语音情感特征提取及其降维方法综述

刘振焘 徐建平 吴 敏 曹卫华 陈略峰 丁学文 郝 曼 谢 桥

(中国地质大学(武汉)自动化学院 武汉 430074)

(复杂系统先进控制与智能自动化湖北省重点实验室 武汉 430074)

摘 要 情感智能是人工智能的重要发展方向,随着人工智能的迅速发展,情感智能已成为当前人机交互领域的研究热点.语音情感是人们相互情感交流最直接、最高效的途径,越来越多的研究者投入到语音情感识别的研究中.该文综述了国内外近几年语音情感特征提取及降维领域的最新进展.首先,介绍了语音情感识别中常用的特征,将语音特征分为韵律特征、基于谱的特征等,并提出以个性化与非个性化的方式对语音情感特征进行分类.然后,对其中广泛应用的特征提取方法进行了详细地比较与分析,阐述了各类方法的优缺点,并对最新的基于深度学习方法的语音特征提取的相关研究进行了介绍.同时,介绍了常用的语音情感特征降维方法,并在此基础上分析了这些特征降维方法时间复杂度,对比了各类方法的优缺点.最后,对当前语音情感识别领域的研究现状与难点进行了讨论与展望.

关键词 语音;情感特征提取;非个性化特征;特征分类;特征降维;情感智能

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2018.02833

Review of Emotional Feature Extraction and Dimension Reduction Method for Speech Emotion Recognition

LIU Zhen-Tao XU Jian-Ping WU Min CAO Wei-Hua CHEN Lue-Feng
DING Xue-Wen HAO Man XIE Qiao

(*School of Automation, China University of Geosciences, Wuhan* 430074)

(*Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan* 430074)

Abstract To make it easier and more natural to interact with robots, new demands are put forward in human-robot interaction (HRI), and the primary demand that the HRI faces is similar to the communication between humans, which is the emotional intelligence capability, i.e., the ability to identify and feedback emotional information. Emotional intelligence is an important development direction of Artificial Intelligence, which has become a hot topic in the field of human-robot interaction. As a main communication way, speech conveys both semantic and emotional information, so that speech emotion recognition is attracting more attentions. Recent progress in emotional speech feature extraction and feature dimension reduction is reviewed. Firstly, commonly used features in speech emotion recognition are introduced, which are divided into prosody features, spectral related features, personalized features, and non-personalized features. Prosodic features that are closely related to tone, stress, and juncture are commonly

收稿日期:2016-07-26;在线出版日期:2017-07-18.本课题得到国家自然科学基金(61403422,61603356)、湖北省自然科学基金创新群体项目(2015CFA010)、武汉市科技计划项目(2017010201010133)、高等学校学科创新引智计划项目(B17040)、中国地质大学(武汉)自主创新资助计划(1610491T09)资助.刘振焘,男,1981年生,博士,讲师,主要研究方向为计算智能、情感计算.E-mail:liuzhentao@cug.edu.cn.徐建平,男,1993年生,硕士研究生,主要研究方向为语音情感识别、人机交互.吴 敏,男,1963年生,博士,教授,主要研究领域为过程控制、鲁棒控制、智能系统.曹卫华(通信作者),男,1972年生,博士,教授,主要研究领域为过程控制、智能控制、多智能体系统.E-mail:weihuacao@cug.edu.cn.陈略峰,男,1986年生,博士,副教授,主要研究方向为人机交互、计算智能、意图理解.丁学文,男,1995年生,硕士研究生,主要研究方向为人脸表情识别、人机交互.郝 曼,女,1994年生,博士研究生,主要研究方向为情感机器人、人机交互.谢 桥,男,1992年生,硕士研究生,主要研究方向为情感模型、脑电情感识别.

used for emotion recognition with high accuracy, including fundamental frequency, formant, short-term energy, etc. Spectral-based features describe the characteristics of the speech signal from the frequency domain, in which the commonly used features are LPCC (Linear Predictor Cepstral Coefficients), MFCC (Mel Frequency Cepstral Coefficient), and so on. We innovatively sort emotional features in speech by personalized features and non-personalized features. Different speakers have different characteristics of speech, which results in difference in speech features from different speakers with the same emotional states. To build a robust speech emotion recognition model for different people, non-personalized emotional characteristics are proposed based on the personalized features, which are extracted using a derivative-based method. Secondly, some widely used feature extraction methods in speech emotion recognition are compared in terms of their advantages and disadvantages. Pitch, formant, and MFCC features are commonly used prosodic features and their extraction methods are introduced in detail. For the non-personalized feature extraction, we present an extraction method based on derivative for speaker-independent speech emotion recognition. Speech is produced by nonlinear air flow in the vocal system, thus Teager-Energy-Operator (TEO) with nonlinear characteristic can be used for speech signal processing. Speech emotion recognition can be regarded as a static or dynamic classification problem with massive data, and deep learning can be used to figure out this problem, in which Recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) RNNs have been used effectively for speech recognition and natural language processing. These architectures are able to deal with high-dimensional inputs and automatically learn features from these inputs, which are different from traditional features. Thirdly, five feature dimension reduction methods in speech emotion recognition are presented, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS), Isometric Mapping (Isomap), and Locally Linear Embedding (LLE). PCA and LDA are linear methods, while MDS, Isomap, and LLE are nonlinear methods. Comparison experiments of these five methods are carried out to analyze their time complexities, moreover, their advantages and disadvantages are summarized. Finally, the development directions of speech emotion recognition are put forward including correlation analysis between speech features and emotions, speaker-independent speech emotion recognition, combination of semantic cues and emotional speech features for emotion recognition, and so on. In addition, the potential applications of speech emotion recognition are discussed.

Keywords speech; emotional feature extraction; non-personalized features; features classification; feature dimension reduction; emotional intelligence

1 引 言

随着计算机技术的迅速发展,以其为载体的人工智能研究也日新月异.计算机系统如何能够履行仅依靠人类大脑才能够完成的任务,成为人工智能的重点研究内容[1].结合多媒体信息感知、情感识别、智能机器等技术,人工智能的最新研究成果也不断地应用于交通、教育、医疗等领域中[2].而情感是智能的一部分,它是一种特殊的智能[3],因此,计算机在实现智能的同时,也应具有识别、理解和表达人类的情感的能力[4].自从 Picard 教授提出"情感计算(Affective Computing)[5]"的概念以来,情感计算引起了国内外研究者的广泛关注.情感计算是一门涉及认知科学、心理科学等领域的高度综合化研究领域[6],致力于研究机器的情感智能,赋予机器人识别、理解、表达和适应人的情感的能力[7].其中,情感识别是情感计算的关键内容.根据情感信息来源的不同,情感识别可分为情感语音、人脸表情、姿势和生理信号等情感信息的识别.

作为相互传递信息最方便、最基本和最直接的途径，语音包含了丰富的情感信息[8]. 语音不仅传达了语义信息，同时也传递了说话者的情感状态. 例如，愤怒时，说话音量高、语气重、语速加快，而悲伤时语调低沉、语速慢. 传统的语音研究着眼于语音合成、自然语音处理，语音中情感的研究是一个新兴的领域. 人们带着不同的感情说同一句话时，其表达的信息也有所不同，因此，为了让计算机更好地理解人的情感、与人进行更加自然和谐的交互，语音情感的研究是十分必要的.

语音情感识别在人机交互中应用十分广泛. 例如，为节省人力和提高效率，众多公司采用了自动客户服务系统，当用户的情绪较为激烈时，语音情感分析程序可以为用户及时转接人工服务，从而改善服务质量，同时，人工客服的情绪受到客户影响时，语音情感分析程序也能及时提醒客服保持良好的服务态度[9]. 在汽车驾驶中，语音情感识别系统根据司机的语速、音量等信息实时监控司机的情绪状况，并提醒司机保持冷静、安全驾驶，解决“路怒症”问题，防止交通事故的发生[10]. 在远程教育系统中，语音情感识别程序可以识别学生的状态，当学生对课程内容较为困惑时，系统可以适当调整教学难度和进度，从而提高教学质量[11]. 在医学上，语音情感识别系统可以用于辅助残疾人讲话，VAESS[12] 工程开发了一种帮助失语者讲话和表达情感的便携式语音合成器. 情感除了与说话者的语音相关，也与说话者所处的情景有关，研究者将影响我们情绪的环境分为三类：强敏感场景因子、弱敏感场景因子和非敏感场景因子，在心理医学上，基于情景分析的语音情感识别通过对患者所处情景的分析获取情感状态，并给予医师一定的指导，从而更早地将患者的情绪向好的方面引导，帮助人们对不良的情绪进行排解，有效地防止抑郁症的产生[13].

近些年研究者们开展了许多与语音情感识别相关的研究[14-16]. 在这些研究中，大部分采用韵律特征[17]作为语音情感识别的特征参数. 例如 Gharavian 等人[18]提取了基频、共振峰、Mel 系数等参数，并对它们进行相关性分析处理，将得到的 25 维向量经过 FAMNN 分类算法分类后获得了较好的情感识别结果. Li 等人[19]将 TEO 算子应用于非线性语音特征的提取，结合 Mel 倒谱系数(MFCC)特征提取出了 ST_MFCC 特征. Devi 等人[20]总结了语音信号预处理的技巧、常用的短时能量、MFCC 特征以及它们在语音情感识别中的应用. 上述研究中提取的情感语音特征多数针对个性化的语音情感识别，而对于非个性化语音情感识别的特征提取仍然是一个难题[21].

语音情感识别的流程如图 1 所示，包括语音情感特征的提取，语音情感特征的降维和语音情感的分类[8-11,14,22-24]. 语音特征提取的好坏关系到语音情感识别的准确性，合理的特征选择不仅能提高系统的准确性也能改善系统的实时性[25-26]. 原始提取的特征矩阵维度大，直接输入到分类器中进行分类，效果不好，占用的计算资源也比较多，因此原始特征需要进行降维处理. 数据降维不仅可以去除数据间的冗余，起到数据压缩、降低存储空间、消除数据噪声的作用，而且能够提高系统运算效率和识别率[27]. 情感分类是对降维后的特征进行识别处理得到情感结果的过程，常用的识别方法有神经网络、支持向量机等.

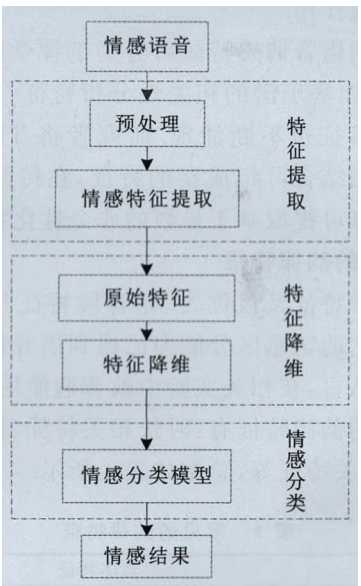

图 1　语音情感识别流程

本文首先对语音情感特征进行了全面地分类与整理. 然后，着重阐述了个性化特征与非个性化语音情感特征的提取方法. 针对基频、共振峰、Mel 倒谱系数三种主要的情感特征，本文详细描述了这些特征与情感语音之间的联系，以及它们的提取方法；针对不同说话人语音情感识别率较低的问题，本文介绍了基于导数的非个性化语音特征提取方法；针对非线性能量算子，本文叙述了其与传统语音特征结合的特征提取方法，同时本文也介绍了深度学习的相关方法在语音情感识别中的应用. 其次，本文评述了几种语音特征的降维方法. 最后，本文分析了当前

语音情感识别领域遇到的挑战与机遇.

本文第 2 节介绍语音情感特征的分类;第 3 节详细阐述几种典型的语音情感特征参数提取方法;第 4 节整理语音情感特征的降维方法;第 5 节总结当前语音情感研究的难点,指出未来语音情感识别的研究趋势.

2 语音情感特征的分类

语音信号包含的信息主要包括语义信息和声学信息. 语义即语音中语言文字的信息,目前,语音文字识别的研究较多. 相比语义信息,语音声学特征包含了更多情感,如说话人的语气、语调,语音情感变化主要是通过声学特征的差异体现,因此,声学特征是语音情感识别中至关重要的因子[28]. 而特征提取是语音情感识别的关键步骤. 通过特征提取得到的能够表现语音情感状态的声学特征在情感信息传递中起了关键作用.

传统的语音情感特征可分为韵律学特征[29]、音质特征[30]和基于谱的相关性分析特征[31]. 近年来,新的语音特征也不断涌现. 研究者将 TEO 算子与传统特征相结合可组成新的特征,在传统特征中添加导数信息可提取基于导数的非个性化特征.

2.1 常见的韵律特征

韵律学特征又被称为"超音段特征"或"超语言学特征",它的情感区分能力已得到语音情感识别领域的广泛认可,在相关实验中被普遍使用[32-34]. 其中最为常用的韵律特征有:时长相关特征、基频相关特征、能量相关特征等,总结如表 1 所示.

表 1 常见的韵律特征

特征类型	具体特征
基频相关特征	基音频率及其均值、变化范围、变化率、均方差
能量相关特征	短时平均能量、短时能量变化率、短时平均振幅、振幅平均变化率、短时最大振幅
时长相关特征	语速、短时平均过零率

2.2 基于谱的相关性分析特征

基于谱的相关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[35],目前基于谱的相关特征主要有线性预测倒谱系数(LPCC)[36]、Mel 频率倒谱系数[37].

通过对传统语音的典型特征的构造,研究人员分别从共振峰构造、基音构造、能量构造、时间构造、Mel 频率倒谱系数(MFCC)、Mel 频谱能量动态系统中提取出大量的语音情感特征,如表 2 所示.

表 2 构造的语音情感特征[38]

特征类型	具体特征
时间构造	短时平均过零率、无声部分时间比率
振幅构造	短时平均能量、短时能量变化率、短时平均振幅、振幅平均变化率、短时最大振幅
基频构造	基频轨迹曲线的最大值、整个曲线的基频平均值、平均变化率、均方差、基音频率 1/3 分位点、1/4 分位点以及基音变化的 1/3 分位点、1/4 分位点
共振峰构造	第一共振峰频率、第二共振峰频率、第三共振峰频率的最大值、平均值、动态变化范围、平均变化率、均方差、1/3 分位点、1/4 分位点
MFCC 系数	12 阶的 MFCC 系数、一阶差分 MFCC 系数、二阶差分 MCFF 系数

2.3 个性化与非个性化特征

在以往情感特征分析的相关研究中,通常存在不同研究者的实验结果具有较大差别[39]的问题,这种研究成果之间较差的可比性一般是由语料库不统一造成的. 当在一个数据库上行之有效的特征,迁移到另一组语料库上就不能获得同样的性能. 因此在近些年的研究中,研究者们继而沿着跨数据库的扩展性方面进行研究. 此外,在语音情感的研究中,还应当考虑人种的差异性,解决不同民族之间、不同语种之间情感表达的无差异问题仍然是一个极大的挑战[40]. 因此,研究人员们就引入了个性化与非个性化特征的概念.

研究人员通过对上述构造特征进行分析发现:根据特征是否受说话者个人说话特征的影响,可将表 2 中的基于声学的语音情感特征分为个性化[41]和非个性化语音情感特征[42]两大类. 第一类特征是直接反映数值大小的语音情感特征,这类情感特征称为个性化语音情感特征;第二类特征是反映情感特征在说话过程中变化情况的特征,这类的特征称为非个性化语音情感特征.

一般认为,对于直接反映数值大小的语音情感特征是不具有代表不同人的共性特征信息,把这类特征归纳为个性化语音情感特征,个性化语音情感特征一般携带了大量的个人情感信息,反映了说话者的说话特点[43],如表 3 所示.

表 3 个性化语音情感特征[42]

特征类型	个性化语音情感特征
时间构造	短时平均过零率
振幅构造	短时平均能量、短时平均振幅、短时最大振幅
基频构造	基频轨迹曲线的最大值、整个曲线的基频平均值、变化范围以及基音频率的 1/4 分位点、3/4 分位点、1/3 分位点和 2/3 分位点
共振峰构造	第一共振峰频率、第二共振峰频率、第三共振峰频率的最大值、平均值、动态变化范围、1/4 分位点、3/4 分位点、1/3 分位点和 2/3 分位点
MFCC 系数	12 阶的 MFCC 系数
Mel 频谱能量动态系数	12 个等间隔的频带上的频谱能量动态系数

个性化语音情感特征在语音情感识别中占多数，因为其可包含特定说话者的情感信息，且利用这些个性化语音情感参数得到的特征向量对于特定说话者具有较高的识别率. 目前大多数研究集中在特定人的语音情感识别上. 虽然可以运用改良的语音情感识别算法获得较高的识别率，却在一定程度上影响了语音情感识别技术在真实自然语境下非特定人语音情感识别的现实化应用.

为了消除个性化语音情感特征对于不同人所呈现出来的数值大小的差异，研究者们引入变化率来反映说话过程中情感的变化情况，即导数概念，以此获得不依赖于人、具有相通性和稳定性的语音情感特征. 这类特征包含了一定的情感信息，又不易受说话者的影响，被称为非个性化语音情感特征. 研究者们通过实验提取出语音情感特征，根据这些特征受说话人变化的干扰情况，确定了基频平均变化率、短时能量平均变化率等非个性化特征，然后对语音情感特征进行排序[44]. 根据这类特征的特点，总结目前的主要非个性特征如表 4 所示.

表 4　非个性化语音情感特征[42]

特征类型	非性化语音情感特征
时间构造	无声部分时间与有声部分时间比率
振幅构造	短时能量平均变化率、振幅平均变化率
基频构造	基频平均变化率、标准方差，基频变化率的 1/4 分位点、3/4 分位点、1/3 分位点和 2/3 分位点
共振峰构造	第一、二、三共振峰频率的平均变化率的 1/4 分位点、3/4 分位点、1/3 分位点和 2/3 分位点
MFCC 系数	1～12 阶的 MFCC 系数

基频特征是反映声门振动的本质特征，对语音情感识别有着重要的作用，不同人的基频差别较大，因此基频特征常应用于说话人识别. 共振峰在决定音质和音色上起重要作用，它在声乐教学以及声纹鉴定中有着广泛的应用. Mel 倒谱系数反应了人耳的听觉特性，在语音识别中被广泛使用. 在语音情感识别方面，传统的语音特征如基频、共振峰、Mel 倒谱系数在语音情感识别中应用广泛. 新的参数如基于 TEO 的特征参数也能产生较好的情感识别结果，部分研究者将深度学习方法应用于语音特征提取上，也取得了较好的效果.

3　语音特征的提取

典型的语音情感识别系统主要包括情感特征的提取、识别，其中情感特征提取的好坏直接影响情感识别的正确率. 由于语音信号的声学特性复杂多样化，因此，正确地从语音信号中找出可体现情感差异的特征参数并且准确地将其提取出来，直接关系到后续情感识别的效果. 目前针对语音情感特征提取的方法众多，对于不同的特征，提取方法也有所不同. 研究人员利用传统声学语音情感特征提取方法提取个性化和非个性化语音特征，然后通过基于导数的非个性语音情感特征提取方法补充了非个性化语音特征.

目前一些研究人员对语音情感特征中相关参数的研究结果表明：语音信号中某些特定的参数和相应情感状态有着明显的联系[45]. 基频相关、共振峰相关以及频率相关的语音情感特征是目前研究者公认的对语音情感识别具有较大贡献的情感特征[46]. 此外，也有较多研究文献对非线性特征参数进行研究，并取得了优于线性特征的研究成果[47-48]. 本节将分别叙述基频、共振峰、MFCC 特征、非个性化特征、TEO 算子特征、基于深度学习的特征的提取方法.

3.1　基频特征提取

基音周期(Pitch)是声带振动频率的倒数. 它指的是人发出浊音时，气流通过声道促使声带振动的周期. 声带振动产生的准周期性的脉冲气流，激励声道发出浊音. 声带震动的周期即为基音周期. 基音周期的估计称为基音检测(Pitch Detection). 基音周期描述了语音激励源的重要特征，是表征语音信号的本质特征的参数，它在语音分析与合成、说话人识别、语音压缩等方面应用广泛. 基频包含了表征语音情感的大量有用信息，是反映情感变化的重要特征之一[49]，因此基音检测对语音情感识别具有重要意义.

Lee 等人[50]提取基频参数，并将这些参数用图形表示出来，分析表明同一语句由同一人用不同情感表达出来时不尽相同，不同人表达同一语句时的基音变化也不相同，最后，通过对语音进行音节划分，对这些音节中的基频进行聚类可以获得较好的情感识别结果. Ali 等人[51]对包含 4 种情感的巴基斯坦语言的情感数据库进行了研究，他们对比了基频和共振峰以及其它声学特征，结果表明基频可以显著提高语音情感识别的准确率，基频在语音情感识别中起到了至关重要的作用.

基音的变化范围大，从 50 Hz～500 Hz，不同人的声道特征也不同，人们在不同情感状态下基音周期也会发生变化，并且基音周期也会受发音词汇的

影响,因此,在情感语音信号中提取基音周期参数依然面临一些问题,主要体现在:

(1)声门信号并不是完整的周期序列,在发音的起始和结束点信号不是周期性的,部分清音和浊音的过渡帧很难判断周期性.

(2)基音变化范围大,从 50 Hz～500 Hz.

(3)声道共振峰太强烈会改变声门的结构,从而影响激励信号的谐波结构,给基音检测造成困难.

由于基音检测的诸多困难,目前尚未有完整的适用于不同年龄、不同语种的基音频率检测方法.研究者们提出了自相关函数法(ACF)[52]、平均幅度差法(AMDF)[53]、小波法[54]等基音频率检测方法.这 3 类方法是目前常用的基频提取方法,ACF 和 AMDF 属于时域检测算法,它们根据基音周期短时周期性的特点进行提取的.小波法属于频域检测算法.ACF 中的短时自相关函数为

$$R_i(k)=\sum_{m=1}^{N-k}S_i(m)S_i(m+k) \tag{1}$$

其中,$R_i(k)$表示第 i 帧自相关函数,$S_i(m)$表示一帧语音信号的第 m 个采样值,N 表示帧长,k 表示时间的延迟量.自相关函数与原语音信号的周期一致,通过寻找自相关函数波峰的延迟即可找到原信号的周期.ACF 方法在文献[55-56]中得到了应用.

AMDF 中的平均幅度差函数为

$$D_i(k)=\sum_{m=1}^{N-k}|S_i(m+k)-S_i(m)| \tag{2}$$

其中,$D_i(k)$表示第 i 帧平均幅度差函数.平均幅度差函数与原语音信号的周期性一致,它在周期的整数倍上取极小值,通过寻找波谷之间的延迟即可找到原信号的周期.AMDF 方法在文献[57]中得到了应用.

语音信号经过小波变换后极值点对应原信号的不连续点.发声时,来自肺部周期性的气流冲击声门,使声门产生周期性的开启或闭合,这样,语音信号就产生了不连续点.通过寻找小波变换后极值点之间的距离即可确定基音周期.小波法在文献[58]中得到了应用.

ACF 方法通过计算语音信号的自相关函数,寻找周期函数的峰值来确定基音周期,当语音信号信噪比降低时,基音频率和第一共振峰频率比较接近,这时 ACF 方法会带来半频和倍频检测误差[59].AMDF 方法通过计算语音信号的平均幅度差函数,寻找周期函数的波谷值来确定基音周期,这种方法运算量小,但当语音信号变化幅度较快时,也容易出现半频和倍频检测误差.小波法利用小波变换对信号进行分析,语音信号变换成为正弦信号,正弦信号的振荡周期即为基因周期.小波法对一整段语音信号进行分析,能够提取基频周期的包络,基频包络能够精确地反映基频的变化,但是小波法的计算较为复杂,不能抑制噪声的干扰.由于检测的诸多困难,目前还没有一种算法能够在不同的环境下精确地检测基音频率,研究者也提出了新的检测方法,如 Wang 等人[60]提出的 NACF-CBS 算法、Hu 等人[61]提出的 SWT-HS 算法,这些新的算法实用性更好且具有更好的鲁棒性.

3.2 共振峰特征提取

根据声学观点,声道可以看作非均匀截面的声管,当声音激励信号的频率与声道频率一致时,声道将发生共振,产生的波形称为共振峰.共振峰是语音信号处理最重要的参数之一[62],它决定着元音中的音质.共振峰参数包括共振峰频率和共振峰带宽.和基频一样,它反映声道系统的物理参数,在语音识别、声纹鉴定中应用广泛.

文献[63-65]研究表明,共振峰对语音中的情感表达起着关键的作用.在不同的情感中,人们的说话方式会发生很大的改变.而这一改变会直接导致声道的形状发生改变,从而改变了说话者的固有频率,这样就会得到不同的共振峰.不同情感发音的共振峰位置不同,情感状态发生变化时前三个共振峰的峰值变化较大,且其峰值从低到高依次为第一共振峰、第二共振峰和第三共振峰[66],因此,研究人员一般选取第一共振峰、第二共振峰、第三共振峰的平均值、最大值、最小值、动态变化范围、平均变化率、均方差,共振峰频率的 1/4 分位点、1/3 分位点以及共振峰变化的 1/3 分位点、1/4 分位点等统计特征作为研究的对象.

目前,若要精确提取共振峰的参数仍然较为困难,主要表现在:

(1)共振峰决定了频谱包络的最大值,有时候虚假峰值的出现会影响共振峰的估计值.

(2)高音调语音的谐波间隔较大导致频谱包络的样本点太少,这会给提取频谱包络带来困难.

(3)相邻两个共振峰间隔很近时难以分辨.

共振峰信息包含在信号频谱包络之中,频谱包络的估计是共振峰提取的关键,一般认为,频谱包络的最大值即为共振峰.目前常用的共振峰提取方法包括:倒谱法[22]、线性预测(LPC)[67]分析方法以及带通滤波组法[68]等.直接对信号进行离散傅里叶变

换(DFT)，运用 DFT 频谱来提取共振峰，但是 DFT 谱受基频谐波的影响，最大值会出现在谐波频率上，误差较大. 倒谱法采用同态解卷技术，将基音信息和声道信息分离开来，从而可以直接求取共振峰参数，这种方法相对直接进行 DFT 运算求取共振峰更加精确，避免了由基音谐波频率产生的误差，其在文献[69-70]中得到了应用. LPC 法的基本思想是语音信号可由过去若干个语音采样点的线性组合来逼近，通过使预测的采样值与实际输出值的方差最小可以求取一组线性预测系数，由此可得到声道的传递函数为

$$H(z)=\frac{G}{A(z)}=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{3}$$

其中，G 为增益，a_k 为模型的系数，p 为模型的阶数. 这是一个全极点模型. 对 $H(z)$ 取模可以得到声道传递函数的功率谱，根据功率谱可以较为精准地检测出带宽和中心频率. 线性预测(LPC)分析方法的主要特点在于能够在由预测系数构成的多项式中精确地估计共振峰参数[71-72]. 线性预测提供了一组优良的语音信号模型参数，比较精确地表征了语音信号的幅度谱[73]. LPC 法在文献[74-75]中得到了应用. 然而 LPC 法求得的声道传递函数是一个全极点模型，对于某些含有零点的语音信号来说，$A(z)$ 的根反映了零极点的复合效应，无法区分这些根是否是声道的谐振点.

倒谱法计算量小，目前被广泛使用，但是它易受合并共振峰和虚假共振峰的影响. LPC 法计算量大，有研究者对其改进来简化计算，但精确度不高. 同时，新的方法也不断提出，Gowda 等人[76]提出了 dSWLP-GD 方法，Smit 等人[77]提出了 wGDF-CI 方法，这些新方法具有更好的效果和鲁棒性.

3.3　Mel 频率倒谱系数提取

Mel 频率倒谱系数(MFCC)是根据人的听觉机理发现的特征参数[78]，它与频率成非线性对应关系. 人耳听觉机理的研究表明，人耳对不同频率声波的敏感程度是非线性的. 在 1000 Hz 以下，人耳对声音的感知能力与频率成线性关系，而在 1000 Hz 以上，人耳对声音的感知能力与频率成非线性关系. Mel 频率倒谱系数就是利用了这种非线性关系，得到频谱特征，它是基于人耳听觉特性的、鲁棒性较好的频域语音特征参数，其频率的对应关系为

$$Mel(f)=2595\log_{10}\left(1+\frac{f}{700}\right) \tag{4}$$

语音的音高在客观上采用频率来表示，但人耳的听觉效应是非线性的，人们感受到的音高和频率不成比例. 人耳主观上用 Mel 来度量音高的大小. 规定 1000 Hz，40 dB 的语音信号音高为 1000 Mel. 在 Mel 刻度上人耳对语音音高的主观感受是线性的. 人耳基底膜相当于一个非均匀滤波器组，它不同地方的细胞膜对频率的响应不同，每一部分对应一个滤波器群，每一个滤波器群对应一个中心频率和带宽，而每个滤波器的带宽大约为 100 Mel. 为了模拟人耳的特点，研究者们根据人耳滤波器组的中心频率和带宽设计了一组 Mel 滤波器，其波形如图 2 所示.

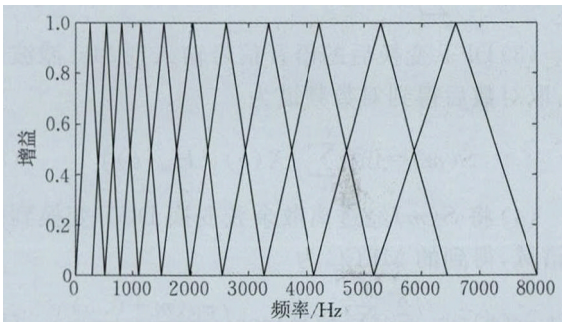

图 2　Mel 滤波器组

语音信号经过 Mel 滤波器处理后可以得到近似人耳的信号. 在语音频谱范围内设置若干个带通滤波器 $H_m(k)$，$0\le m<M$，M 为滤波器的个数，通常范围为 24～40，每个滤波器为三角形滤波特性，中心频率为 $f(m)$，随着 m 的增加，滤波器的间隔逐渐增大. 每个带通滤波器的传递函数为

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)<k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k\le f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{5}$$

其中，

$$f(m)=\left(\frac{N}{f_s}\right)Mel^{-1}\left(Mel(f_l)+m\frac{Mel(f_h)-Mel(f_l)}{M+1}\right) \tag{6}$$

其中，N 为帧长，f_s 为采样频率，Mel^{-1} 为 Mel 函数的逆函数，f_l 为频率范围内的最低频率，f_h 为最高频率.

在 3.2 节共振峰提取中介绍了语音倒谱分析的概念，在语音倒谱分析中加入 Mel 滤波器，所得到的结果即为 Mel 频率倒谱系数，具体过程如图 3 所示.

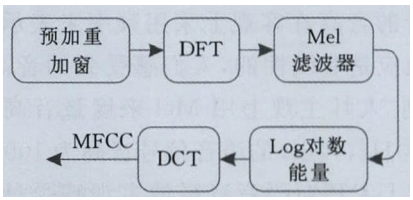

图 3 MFCC 系数提取过程

MFCC 的提取步骤为：

(1) 语音信号经过预加重、分帧、加窗得到预处理后的语音信号 $x(n)$.

(2) 对预处理后的信号进行 DFT 变换得到离散谱 $X(k)$，变换公式为

$$X(k)=\sum_{n=0}^{N-1}x(n)\mathrm{e}^{-\mathrm{j}2\pi nk/N},\ k=0,1,\cdots,N-1 \quad (7)$$

(3) DFT 变换后的语音信号输入到 Mel 滤波器组，取对数后得到对数频谱为

$$S(m)=\ln\Big(\sum_{k=0}^{N-1}|X(k)|^2H_m(k)\Big) \quad (8)$$

(4) 将 $S(m)$ 经过离散余弦变换 DCT 变换到倒频谱域，得到的 MFCC 为

$$c(n)=\sqrt{\frac{2}{M}}\sum_{k=0}^{M-1}S(m)\cos\Big(\frac{\pi n(m+0.5)}{M}\Big) \quad (9)$$

MFCC 系数较好地模拟了人耳听觉系统感知信号的能力，具有鲁棒性强、识别率高的特点，广泛应用于语音处理系统中. 但语音是动态变化的信号，MFCC 参数没有考虑语音信号分帧后帧与帧之间的关系以及同一帧语音信号 MFCC 参数之间的关系. 为了反映语音信号的动态特性，研究者们提出一阶差分 MFCC 系数，其计算方法为

$$\Delta c(n)=\frac{\sum_{k=-K}^{K}k\cdot c(n+k)}{\sum_{k=-K}^{K}k^2} \quad (10)$$

其中，$c(n)$ 为 MFCC 系数，K 为常数，通常取 2. 一阶差分 MFCC 是 MFCC 的一阶导数，该计算方法是使用最小二乘法对局部斜率进行估计，从而得到一阶导数的平滑估计. 除了一阶 MFCC 系数，研究者也提出了二阶 MFCC 系数，将一阶 MFCC 系数代入式(10)就得到二阶 MFCC 系数.

近年来国内外很多的学者研究 Mel 频率倒谱系数，并把它运用到了语音情感识别中[79-82]. 传统的 Mel 系数提取方法就是使用滤波器从低频到高频带内提取信号并做进一步的处理后作为语音信号特征[83]. 研究数据表明，基于频域的参数对于情感识别是非常有效的[23]. Mel 频率尺度的值大体上对应于实际频率的对数分布关系，更符合人耳的听觉特性.

3.4 基于导数的非个性语音情感特征提取

语音信号是由气流冲击声门产生的. 由于声道的差异，不同人具有不同特点的情感特征. 基频、共振峰是根据发声系统得出来的物理参数，它们反映了声道的结构特性，这类参数因人而异，并且携带了大量的说话人的个性化的信息，在说话人身份识别中有很多应用[24]. 将它们应用于说话人的情感识别，对于特定人的识别往往是有效的，对于非特定人来说，效果不如特定人好. 不同人有自己独特的语音特征，某个人的特征模型往往不适用于其他人.

对于同一区域的人，情感是相通的，他们的情感表达方式相同. 因此，研究者试图寻找他们的共性特征，这一特征称之为非个性化特征. 非个性化特征具有不依赖于人、通用性较好的特点.

个性化特征包含了特定人丰富的情感信息，为了消除其在不同人之间的差异性，引入变化率来反映这类特征的变化情况[38]. 变化率的引入消除了个性化的影响，并且反映了一种趋势，而这种趋势是共同的，这也符合同一地区情感表达方式一致的特点，因此，非个性化特征也包含了一定的情感信息，这种信息不受说话者的影响.

文献[41]提出了非个性化特征的概念，并将其应用于语音情感识别中. 文中提取了基频、短时能量、共振峰的变化率以及它们的变化范围、方差等统计值作为非个性化特征，同时提取了传统的基频、共振峰等个性化特征，并用这两类特征分别进行实验，结果表明非个性化特征同样对语音情感识别有着很大的作用，并且这类特征受不同说话者的影响更小.

3.5 基于 Teager 能量算子(TEO)非线性特征提取

TEO(Teager Energy Operator)能量算子是由 Teager 提出的一种非线性算子[84]，表述为

$$\psi[x(n)]=x^2(n)-x(n+1)x(n-1) \quad (11)$$

TEO 已经成功地应用于语音情感识别的信号处理中. 研究人员在 TEO 算子的基础上，做了很多改进的算法，使得情感识别率有了显著的提升[85-89].

声学理论认为来自肺部的气流激励声门发出声音，气流在声道内以平面波的形式传播，由于声道具有不同的特征，不同人的声音特点也不相同. Teager 的研究表明，气流通过声道时会出现气流的附着、分离，形成涡流，涡流与平面波一起构成语音信号的影响因素. 这种涡流被认为是非线性的，而且广泛存在于声道内. Teager 根据这些特点提出了 Teager 能

量算子，Kaiser 给出了 TEO 算子的运算公式，式(11)是 TEO 算子的离散形式，语音信号的 TEO 算子只与该样本点和它前后各一个样本点有关，因此 TEO 的计算十分简单. 情感语音可以分为线性和非线性部分，将 TEO 算子引入语音信号的研究中，有助于解决语音信号的非线性问题.

设正弦信号为 $x_n=A\cos(\omega n+\phi)$，其中，A 为振幅，ϕ 为初始相位，$\omega=2\pi f/f_s$，f 为信号频率，f_s 为采样频率. 该信号的 TEO 变换为

$$\psi[x(n)]=A^2\sin^2(\omega)\approx A^2\omega^2 \tag{12}$$

由式(12)可知 TEO 变换后包含振幅和角速度信息. TEO 变换只采用 3 个点便可跟随信号幅值和角频率的变化.

设 $x(n)=s(n)+\sigma(n)$，$s(n)$为原始信号，$\sigma(n)$为期望为 0 的噪声信号，$x(n)$经过 TEO 变换后的期望为

$$\psi[x(n)]=\psi[s(n)]+\psi[\sigma(n)]+2\bar{\psi}[s(n),\sigma(n)] \tag{13}$$

其中，

$$\bar{\psi}[s(n),\sigma(n)]=s(n)\sigma(n)-0.5s(n-1)\sigma(n+1)-0.5s(n+1)\sigma(n-1),$$

其中，$s(n)$和 $\sigma(n)$为互相独立的两个信号，因此，$E(\bar{\psi}[s(n),\sigma(n)])=0$，则式(13)的期望值为

$$E(\psi[x(n)])=E(\psi[s(n)])+E(\psi[\sigma(n)]),$$

其中，$E(\psi[\sigma(n)]$远小于 $E(\psi[s(n)])$，因此 $E(\psi[x(n)])\approx E(\psi[s(n)])$. 由此可见，TEO 算子可以消除噪声的影响，其计算量也较小.

TEO 算子可与传统的基频、共振峰特征相结合形成新的语音特征. 文献[90]根据当前的涡流非线性理论和目前已经成熟的线性理论，采用非线性能量算子与传统语音分析参数相结合的方法，探索出有效的情绪识别特征. 语音信号经过 TEO 变换后，再利用相关的特征提取方法可以获得基于 TEO 的语音情感特征. 文中提取了 5 个基于 TEO 的特征：基于频域 TEO 的 Mel 倒谱参数 NFD_Mel、基于幅频特性的 Mel 倒谱参数 AF_Mel、基于微分幅频特性的 Mel 倒谱参数 DAF_Mel、基于幅度调制的子带倒谱参数 AM_SBCC、基于幅频调制的子带倒谱参数 AMFM_SBCC. 结果表明，基于 TEO 的这 5 类特征的情感识别效果优于传统的 MFCC 特征.

3.6　基于深度学习的特征提取

深度学习方法在处理复杂的海量数据建模上有很大优势，近年来，计算能力的提升使深度学习方法广泛应用于语音识别和图像处理等领域. 深度学习方法多用于数据分类的建模，部分研究者将深度学习应用于语音特征的提取，并取得了一定的成果.

文献[91]将瓶颈结构(Bottle-Neck，BN)和深度信任神经网络(Deep Belief Network，DBN)相结合，提出了一种新的特征提取方法，称为 BN-DBN 方法. BN-DBN 的网络组成结构上通常被设定为一个奇数层的多层 ANN，并将其中神经元个数相对于其它层较少的一层命名为瓶颈层. 基于 BN-DBN 的语音特征提取方法首先利用原始语音信号提取的 MFCC 作为输入，经过若干个显层和隐层对网络进行训练，并采用类似传统 BP 神经网络的监督学习方式，对整个 DBN 进行由后至前的微调，最终建立 DBN. 在得到训练好的瓶颈层之后，将瓶颈层之后的网络去掉，将原瓶颈层的输出作为新的语音特征. BN-DBN 具有 DBN 和瓶颈特征的共同优点，主要表现为：

(1) 具有强大的对数据内部结构和统计特征的表征能力，提取的特征更能反映不同数据的本征特征；

(2) 预训练部分采用了无监督学习模式，因此可以有效利用未标签数据；

(3) 微调部分采用监督式学习的方法，可以根据标签信息对神经网络参数进行调整，使得提取的特征更具有判决性，利于分类；

(4) 人为设定的参数少，可以由机器主动学习进行训练得到特征.

文献[92]将卷积神经网络(Convolutional Neural Network，CNN)运用到语音特征提取中. 卷积神经网络是由一对或者多对卷积层(C 层)和最大层 pooling 层(S 层)组成的. 卷积层主要采用的是一个卷积核对输入信号来进行滤波，卷积核在整个信号空间中重复过滤. S 层类似一个最大值的过滤器，对卷积后的结果进行二次采样，选取一定范围内的最大值作为滤波输出. 语音特征提取的 CNN 网络首先要建立由 C 层和 S 层组成的多层网络，随机求取网络各节点初始值. 然后在网络最底层链接 softmax 分类器，以声韵母为基元标签采取 BP 算法对参数进行有监督微调. 最后去掉 softmax 分类器，截取剩下的网络作为 CNN 特征变换映射过程. 在语音特征提取时，首先对语音信号进行基本的特征提取得到 Mel 滤波器的输出特征，将这些基本特征作为 CNN 神经网络的输入，通过 softmax 分类器进行 BP 算法训练后，选取倒数第二层神经元数据作为 CNN 特征输出，以此产生新的语音特征. 输入信号先经过 C 层，卷积过程会把噪声维度上的差异均摊

到周围的维度上，模糊同一基元特征间的差异性，从而降低基元匹配错误的概率；S 层通过降采样的方式降低卷积结果的维度，从而降低计算负担. S 层可以解决非特定人语音识别中由于不同人器官造成和发生习惯的不同而导致的差异. CNN 网络具有类内收敛，类间发散等特点，更适用于语音识别的分类任务.

除了上述两种特征提取方法以外，循环神经网络(Recurrent Neural Networks，RNN)和长短时记忆(Long Short Term Memory，LSTM)等方法也能进行语音特征的提取. 文献[93]使用 RNN 网络对语音特征进行提取，并将 RNN 网络的输出结果作为 LPC 和 MFCC 的输入，使用这种方法获得了较好的效果. 文献[94]使用 LSTM 方法对语音特征进行提取，其利用 LSTM 网络的增强特性进行区分性训练，大大提高了特征的有效率.

4 常用特征降维算法

根据上述语音特征提取方法得到的语音情感特征数据维度较高，不同特征之间相关性强，存在特征冗余的问题. 并且维度的数据对数据样本的规模提出了更高的要求，当数据维度远大于样本数目时，易造成过度拟合的问题，不利于情感模型的建立，从而影响分类器的精度[95]. 此外，过高维度的特征数据会增加训练时间，降低分类的实时性，也会耗费计算机资源的消耗. 因此，对于提取的语音情感特征数据应首先进行选择分析与降维处理. 通过对语音情感特征的选择与降维，冗余特征被具有表征同一特性的最优特征筛除，最终得到对各类情感具有较强贡献的语音情感特征组合.

目前常用的特征降维方法分为线性与非线性方法. 通过线性降维方法得到的低维数据仍可保持高维数据之间的线性关系. 线性方法主要包括主成分分析法(PCA)[96]、线性判别分析法(LDA)[97]、局部保留投影法(LPP)[98]等. 但是针对高度非线性结构的数据集合来说，非线性降维方法更能揭示数据的内在特征. 语音情感特征数据的非线性降维方法主要有多维尺度分析法(MDS)、等距映射法(Isomap)[99]、局部线性嵌入法(LLE)[100]、拉普拉斯特征映射法(Laplacian Eigenmaps)[101]等. 上述降维方法中，LPP、LLE 与 Laplacian Eigenmaps 均为局部的数据降维方法，而 PCA、LDA、Isomap 为全局降维方法. 其中，局部方法仅考虑样本点与邻近点之间的关系. 全局方法既考虑样本集合的局部信息，同时又综合了样本集合的全局信息，考虑了样本点与非临近点之间的关系，从而使得数据流形上相距较远的样本特征也较远.

(1) 主成分分析法(PCA)

PCA 可将高维数据映射到低维空间中，它是通过协方差矩阵进行特征分解，得出数据的主成分和权值，再根据数据的权值选择具有代表性的特征向量. 它用数据投影后方差的大小衡量特征代表信息量的多少，方差越大，代表携带的信息就越多，选取前 k 维方差较大的特征，投影后的数据满足方差最大化，即数据点的分布稀疏，便于分类，此种数据降维方法可最大程度接近原始数据，但其并不着重探索数据的内部结构特征. 图 4 是 5 个点的 PCA 分析过程，图中直线是这 5 个点的投影方向，数据点在该投影线上的投影后数据点方差是最大的，这样数据点投影后分散地最开，数据更容易区分.

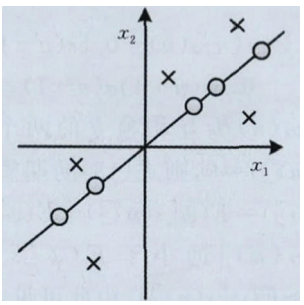

图 4 五个样本内 PCA 过程

设样本点的数量为 m，维度为 n，数据降维后的维度为 k，投影直线的方向向量为 $\boldsymbol{u}$. 首先对样本求取平均值，然后用样本点减去平均值，处理后样本点的平均值为 0，那么样本点投影后的方差为

$$\delta=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u}=\boldsymbol{u}^{\mathrm{T}}\left(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\boldsymbol{u} \tag{14}$$

其中，δ 表示数据投影后的方差，$\boldsymbol{x}_i$ 为预处理后的样本数据，$\boldsymbol{u}$ 为投影直线的方向向量，可令 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1$. $\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ 为样本的协方差矩阵，用 $\boldsymbol{A}$ 来表示，则式(14)可以表示为

$$\delta=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u} \tag{15}$$

式(15)中 δ 的最大值即为投影点对应的方差最大值. 运用拉格朗日极值法，式(15)可变换为

$$\boldsymbol{A}\boldsymbol{u}=\delta\boldsymbol{u} \tag{16}$$

其中，投影方差 δ 为样本点协方差矩阵 $\boldsymbol{A}$ 的特征值，投影直线的方向向量 $\boldsymbol{u}$ 为 $\boldsymbol{A}$ 的特征向量. 对 $\boldsymbol{A}$ 的特征值进行排序，取前 k 个较大的特征值并计算它们

对应的特征向量，这些向量就是样本点投影直线对应的方向向量. 这些特征向量构成了投影矩阵，样本数据与该投影矩阵相乘就得到了降维后的 k 维向量. 其在文献[102]中得到了应用.

PCA 是一种无监督的线性降维方法，适用于有线性关系数据的降维，在人脸识别和图像处理方面有广泛的应用. 该方法概念简单，计算方便，时间复杂度较低，降维后的数据保留了大部分原始数据特征，但降维后的数据维度的确定没有明确准则.

(2) 线性判别分析法(LDA)

LDA 又叫 Fisher 线性判别法，其基本原理是通过 Fisher 准则函数选择某一个最佳的投影方向，使得样本投影到该方向后有最大的类间区分度和最小的类内离散度，以达到抽取分类信息和类别之间最佳的可分离性. 此方法的输入数据带有标签，为有监督的降维方法. 图 5 是二维平面中两类数据的 LDA 降维过程，图中两块椭圆区域分别表示需要降维的二个数据集，图中直线为投影方向，可以看出，数据投影到该方向后可以用投影直线上的一维数据点来表示这两类数据，并且数据集类间的距离最大，数据集的类内距离最小.

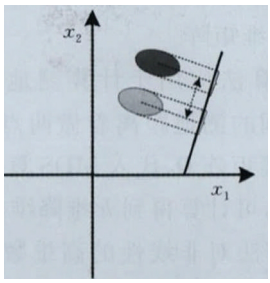

图 5　两类数据的 LDA 分析

设样本点的数量为 m，维度为 n. 对于二类降维问题，投影直线为 $y=\boldsymbol{\omega}^{\mathrm{T}}x$，样本投影后的中心点为

$$\tilde{\boldsymbol{\mu}}_i=\frac{1}{N_i}\sum_{j\in\boldsymbol{\omega}_i}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_j=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\mu}_i \tag{17}$$

其中，N_i 表示第 i 类数据的个数，$\boldsymbol{\omega}_i$ 为第 i 类数据的数据集，$\boldsymbol{\mu}_i$ 为原数据的中心点. 投影后样本的类内方差为

$$\tilde{s}_i^2=\sum_{j\in\boldsymbol{\omega}_i}(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}_i)^2=\sum_{j\in\boldsymbol{\omega}_i}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\omega} \tag{18}$$

LDA 要求样本投影后具有最小的类内距离和最大的类间距离，综合两者考虑，其评价函数为

$$J(\boldsymbol{\omega})=\frac{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2}{\tilde{s}_1^2+\tilde{s}_2^2} \tag{19}$$

其中，

$$|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2=\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\omega} \tag{20}$$

记 $\boldsymbol{S}_B=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}$，$\boldsymbol{S}_i=\sum_{j\in\boldsymbol{\omega}_i}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}$，$\boldsymbol{S}_w=\boldsymbol{S}_1+\boldsymbol{S}_2$，式(19)的评价函数可表示为

$$J(\boldsymbol{\omega})=\frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}} \tag{21}$$

其中，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{\omega}$ 表示投影后的类间分散，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}$ 表示投影后的类内分散. $J(\boldsymbol{\omega})$的最大值就代表了类间距离的最大值和类内距离的最小值. 令 $|\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}|=1$，构造 $J(\boldsymbol{\omega})$的拉格朗日函数

$$L(\boldsymbol{\omega})=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{\omega}-\lambda(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}-1) \tag{22}$$

$L(\boldsymbol{\omega})$对 $\boldsymbol{\omega}$ 求导得

$$\boldsymbol{S}_w^{-1}\boldsymbol{S}_B\boldsymbol{\omega}=\lambda\boldsymbol{\omega} \tag{23}$$

因此，所求直线的方向向量就是矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_B$ 的特征向量. 对于具有多个类别的降维问题，这类别数量为 C，投影后的类内距离为 $\boldsymbol{\omega}^{\mathrm{T}}\sum_{i=1}^{C}\boldsymbol{S}_i\boldsymbol{\omega}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}$，类间距离为

$$\sum_{i=1}^{C}N_i(\tilde{\boldsymbol{\mu}}_i-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_i-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{\omega} \tag{24}$$

其中，$\tilde{\boldsymbol{\mu}}$ 为投影后所有数据的中心点，$\tilde{\boldsymbol{\mu}}=\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{\mu}}_i$，$N$ 为样本的总个数. 多分类问题的降维目标函数依然为式(21). 因此，对矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_B$ 的特征值进行排序，取前 k 个特征值对应的特征向量即为最佳的 k 条投影曲线.

线性判别分析法(LDA)能够保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离，即模式在该空间中有最佳的可分离性，该方法属于监督学习的算法，但是对于不满足高斯分布的样本并不适用. LDA 在文献[103]中得到了应用.

(3) 多维尺度分析法(MDS)

MDS 降维算法可以解决非线性数据的降维问题，它的原理是通过输入相似程度矩阵，在低维空间中找到相对位置坐标，利用欧氏距离来计算两点之间的距离，根据距离的长短来判断相似程度的大小. 它的基本思想是保留数据之间的相似性，可以分为经典 MDS、度量性 MDS 和非度量性 MDS. 下面主要介绍经典 MDS 算法.

设有 m 个 d 维样本，它们之间的欧氏距离定义为

$$\boldsymbol{\Delta}=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm}\end{pmatrix} \tag{25}$$

其中，$\sigma_{ij}=\sum_{p=1}^{d}\sqrt{(r_{ip}-r_{jp})^2}$ 为第 i 个样本和第 j 个样本之间的距离. MDS 的思想是给定样本的距离矩阵 $\boldsymbol{\Delta}$，寻找低维特征向量 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$，使得

$$|\boldsymbol{x}_i-\boldsymbol{x}_j|\approx\sigma_{ij} \tag{26}$$

这些向量与原数据距离尽可能接近，故其评价函数为

$$\min\sum_{i<j}(|\boldsymbol{x}_i-\boldsymbol{x}_j|-\sigma_{ij})^2 \tag{27}$$

降维后的矩阵 $\boldsymbol{X}$ 的求解是对矩阵 $\boldsymbol{\Delta}$ 的双重中心化矩阵进行奇异值分解得到的，矩阵 $\boldsymbol{\Delta}$ 的双重中心化矩阵为

$$\hat{\boldsymbol{\Delta}}=-\frac{1}{2}J\boldsymbol{\Delta}^{(2)}J=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \tag{28}$$

其中，$J=E-\frac{1}{m}$，$\Delta_{ij}^{(2)}=\sigma_{ij}^2$，$\hat{\boldsymbol{\Delta}}$ 的元素为

$$\begin{aligned}\hat{\Delta}_{ij}&=-\frac{1}{2}\left(\sigma_{ij}^2-\frac{1}{m}\sum_{k=1}^{m}\sigma_{ik}^2-\frac{1}{m}\sum_{l=1}^{m}\sigma_{jl}^2+\frac{1}{m^2}\sum_{l=1}^{m}\sum_{k=1}^{m}\sigma_{lk}^2\right)\\&=\boldsymbol{x}_i\cdot\boldsymbol{x}_j\end{aligned} \tag{29}$$

矩阵 $\hat{\boldsymbol{\Delta}}$ 是对称且半正定的，对矩阵 $\hat{\boldsymbol{\Delta}}$ 进行奇异值分解

$$\hat{\boldsymbol{\Delta}}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^{\mathrm{T}} \tag{30}$$

其中，$\boldsymbol{\Lambda}$ 为 $\hat{\boldsymbol{\Delta}}$ 的特征值组成的对角矩阵，$\boldsymbol{U}$ 为 $\hat{\boldsymbol{\Delta}}$ 的特征向量. 对矩阵 $\hat{\boldsymbol{\Delta}}$ 的特征值进行由大到小排序，选取前 k 个较大的特征值和它们对应的特征向量，令 $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$ 即求出降维后的样本 $\boldsymbol{X}$.

多维尺度分析法(MDS)是一种将多维空间的研究对象(样本或变量)简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法. 它不仅适用于探索变量之间的潜在规律性联系，也能处理名称变量和顺序变量数据，而且不要求数据满足多元正态分布假设，现已被广泛应用于各研究领域.

(4) 等距映射法(Isomap)

针对语音情感特征数据的非线性降维方法，文献[104]研究了基于增强型核函数的 Isomap 数据降维方法，其可从原始高维语音特征数据中提取出低维度的、强鉴别能力的嵌入式数据特征，显著地提高了语音情感的识别精度. Isomap 是对 MDS 算法的一种改进，MDS 算法适用于欧式空间，它用欧式距离来衡量两点之间的距离大小，而对于流形结构，欧氏距离不再适用，故 Isomap 采用测地线来计算流形中的距离. 对于 m 个 d 维样本集合 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$，降维后为维度 k 的集合 $\{\boldsymbol{y}_1,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_m\}$，Isomap 的计算步骤如下：

① 构造邻接图

计算两两样本点之间的欧氏距离 d_{ij}. 当 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的前 p 个距离最近的点或者当 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_i$ 之间的欧氏距离 d_{ij} 小于某个特定的常数 ε 时，称 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的近邻点. 根据点与点之间的近邻关系构造领域连接图，当 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 互为近邻关系时，用一条直线连接两点，认为这两点之间的边长为 d_{ij}，否则，两点之间没有边，边长为 ∞.

② 计算两点之间的测地距离

两点之间的测地距离计算公式为

$$d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)=\begin{cases}d_{ij}, & \boldsymbol{x}_i\text{与}\boldsymbol{x}_j\text{互为近邻点}\\ \min\{d_G(\boldsymbol{x}_i,\boldsymbol{x}_j),\ d_G(\boldsymbol{x}_i,\boldsymbol{x}_l)+d_G(\boldsymbol{x}_l,\boldsymbol{x}_j)\} & \text{其他}\end{cases} \tag{31}$$

如果 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间有边连接，那么测地距离 $d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 就是两点之间的欧氏距离，否则先令 $d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)=\infty$，通过 Dijkstra 算法计算测地距离为

$$d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)=\min\{d_G(\boldsymbol{x}_i,\boldsymbol{x}_j),d_G(\boldsymbol{x}_i,\boldsymbol{x}_l)+d_G(\boldsymbol{x}_l,\boldsymbol{x}_j)\},\quad l=1,2,\cdots,m \tag{32}$$

上式通过两点之间的最短路径来逼近测地距离，两点之间的测地距离可组成矩阵 $\boldsymbol{D}_G$.

③ 计算 k 维矩阵

将 MDS 算法应用于计算测地距离后的样本中. 把两点之间的测地距离看做两点之间的欧氏距离，将测地距离矩阵 $\boldsymbol{D}_G$ 代入 MDS 算法中，替换掉欧式距离矩阵 $\boldsymbol{\Delta}$，可计算得到 k 维降维后的样本.

Isomap 算法对非线性的高维数据具有较好的处理能力，将相距较近的点之间的测地距离用欧式距离表示，相距很远的点间的测地距离采用最短路径逼近，并用最大限度的保留降维后的数据样本间的全局测地距离信息，该方法在保证误差最小的同时，可实现情感语音特征数据的降维，因此该方法常应用于特征数据的非线性降维处理.

(5) 局部线性嵌入法(LLE)

LLE 是从局部的角度构建数据间的关系，其能够突破 PCA 降维对非线性数据的局限性，较好的表达情感语音特征数据内部的流形结构，保留本质特征，并且参数的优化也较为简单. 它的基本思想是每个数据点和与它邻近的 k 个点落在一个局部线性化的区域内，用这 k 个点逼近该点会得到 k 个权重系数，用这组系数可以刻画图形局部的几何性质，接着可以建立一个权值矩阵. 对于 m 个 d 维样本 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$，其实施步骤为

① 选取 k 个领域样本点

计算两样本点之间的欧氏距离 d_{ij}，选取前 k 个与 $\boldsymbol{x}_i$ 距离相近的点 $\boldsymbol{x}_{ij}$，$j=1,2,\cdots,k$.

② 计算权值矩阵

对于样本 $\boldsymbol{x}_i$，用其 k 个邻近点线性近似，定义误差评价函数

$$\min \varepsilon(\boldsymbol{W})=\sum_{i=1}^{m}\left|\boldsymbol{x}_i-\sum_{j=1}^{k}\omega_{ij}\boldsymbol{x}_{ij}\right|^2 \tag{33}$$

其中，ω_{ij} 为 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{ij}$ 之间的权值. 令 $\sum_{j=1}^{k}\omega_{ij}=1$，式(33)可变换为

$$\begin{aligned}\min \varepsilon(\boldsymbol{W})&=\sum_{i=1}^{m}\left|\sum_{j=1}^{k}\omega_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_{ij})\right|^2\\&=\sum_{i=1}^{m}|\boldsymbol{\omega}_i(\boldsymbol{x}_i-\boldsymbol{x}_{ij})|^2\\&=\sum_{i=1}^{m}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{Z}_i\boldsymbol{\omega}_i\end{aligned} \tag{34}$$

其中，$\boldsymbol{Z}_i=(\boldsymbol{x}_i-\boldsymbol{x}_{ij})^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_{ij})$ 是第 i 个样本点的局部协方差矩阵，$\boldsymbol{\omega}_i=\{\omega_{i1},\omega_{i2},\cdots,\omega_{ik}\}^{\mathrm{T}}$ 是第 i 个样本点的权值向量. 构造该评价函数的拉格朗日函数为

$$L(\boldsymbol{W})=\sum_{i=1}^{m}\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{Z}_i\boldsymbol{\omega}_i+\lambda\left(\sum_{j=1}^{k}\omega_{ij}-1\right) \tag{35}$$

上式对 ω_{ij} 求偏导为

$$2\boldsymbol{Z}_i\boldsymbol{\omega}_i+\lambda=0 \tag{36}$$

令 $\boldsymbol{Z}_i\boldsymbol{\omega}_i=1$ 结合 $\sum_{j=1}^{k}\omega_{ij}=1$ 可计算出 $\boldsymbol{\omega}_i$.

③ 求取低维特征 $\boldsymbol{Y}$

降维后的向量为 $\{\boldsymbol{y}_1,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_m\}$，维度为 p. 降维后的特征与降维前的样本的局部特征一致，降维后的特征评价函数为

$$\min \varepsilon(\boldsymbol{Y})=\sum_{i=1}^{m}\left|\boldsymbol{y}_i-\sum_{j=1}^{n}\omega_{ij}\boldsymbol{y}_{ij}\right|^2 \tag{37}$$

其中，$\boldsymbol{y}_i$ 是 $\boldsymbol{x}_i$ 降维后的向量，$\boldsymbol{y}_{ij}$ $(j=1,2,\cdots,k)$ 是 $\boldsymbol{y}_i$ 的 k 个近邻点，$\boldsymbol{y}_i$ 满足

$$\begin{cases}\sum_{i=1}^{m}\boldsymbol{y}_i=\boldsymbol{0}\\\dfrac{1}{m}\sum_{i=1}^{m}\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{I}\end{cases} \tag{38}$$

权值 ω_{ij} 可以存储在 $m\times m$ 的稀疏矩阵 $\boldsymbol{W}$ 中，当 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的近邻点时，$W_{ij}=\omega_{ij}$，否则，$W_{ij}=0$，记 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_m\}$，式(37)可改写为

$$\begin{aligned}\varepsilon(\boldsymbol{Y})&=\sum_{i=1}^{m}|\boldsymbol{Y}(\boldsymbol{I}_i-\boldsymbol{W}_i)|^2\\&=\sum_{i=1}^{m}\boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}\end{aligned} \tag{39}$$

其中，$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}$，构造拉格朗日函数

$$L(\boldsymbol{Y})=\boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}+\lambda(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}-m\boldsymbol{I}) \tag{40}$$

$L(\boldsymbol{Y})$ 对 $\boldsymbol{Y}$ 求导得

$$\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}=\lambda\boldsymbol{Y}^{\mathrm{T}} \tag{41}$$

计算矩阵 $\boldsymbol{M}$ 的特征值和特征向量，对其特征值进行从小到大的排序，取前 p 个非零特征值对应的特征向量即得到了降维后的样本. LLE 将全局非线性转化为局部线性，相互重叠的局部领域能够提供全局结构的信息.

文献[105]指出，LLE 作为最具代表性的流形数据降维方法之一，在语音情感识别上的效果相对于经典的 PCA 分析法效果欠佳，因此在原始的 LLE 的基础上提出了一种监督式的局部线性嵌入方法 SLLE，其通过修改 LLE 第一步的欧式距离，使得属于不同类别的两个点间的距离远远大于属于同一类别两点间的距离. 结果证明，SLLE 仅需要 11 个嵌入式特征，便能获取高于 80% 的语音情感识别准确率，而 LLE 方法需要 19 个情感语音特征，识别准确率远低于 SLLE，PCA 与 Isomap 方法均降维得到 12 维情感语音特征，识别准确率略高于 LLE.

各个降维算法的原理和结构的差异带来了算法之间运算复杂度的差异，而运算复杂度的差异决定了算法的运算效率，通常用时间复杂度来衡量算法的运算效率. 时间复杂度与样本点的个数 n、原始数据维度 D 以及算法所选临近点的个数 k 有关. 表 5 给出了各个算法的时间复杂度[106]. Isomap、MDS、LLE 的时间复杂度较高，PCA 和 LDA 的时间复杂度较低.

表 5　各类降维算法的优缺点

降维算法	时间复杂度	时间/s
PCA	$O(nD)+O(D^3)$	0.017
LDA	$O(nD)+O(D^3)$	0.015
MDS	$O(n^3)$	0.016
Isomap	$O(Dn\log n)+O(nk+n\log n)+O(n^3)$	0.164
LLE	$O(Dn\log n)+O(nk^3)+O(pn^2)$	2.213

为了验证各类算法的时间复杂度，选取了 CASIA 语音情感数据库进行实验，该数据库录制了 4 位实验者(2 男 2 女)的 6 种情感语音(高兴、悲哀、生气、惊吓、难过、中性)，选取了其中 300 句情感语音，采用 openSMILE 提取了基频、共振峰、短时能量等以及它们的衍生参数共 384 维特征. 当 PCA 的降维后维度选择指标达到 85% 时，降维后的维度为 62 维，以此为基准，将所有降维算法降维后的维度设置为 62 维，得到各算法所消耗的时间如表 5 所示. 表中与理论分析基本一致，Isomap 和 LLE 这类非线性

降维算法远比 PCA 和 LDA 这类线性降维算法消耗的时间多. MDS 算法虽然是非线性降维算法,但由于样本数量不大,所花费的时间与 PCA 相当.

表 6 给出了这 5 种降维算法的优缺点. 线性降维方法的时间复杂度较低,并且理论完善,概念简单,计算方便,但是也存在缺陷,如 PCA 对于主成分个数的选择没有明确的准则;LDA 降维后特征的最大维度为 $C-1$,C 为样本的类别数,当样本不满足高斯分布时,LDA 投影后不能把数据区分开来;MDS 较好的保留了数据的结构特性,但其计算量较大. 非线性降维方法复杂度较高,但其特殊的处理方法适用于非线性数据,Isomap 以流形上测地距离代替欧氏距离,可以更好地保留数据的几何结构,但它有结构不确定性[107],短环路也会严重影响其效果[108];LLE 对数据平移和旋转能够保持结构不变性,对于短环路的情况比 Isomap 改善了很多,但其要求样本采样率要高,所学习的流形是不闭合的而且是局部线性的.

表 6 各类降维算法的优缺点

算法	优点	缺点
PCA	概念简单、计算方便	降维后维度的选择没有明确的规则
LDA	有监督降维方法	降维后最大维度为 $C-1$,C 为样本类别数,对于不满足高斯分布的样本不适用
MDS	较好地保留了数据的内部结构	计算量较大
Isomap	用测地距离代替欧氏距离,较好地保留了流行数据的几何结构	计算量大,具有拓扑不稳定性,受短环路的影响
LLE	参数少,具有平移、旋转不变性,保留了数据的内在结构	需要数据样本是稠密而且是局部线性的,k 和 d 选择会影响降维结果,且对噪声敏感

对于语音情感特征数据的选择与降维方法的选取问题,不同降维方法有不同的视角,也会获得不同的识别效果,因而对数据与分类方法进行全面的分析与实验尤为重要.

5 总 结

语音情感识别是目前人机交互领域的研究热点,而与语音情感识别相关的研究过程是一个较为复杂的过程. 从语音情感特征参数的提取、特征的选择与降维分析,到最终的语音情感识别,每一步都至关重要. 其中,语音情感特征的提取作为语音情感识别整个过程的开始阶段,占据着十分重要的地位,其提取特征的准确与否将决定着语音情感识别的最终效果. 而特征的选择与降维分析大大减少了输入到系统中变量的维度,有助于提高系统的实时性.

本文对当今语音情感识别中特征提取及降维领域的研究进展进行了综述,对语音情感特征以个性化与非个性化的方式分类,并详细地对特征提取方法进行了分析与总结. 简要明了地叙述了目前存在的主流情感特征提取方法,着重介绍了基于导数的非个性语音情感特征与基于 Teager 能量算子(TEO)非线性特征提取方法. 从线性降维和非线性降维两个方面介绍了 4 种主要的降维方法,并对这些方法的优缺点做了对比.

通过对语音特征提取和降维领域的综述,发现语音情感识别已经取得了一些成就. 但情感识别是一个跨生理学、心理学、认知科学等领域的学科,目前还有很多问题值得研究,这也是语音情感识别未来的发展趋势,概括如下:

(1) 目前关于语音声学特征和语音情感之间缺乏相关性分析的研究. 很多研究者从各自的角度做了实验表明了某些特征与情感相关联,但不具有普遍性,研究者们也没有从机理上深入分析影响语音情感的特征.

(2) 基频、共振峰等部分语音声学特征的提取较为困难,目前缺乏一种精确的且适用于各种场合的特征提取方法. 传统的方法不能完美地解决复杂情况下声学特征的提取,新的方法能够较好的解决这一问题,但仍存在着不足,新的鲁棒性更好的方法亟待被提出.

(3) 不同研究者采用的语音情感数据库也不尽相同,这些数据库存在着语言、情感种类以及采集人数等差异,研究者们的语音情感识别率也不尽相同,缺乏统一的标准.

(4) 结合语义信息的情感识别. 语音中声学信息表达了大部分情感,但语义信息也能传递说话人的信息. 结合语义信息,在更高的层面上把握说话人的情感是一个重要的研究方向.

(5) 基于非个性化特征的语音情感识别. 非个性化特征在语音情感识别领域占据着越来越重要的地位,提取非个性化语音特征将有助于识别不同说话人的情感,非个性化特征的提取将成为语音情感识别领域的研究热点.

语音情感识别在安全驾驶、远程教育、健康医疗以及电子商务方面有着广泛的应用. 相信随着语音情感识别技术的快速发展,研究人员有所突破,探索更加高效、精确的语音情感特征提取方法,从而提取

出更加丰富、多元化的语音情感特征，为语音情感识别技术提供有力的支持.

参 考 文 献

[1] Müller V C. Fundamental Issues of Artificial Intelligence. Cham, Switzerland: Springer, 2016

[2] Zeng Yi, Liu Cheng-Lin, Tan Tie-Niu. Retrospect and outlook of brain-inspired intelligence research. Chinese Journal of Computers, 2016, 38(1): 212-222(in Chinese)
(曾毅，刘成林，谭铁牛. 类脑智能研究的回顾与展望. 计算机学报, 2016, 38(1): 212-222)

[3] Schutte N S, Malouff J M, Thorsteinsson E B. Increasing emotional intelligence through training: Current status and future directions. International Journal of Emotional Education, 2013, 5(1): 56-72

[4] Kwak S S, Kim Y, Kim E, et al. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot// Proceedings of the 2013 IEEE RO-MAN. Gyeongju, Korea, 2013: 180-185

[5] Picard R W. Affective Computing. Cambridge, UK: MIT Press, 1997

[6] Dai W, Han D, Dai Y, et al. Emotion recognition and affective computing on vocal social media. Information & Management, 2015, 52(7): 777-788

[7] Cambria E. Affective computing and sentiment analysis. IEEE Intelligent Systems, 2016, 31(2): 102-107

[8] Fu L Q, Mao X, Chen L J. Speaker independent emotion recognition based on SVM/HMMS fusion system//Proceedings of the Audio, Language and Image Processing. Shanghai, China, 2008: 61-65

[9] Lee C M, Narayanan S S. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech & Audio Processing, 2005, 13(2): 293-303

[10] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture// Proceedings of the International Conference on Acoustics. Montreal & Quebec, Canada, 2004: 577-580

[11] Ververidis D, Kotropoulos C. Speech communication: Emotional speech recognition: Resources, features, and methods. Speech Communication, 2006, 4(3): 1162-1181

[12] Bertenstam J, Granström B, Gustafson K, et al. The VAESS communicator: A portable communication aid with new voice types and emotion//Proceedings of the Swedish Phonetics Conference. Umea, Sweden, 1997: 57-60

[13] Gert R, Hans S. Handbook of Communication Competence. Germany Walter de Gruyter, 2008

[14] Wang K, An N, Li B N, et al. Speech emotion recognition using Fourier parameters. IEEE Transactions on Affective Computing, 2015, 6(1): 69-75

[15] Han J, Ji X, Hu X, et al. Arousal recognition using audio-visual features and FMRI-based brain response. IEEE Transactions on Affective Computing, 2015, 6(4): 337-347

[16] Kumar S S, RangaBabu T. Emotion and gender recognition of speech signals using SVM. International Journal of Engineering Science and Innovative Technology, 2015, 4(3): 128-137

[17] Rao K S, Koolagudi S G, Vempada R R. Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology, 2013, 16(2): 143-160

[18] Gharavian D, Sheikhan M, Nazerieh A, et al. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computing and Applications, 2012, 21(8): 2115-2126

[19] Li X, Li X, Zheng X, et al. EMD-TEO based speech emotion recognition. Lecture Notes in Computer Science, 2010, 6329: 180-189

[20] Devi J S, Yarramalle S, Nandyala S P. Speaker emotion recognition based on speech features and classification techniques. International Journal of Image, Graphics and Signal Processing, 2014, 6(7): 61-77

[21] Picard R W. Toward computers that recognize and respond to user emotion. IBM Technical Journal, 2000, 38(2): 705-719

[22] Welling L, Ney H. Formant estimation for speech recognition. IEEE Transactions on Speech and Audio Processing, 1998, 6(1): 36-48

[23] Zhou C, Hansen J, Kaiser J F. Nonlinear feature based and classification of speech under stress. IEEE Transactions on Speech and Audio Processing, 2001, 9(2): 201-216

[24] Kim J B, Park J S. Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. Engineering Applications of Artificial Intelligence, 2016, 52: 126-134

[25] Kyung H H, Eun H K, Yoon K K. Emotional feature extraction method based on the concentration of phoneme influence for human—robot interaction. Advanced Robotics, 2010, 24(1): 47-67

[26] Aher P, Cheeran A. Auditory processing of speech signals for speech emotion recognition. International Journal of Advanced Research in Computer and Communication Engineering, 2014, 3(5): 6790-6793

[27] Fewzee P, Karray F. Dimensionality reduction for emotional speech recognition//Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom). Amsterdam, Netherlands, 2012: 532-537

[28] Zeng Z H, Pantic M, Roisman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expression. Pattern Analysis and Machine Intelligence, 2009, 31(1): 39-58

[29] Rao K S, Koolagudi S G. Robust Emotion Recognition Using Spectral and Prosodic Features. New York, USA: Springer, 2013

[30] Shriberg E, Ferrer L, Kajarekar S, et al. Modeling prosodic feature sequences for speaker recognition. Speech Communication, 2005, 46(3): 455-472

[31] Han Wen-Jing, Li Hai-Feng, Ruan Hua-Bin, et al. Review on speech emotion recognition. Journal of Software, 2013, 25(1): 37-50(in Chinese)
(韩文静, 李海峰, 阮华斌等. 语音情感识别研究进展综述. 软件学报, 2013, 25(1): 37-50)

[32] Schroder M, Cowie R. Issues in emotion-oriented computing toward a shared understanding//Proceedings of the Workshop on Emotion and Computing. Bremen, Germany, 2006: 1-4

[33] Iliou T, Anagnostopoulos C N. Statistical evaluation of speech features for emotion recognition//Proceedings of the 4th International Conference on Digital Telecommunications. Colmar, France, 2009: 121-126

[34] Luengo I, Navas E, Hernáez I, et al. Automatic emotion recognition using prosodic parameters//Proceedings of the European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005: 493-496

[35] Benesty J, Sondhi M M, Huang Y. Springer Handbook of Speech Processing. Berlin, Germany: Springer, 2008

[36] Tsai S M. A robust zero-watermarking algorithm for audio based on LPCC//Proceedings of the International Conference on Orange Technologies(ICOT). Tainan, China, 2013: 63-66

[37] Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC//Proceedings of the International Conference on Computer Graphics, Simulation and Modeling. Pattaya, Thailand, 2012: 135-138

[38] Liu Z T, Li K, Li D Y, et al. Emotional feature selection of speaker-independent speech based on correlation analysis and Fisher//Proceedings of the 34th Chinese Control Conference. Hangzhou, China, 2015: 3780-3784

[39] Jin Xue-Cheng. A Study on Recognition of Emotion in Speech [M. S. dissertation]. University of Science and Technology of China, Hefei, 2007(in Chinese)
(金学成. 基于语音信号的情感识别[硕士学位论文]. 中国科学技术大学, 合肥, 2007)

[40] Kim E H, Hyun K H, Kim S H, et al. Improved emotion recognition with a novel speaker-independent feature. IEEE/ASME Transactions on Mechatronics, 2009, 14(3): 317-325

[41] Anagnostopoulos C N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. Artificial Intelligence Review, 2015, 43(2): 155-177

[42] Mao Q, Zhao X, Zhan Y. Extraction and analysis for non-personalized emotion features of speech. Advances in Information Sciences & Service Sciences, 2011, 3(10): 255-263

[43] Zhan Yong-Zhao, Mao Qi-Rong, Lin Qing, et al. Emotion Recognition of Vision Speech. Beijing: Science Press, 2013 (in Chinese)
(詹永照, 毛启容, 林庆等. 视觉语音情感识别. 北京: 科学出版社, 2013)

[44] Sherif M, Steven J, Lin X F, et al. Recognition of emotions in interactive voice response systems//Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland, 2010: 729-732

[45] Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech. Speech Communication, 2003, 40(1): 5-32

[46] Schafer R W, Rabiner L R. System for automatic formant analysis of voiced speech. Journal of the Acoustical Society of America, 1970, 47(2): 634-648

[47] Esposito A, Gennaro A, Jordi T, et al. Recent Advances in Nonlinear Speech Processing. Berlin, Germany: Springer, 2016

[48] Fernandez R, Picard R W. Modeling drivers' speech under stress. Speech Communication, 2003, 40(1): 145-159

[49] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home, 2012, 6(2): 101-108

[50] Lee C, Lui S, So C. Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. Journal of the Acoustical Society of America, 2014, 135(4): 25-34

[51] Ali S A, Khan A, Bashir N. Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus. International Journal of Information Technology and Computer Science, 2015, 7(2): 54-59

[52] Chen Pan-Di, Huang Hua, He Ling. Improved algorithm for pitch detection based on ACF and CEP. Computer Applications and Software, 2015, 32(1): 163-166(in Chinese)
(陈盼弟, 黄华, 何凌. 基于自相关和倒谱法的基音检测改进算法. 计算机应用与软件, 2015, 32(1): 163-166)

[53] Zeng Y M, Wu Z Y, Liu H B, et al. Modified AMDF pitch detection algorithm//Proceedings of the International Conference on Machine Learning and Cybernetics. Xi'an, China, 2003: 470-473

[54] Kadambe S, Boudreaux-Bartels G F. Application of the wavelet transform for pitch detection of speech signals. IEEE Transactions on Information Theory, 1992, 38(2): 917-924

[55] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Hong Kong, China, 2003: 401-404

[56] Schuller B, Arsic D, Wallhoff F, et al. Emotion recognition in the noise applying large acoustic feature sets//Proceedings of the Speech Prosody. Dresden, Germany, 2006: 276-289

[57] Kang G, Guo S. Improving AMDF for pitch period detection //Proceedings of the International Conference on Electronic Measurement & Instruments. Qingdao, China, 2009: 283-286

[58] He L, Lech M, Maddage N C, et al. Time-frequency feature extraction from spectrograms and wavelet packets with application to automatic stress and emotion classification in speech//Proceedings of the International Conference on

Information, Communications and Signal Processing. Macau, China, 2009: 615-622

[59] Yang Y, Zhang H, Guo X. A pitch tracking method mixing ACF & AMDF algorithms based on correlations//Proceedings of the & International Conference on Image Analysis and Signal Processing. Wuhan, China, 2011: 553-556

[60] Wang Q, Zhao X, Xu J. Pitch detection algorithm based on normalized correlation function and central bias function//Proceedings of the 2015 10th International Conference on Communications and Networking in China (ChinaCom). Chongqing, China, 2015: 617-620

[61] Hu H T, Hsu L Y. Robust glottal closure instant detection by jointly exploiting stationary wavelet transform and harmonic superposition. International Journal of Speech Technology, 2015, 18(4): 685-695

[62] Vlasenko B, Prylipko D, Philippou-Hübner D, et al. Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions//Proceedings of the 12th Annual Conference of the International Speech Communication Association. Florence, Italy, 2011: 1577-1580

[63] Zhang Li-Hua, Yang Ying-Chun. Change of emotional speech feature analysis. Journal of Tsinghua University: Natural Science Edition, 2008, 48(S1): 652-657(in Chinese)
(**张立华，杨莹春. 情感语音变化规律的特征分析. 清华大学学报:自然科学版**, 2008, 48(S1): 652-657)

[64] Montero-Martínez J M, Gutiérrez-Arriola J M, Pasamontes J C, et al. Development of an emotional speech synthesizer in Spanish//Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest, Hungary, 1999: 2099-2102

[65] Murray I R, Arnott J L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication, 1995, 16(4): 369-390

[66] Titze I R, Baken R J, Bozeman K W, et al. Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. The Journal of the Acoustical Society of America, 2015, 137(5): 3005-3007

[67] Snell R C, Milinazzo F. Formant location from LPC analysis data. IEEE Transactions on Speech and Audio Processing, 1993, 1(2): 129-134

[68] Hunt A, Howard D, Worsdall J. Real-time interfaces for speech and singing//Proceedings of the Euromicro Conference. Maastricht, Netherlands, 2000: 356-361

[69] Gharavian D, Sheikhan M, Ashoftedel F. Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model. Neural Computing and Applications, 2013, 22(6): 1181-1191

[70] Bozkurt E, Erzin E, Erdem Ç E, et al. Formant position based weighted spectral features for emotion recognition. Speech Communication, 2011, 53(9-10): 1186-1197

[71] Zhao Li. Speech Signal Processing. Beijing: China Machine Press, 2009(in Chinese)
(**赵力. 语音信号处理. 北京：机械工业出版社**, 2009)

[72] Messaoud Z B, Hamida A B. Combining formant frequency based on variable order LPC coding with acoustic features for TIMIT phone recognition. International Journal of Speech Technology, 2011, 14(4): 393-403

[73] Deng L, O'Shaughnessy D. Speech Processing: A Dynamic and Optimization-Oriented Approach. Boca Raton, America: CRC Press, 2003

[74] Mohamed M, Lee C C, Ahmad I L. Feature extraction of speech signal and heartbeat detection in angry emotion identification. International Journal of Computer Science and Electronics Engineering, 2013, 1(1): 101-105

[75] Sathepathak B, Panat A R. Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person. International Journal of Computer Science Issues, 2012, 9(4): 296-299

[76] Gowda D, Pohjalainen J, Kurimo M, et al. Robust formant detection using group delay function and stabilized weighted linear prediction//Proceedings of the Interspeech. Lyon, France, 2013: 25-29

[77] Smit T, Turckheim F, Mores R. Fast and robust formant detection from LP data. Speech Communication, 2012, 54(7): 893-902

[78] Logan B. Mel frequency cepstral coefficients for music modeling //Proceedings of the International Symposium on Music Information Retrieval. Massachusetts, USA, 2000: 1-11

[79] Zhou Ping, Li Xiao-Pan, Li Jie, et al. Speech emotion recognition based on mixed MFCC characteristic parameter. Computer Measurement Control, 2013, 23(1): 215-227(in Chinese)
(**周萍，李晓盼，李杰等. 混合 MFCC 特征参数应用于语音情感识别. 计算机测量与控制**, 2013, 23(1): 215-227)

[80] Nalini N J, Palanivel S, Balasubramanian M. Speech emotion recognition using residual phase and MFCC features. International Journal of Engineering and Technology, 2013, 5(6): 4515-4527

[81] Nwe T L, Foo S W, Silva L C D. Speech emotion recognition using hidden Markov models. Speech Communication, 2003, 41(4): 603-623

[82] Noll A M. Cepstrum pitch determination. Journal of the Acoustical Society of America, 1967, 41(2): 293-309

[83] Chen G H, Liu J H, Ye J. An improved method of endpoints detection based on energy-frequency-value//Proceedings of the Conference on High Density Microsystem Design and Packaging and Component Failure Analysis. Shanghai, China, 2006: 9-11

[84] Teager H M, Teager S M. Evidence for nonlinear sound production mechanisms in the vocal tract. Speech Production and Speech Modelling, 1990, 55: 241-261

[85] Zhang D X, Wu X P, Guo X J, et al. Endpoint detection of speech signal based on empirical mode decomposition and Teager kurtosis. Chinese Journal of Scientific Instrument, 2010, 146(7): 493-499

[86] Gao H, Chen S G, Su G C. Emotion classification of mandarin speech based on TEO nonlinear features//Proceedings of the International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD). Qingdao, China, 2007: 394-398

[87] Ma Yong-Ling, Han Ji-Qing, Zhang Lei, et al. TEO-Pitch based classification of stressed speech under G-Force. Acta Acustica, 2002, 27(6): 518-522(in Chinese)
(马永林, 韩纪庆, 张磊等. 基于 Teager 能量算子(TEO)基频的应力影响下的变异语音分类. 声学学报, 2002, 27(6): 519-522)

[88] Zhou G, Hansen J H L, Kaiser J F. Classification of speech under stress based on features derived from the nonlinear Teager energy operator//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Washington, USA, 1998: 549-552

[89] Liu Ji-Ming. Noise Speech Emotion Recognition Based on TEO[M. S. dissertation]. Shanghai University, Shanghai, 2009(in Chinese)
(刘记明. 基于 TEO 特征的抗噪语音情感识别[硕士学位论文]. 上海大学, 上海, 2009)

[90] Gao Hui, Su Guang-Chuan, Chen Shan-Guang. Emotional recognition of mandarin speech using nonlinear features based on Teager energy operator (TEO). Space Medicine Medical Engineering, 2005, 18(6): 427-431(in Chinese)
(高慧, 苏广川, 陈善广. 基于 Teager 能量算子(TEO)非线性特征的语音情绪识别. 航天医学与医学工程, 2005, 18(6): 427-431)

[91] Li Jin-Hui, Yang Jun-An, Wang Yi. New feature extraction method based on bottleneck deep belief network and its application in language recognition. Computer Science, 2014, 41(3): 263-266(in Chinese)
(李晋徽, 杨俊安, 王一. 一种新的基于瓶颈深度信念网络的特征提取方法及其在语种识别中的应用. 计算机科学, 2014, 41(3): 263-266)

[92] Mao Q, Dong M, Huang Z, et al. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213

[93] Dutta K, Sarma K K. Multiple feature extraction for RNN-based Assamese speech recognition for speech to text conversion application//Proceedings of the 2012 International Conference on Communications, Devices and Intelligent Systems. Ottawa, Canada, 2012: 600-603

[94] Felix W, Jürgen G, Martin W, et al. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. Computer Speech Language, 2014, 28(4): 888-902

[95] Zhang Bo, Shi Zhong-Zhi, Zhao Xiao-Fei, et al. A transfer learning based on canonical correlation analysis across different domains. Chinese Journal of Computers, 2015, 38(7): 1326-1336(in Chinese)
(张博, 史忠植, 赵晓非等. 一种基于跨领域典型相关性分析的迁移学习方法. 计算机学报, 2015, 38(7): 1326-1336)

[96] Busso C, Deng Z, Yildirim S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information//Proceedings of the 6th International Conference on Multimodal Interfaces. Pennsylvania, USA, 2004: 205-211

[97] Schuller B, Batliner A, Steidl S, et al. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication, 2011, 53(9): 1062-1087

[98] Song P, Zheng W, Liu J, et al. A novel speech emotion recognition method via transfer PCA and sparse coding//Proceedings of the 10th Chinese Conference on Biometric Recognition. Tianjin, China, 2015: 393-400

[99] Zhang L, Song M, Li N, et al. Feature selection for fast speech emotion recognition//Proceedings of the 17th ACM International Conference on Multimedia. Orlando, USA, 2009: 753-756

[100] Huang R S. Information technology in an improved supervised locally linear embedding for recognizing speech emotion. Advanced Materials Research, 2014, 1014: 375-378

[101] Qian Y, Ying L, Pingping J. Speech emotion recognition using supervised manifold learning based on all-class and pairwise-class feature extraction//Proceedings of the 7th IEEE International Conference on Intelligent Computing and Integrated Systems. Kharagpur, India, 2011: 1-5

[102] Wang S, Ling X, Zhang F, et al. Speech emotion recognition based on principal component analysis and back propagation neural network//Proceedings of the Measuring International Conference on Technology and Mechatronics Automation. Changsha, China, 2010: 437-440

[103] Zhang X, Cheng Z, Xu X, et al. Speech emotion recognition based on LDA + kernel-KNNFLC. Journal of Southeast University (Natural Science Edition), 2015, 45(1): 5-11

[104] Zhang S, Zhao X, Lei B. Speech emotion recognition using an enhanced kernel Isomap for human-robot interaction. International Journal of Advanced Robotic Systems, 2013, 10(2): 323-330

[105] Zhang S, Li L, Zhao Z. Speech emotion recognition based on supervised locally linear embedding//Proceedings of the International Conference on Communications, Circuits and Systems. Chengdu, China, 2010: 401-404

[106] Wu Xiao-Ting, Yan De-Qin. Data dimension reduction methods and research. Application Research of Computers. 2009, 26(8): 2832-2835(in Chinese)
(吴晓婷, 闫德勤. 数据降维方法分析与研究. 计算机应用研究, 2009, 26(8): 2832-2835)

[107] Tenenbaum J B, Silva V D, Langford J C. The Isomap algorithm and topological stability-Response. Science, 2002, 295(5552): 7-7

[108] Lee J A, Verleysen M. Nonlinear dimensionality reduction of data manifolds with essential loops. Neurocomputing, 2005, 67(1): 29-53

LIU Zhen-Tao, born in 1981, Ph.D., lecturer. His main research interests include computational intelligence and affective computing.

XU Jian-Ping, born in 1993, M. S. candidate. His main research interests include speech emotion recognition and human-robot interaction.

WU Min, born in 1963. Ph. D., professor. His main research interests includes process control, robust control, and intelligent system.

CAO Wei-Hua, born in 1972, Ph. D., professor. His main research interests include process control, intelligent control and multi-agent control.

CHEN Lue-Feng, born in 1986, Ph. D., associate professor. His main research interests include human-robot interaction, computational intelligence and intention understanding.

DING Xue-Wen, born in 1995, M. S. candidate. His main research interests include facial expression recognition and human-robot interaction.

HAO Man, born in 1994, Ph. D. candidate. Her main research interests include emotional robot and human-robot interaction.

XIE Qiao, born in 1992, M. S. candidate. His main research interests include emotion modeling and EEG emotion recognition.

Background

With the development of computer science and human-robot interaction (HRI), new demands that robots can recognize and generate emotions as humans are produced. Speech emotion recognition (SER) is important for HRI, since speech is a commonly used way to communication, and it has broad application prospects in many areas such as service robot, intelligent driving, and distance education. Much progress has been achieved in SER in the last ten years. However, there are still many challenges and difficulties, for example, universal definition of the speech emotion has not been made, and most of recent researches focus on speaker-dependent speech emotion recognition without studying on how to extract emotional features from speaker-independent speech.

In this paper, we try to classify emotional features in a new way that is different from the traditional classification. We classify them into personalized features and non-personalized features. Fundamental frequency, formants, and Mel cepstral coefficient are three major emotional features, which are closely related to the emotional information in speech are described in details, and the exaction methods of them are introduced as well. To improve the SER accuracy for different people, we introduce a non-personalized speech feature extraction method based on the derivative. In addition, feature extraction methods based on Teager Energy Operator (TEO) as well as deep learning for speech emotion recognition are presented. What's more, some linear dimension reduction methods such as PCA and LDA, and nonlinear methods such as MDS, Isomap, and LLE are introduced. Finally, the challenges and opportunities in the field of speech emotion recognition are analyzed.

This work was supported by the National Nature Science Foundation of China (Grant No. 61403422 and No. 61603356), the Hubei Province Natural Science Foundation of China (Grant No. 2015CFA010), the Wuhan Science and Technology Project (Grant No. 2017010201010133), 111 Project (Grant B17040), and the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) (Grant No. 1610491T09).