymbol{W}_{i+1}-\sum_{j=1}^{i}\langle \boldsymbol{W}_{i+1},\boldsymbol{W}_j\rangle \boldsymbol{W}_j \tag{8}$$

对其进行归一化可得:

$$\boldsymbol{W}_{i+1}=\frac{\boldsymbol{W}_{i+1}}{\|\boldsymbol{W}_{i+1}\|} \tag{9}$$

有了第 i 个权向量,即可得第 i 个主分量:$y_i=\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{X}_l$。

复数帧段 GMM 的输入是由相继的复数帧特征参数矢量按顺序组合成的一个复合特征参数矢量,每个复数帧段特征参数的段移为一帧。这些复数帧段特征参数作为语音输入特征数据在模型训练和识别时使用。

3　实验和结果

本文使用的语音情感数据库是免费的柏林情感语音库,其采样频率为 16 kHz,16 bit 量化[14]。该语音库分别由十名专业演员(5 男,5 女)在不同情感状态下(高兴、无聊、中性、悲伤、恐惧、厌恶、生气)朗读十句不同文本的德语组成。本实验选取其中的高兴、中性、悲伤、恐惧、厌恶、生气六种情感的语句各 60 条。其中每种情感选 30 条作为训练样本,另外 30 条作为待识别样本,而且训练样本和待识别样本中,男女声音样本比例基本为 1∶1,来验证复数帧段输入 GMM 在语音情感识别中的识别效果。

语音情感识别特征选取部分语音韵律特征和音质特征及其衍生参数共 23 个特征参数,构成用于识别的情感特征向量:特征 1~5 维:短时幅度的均值、最大值、最小值、中值、方差;特征 6~10 维:短时能量的均值、最大值、最小值、中值、方差;特征 11~14 维:短时过零率的均值、最大值、中值、方差;特征 15~18 维:短时基音频率的均值、最大值、中值、方差;特征 19~23 维:短时共振峰频率的均值、最大值、最小值、中值、方差。

评价上述 PCANN/GMM 混合结构语音情感识别方法的识别实验主要是把传统的状态输出独立 GMM 和 PCANN/GMM 混合结构模型进行识别准确率比较。PCANN/GMM 模型的输入分别采用 2 帧、4 帧和 6 帧长度的复数帧。识别结果如表 1~表 4 所示,识别率采用四舍五入法取整数。

由表 1~表 4 的识别测试结果可以看出,PCANN/GMM 的识别效果比状态输出独立 GMM 好,识别率均有所提高。2 帧、4 帧和 6 帧宽度 PCANN/GMM 的平均识别率分别为 76.3%、84.2%和 81.2%,几种情况中,对“生气”的情感识别率普遍较高。另外,4 帧宽度 PCANN/GMM 的识别率最高,4 帧的语音长度能较好地描述帧之间的动态特性,帧数太少,不能较全面完整地利用帧间的特性,随着帧数的增加,帧之间的情感相关性随之减弱,有时甚至会发生情感的转变,从而影响识别率。

表 1　状态独立输出 GMM 情感识别结果

测试样本	识别结果					
	高兴	生气	厌恶	悲伤	恐惧	中性
高兴	77%	6%	7%	2%	3%	5%
生气	4%	80%	7%	2%	5%	2%
厌恶	7%	6%	71%	4%	11%	1%
悲伤	1%	3%	4%	72%	9%	11%
恐惧	3%	5%	10%	8%	71%	3%
中性	5%	5%	2%	12%	3%	73%

表 2　2 帧宽度 GMM 情感识别结果

测试样本	识别结果					
	高兴	生气	厌恶	悲伤	恐惧	中性
高兴	79%	6%	7%	2%	3%	3%
生气	3%	81%	5%	3%	5%	3%
厌恶	7%	6%	73%	3%	10%	1%
悲伤	2%	2%	4%	77%	9%	6%
恐惧	4%	6%	10%	6%	73%	1%
中性	5%	4%	3%	10%	3%	75%

表 3　4 帧宽度 GMM 情感识别结果

测试样本	识别结果					
	高兴	生气	厌恶	悲伤	恐惧	中性
高兴	84%	4%	5%	2%	2%	3%
生气	2%	89%	3%	3%	2%	1%
厌恶	5%	4%	85%	3%	1%	2%
悲伤	0%	1%	4%	86%	3%	6%
恐惧	4%	5%	2%	3%	80%	6%
中性	4%	3%	3%	7%	2%	81%

表 4　6 帧宽度 GMM 情感识别结果

测试样本	识别结果					
	高兴	生气	厌恶	悲伤	恐惧	中性
高兴	84%	3%	5%	3%	2%	3%
生气	3%	83%	6%	4%	2%	2%
厌恶	5%	5%	80%	4%	3%	3%
悲伤	3%	1%	4%	83%	3%	6%
恐惧	4%	3%	2%	7%	79%	5%
中性	5%	3%	4%	7%	3%	78%

4　结论

语音情感识别作为情感计算中的一个重要方面,目的是要让计算机能够理解人类语音中所传递

的情感信息。而由于情感信息的社会性、文化性,以及语音信号自身的复杂性,语音情感识别中尚有许多问题需要解决,特别是符合人脑认知结构与认知心理学机理的情感信息处理算法。本文将主分量分析神经网络与高斯混合模型相结合,研究了其在语音情感识别中的学习能力和识别效果。针对高兴、生气、厌恶、悲伤、恐惧和中性六种基本情感,提取了包括韵律特征与音质特征在内的 23 个情感特征。语音情感识别实验证实了引入帧间相关动态信息方法的有效性。建立一个高效合理的语言情感识别模型仍是研究重点,今后需要进一步探讨主分量分析神经网络与高斯混合模型的结合,特别是优化神经网络的拓扑结构方面还存在许多尚未解决的问题。

参考文献:

[1] 赵力,黄程韦. 实用语音情感识别中的若干关键技术[J]. 数据采集与处理,2014,29(2):157-170.

[2] 张会云,黄鹤鸣,李伟,等. 语音情感识别研究综述[J]. 计算机仿真,2021,38(8):7-17.

[3] 赵力. 语音信号处理[M]. 第 3 版. 北京:机械工业出版社,2016:31-43.

[4] Nwe T L, Foo S W, Silva L C D. Speech Emotion Recognition Using Hidden Markov Models[J]. Speech Communication, 2003, 41(4):603-623.

[5] Reynolds D A, Rose R C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models[J]. IEEE Transaction on Speech and Audio Processing, 1995, 3(1):72-83.

[6] Ehkan P, Allen T, Quigley S F. FPGA Implementation for GMM-Based Speaker Identification[J]. International Journal of Reconfigurable Computing, 2011(1):3-10.

[7] 何振亚,顾明亮. 语音信号的主分量特征[J]. 应用科学学报,1999,17(4):427-431.

[8] Hao T, Chu S M, Hasegawajohnson M, et al. Emotion Recognition from Speech VIA Boosted Gaussian Mixture Models[C]//IEEE International Conference on Multimedia and Expo, New York City, NY USA, IEEE, 2009:294-297.

[9] 王卫东,徐金慧,张志峰,等. 基于密度峰值聚类的高斯混合模型算法[J]. 计算机科学,2021,48(10):191-196.

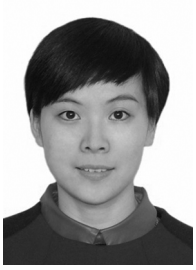
[10] Goodfellow I, Bengio Y, Courville A. 深度学习[M]. 赵申剑,黎彧君,符天凡,译. 北京:人民邮电出版社,2017:30-33.

[11] 郭倩岩,白静. 基于 PCA 鸟群算法的 SVM 参数优化及应用[J]. 计算机工程与设计,2018,39(4):1029-1033.

[12] 李思奇,吕王勇,邓柙,等. 基于改进 PCA 的朴素贝叶斯分类算法[J]. 统计与决策,2022,38(1):34-37.

[13] Oja E. A Simplified Neuron Model as a Principal Components Analyzer[J]. Journal of Mathematical Biology, 1982, 15(3):267-273.

[14] Burkhardt F, Paeschke A, Rolfes M, et al. A Database of German Emotional Speech[C]//Proceedings of Interspeech 2005, Lisbon, Portugul, 2005:1517-1520.



张 霞(1983—),女,工学硕士,莆田学院讲师,研究方向为信号与信息处理、人工智能等,concise.zhang@gmail.com;

赵 力(1958—),男,东南大学信息科学与工程学院教授,博士生导师,研究方向为信号与信息处理等。

杨 勇(1981—),男,河北涉县人,工学博士,现为东南大学信息科学与工程学院博士后,副教授,研究方向为信号与信息处理,YongYang@cumt.edu.cn;