计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400122

语音任务下声学特征提取综述

郑纯军[1,2]　王春立[1]　贾　宁[2]
1 大连海事大学信息科学技术学院　辽宁 大连 116023
2 大连东软信息学院计算机与软件学院　辽宁 大连 116023

摘　要　语音是一种重要的信息资源传递与交流方式,人们经常使用语音作为交流信息的媒介,在语音的声学信号中包含大量的说话者信息、语义信息和丰富的情感信息,因此形成了解决语音学任务的3个不同方向,即声纹识别(Speaker Recognition,SR)、语音识别(Auto Speech Recognition,ASR)和情感识别(Speech Emotion Recognition,SER),3个任务均在各自的领域使用不同的技术与特定的方法进行信息提取与模型设计。文中首先综述了3个任务在国内外早期的发展历史路线,将语音任务的发展归纳为4个不同阶段,同时总结了3个语音学任务在特征提取时所采用的公共语音学特征,并针对每类特征的侧重点进行了说明。然后,随着近年来深度学习技术在各个领域中的广泛应用,语音任务也得到了很好的发展,文中针对目前流行的深度学习模型在声学建模中的应用分别进行了分析,按照有监督、无监督的方式总结了针对3种不同语音任务的声学特征提取方式及技术路线,还总结了基于多通道并融合注意力机制的模型,用于语音的特征提取。为了同时完成语音识别、声纹识别和情感识别任务,针对声学信号的个性化特征提出了一个基于多任务的Tandem模型;此外,提出了一个多通道协作网络模型,利用这种设计思路可以提升多任务特征提取的准确度。

关键词:声学特征提取;声纹识别;语音识别;情感识别;深度学习;多通道融合
中图法分类号　TP183

Survey of Acoustic Feature Extraction in Speech Tasks

ZHENG Chun-jun[1,2],WANG Chun-li[1] and JIA Ning[2]
1 College of Information Science and Technology,Dalian Maritime University,Dalian,Liaoning 116023,China
2 School of Computer & Software,Dalian Neusoft University of Information,Dalian,Liaoning 116023,China

Abstract　Speech isan important means of information transmission and communication,people often use speech as a medium for exchanging information. The acoustic signal of speech contains a large amount of speaker information,semantic information and rich emotional information. Therefore,three different directions of speech tasks,speaker recognition (SR),auto speech recognition (ASR),and speech emotion recognition (SER),are formed. Each of the three tasks uses different techniques and specific methods for information extraction and model design in their respective fields. Firstly,the historical routes of three tasks at the early stage of development at home and abroad were summarized. The development of speech tasks was summarized into four different stages. At the same time,the public phonetics features for three speech tasks were summarized. The focus of each type of feature was explained. Then,with the wide application of deep learning technology in various fields in recent years,the speech task is well developed. The application of the current popular deep learning model in acoustic modeling was analyzed separately. The acoustic features extraction methods and technical routes for three different speech tasks were summarized in two ways,supervised and unsupervised. In addition,a multi-channel fusion model based on attention mechanism for feature extraction was proposed. In order to solve three speech tasks at the same time,a multi-task model based personalized was proposed for speech feature extraction. This paper also proposed a multi-channel cooperative network model. By using this design idea,the accuracy of multi-task feature extraction can be improved.

Keywords　Acoustic features extraction,Speaker recognition,Auto speech recognition,Speech emotion recognition,Deep learning,Multi-channel fusion

到稿日期:2019-04-22　返修日期:2019-08-10　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:辽宁省自然科学基金(20180551068)
This work was supported by Liaoning Natural Science Foundation (20180551068).
通信作者:郑纯军(zhengchunjun@neusoft.edu.cn)

1 引言

语音作为人类交流最方便、最自然的媒介，包含多种不同类型的信息，可以通过语音识别来获得语音中的语义文字信息并通过不同渠道来表达，再对其进行文本分析和理解。鉴于声道结构的多样性，每个人的语音数据都有其自身的特定信息，因此可以使用相关技术来有效地为说话者识别语音内容。此外，可以通过不同语音场景或语音事件来定义情绪，从而识别用户表达的丰富的情感信息[1]。

基于此，形成了语音学领域的3个重要研究方向：语音识别、声纹识别和情感识别。声纹识别也称为说话者识别，其在分析连续语音信号后提取离散语音特征，通过与模板进行匹配来自动确认该语音的说话者。声纹识别任务主要分为说话者确认和说话者身份识别，往往被应用于金融交易等高安全级别场所。语音识别[2]即通过语音发布指令，由计算机根据声学、语言模型及词典，利用特定的算法将其识别后转化成文字等操作命令，目前语音识别被广泛应用于声控、智能对话、医疗服务等行业。语音情感识别旨在通过语音信号来识别说话者的正确情绪状态，由于语音并非是情感生理信号的完整表达形式，在忽略其余感官结果的前提下，如何高效且精确地识别用户表达的情感，是近年来语音学研究的热点问题。

语音识别和声纹识别的研究均在20世纪60年代左右起步，历经半个多世纪的发展，已经从简易演变成繁琐，从科学研究转变为实际应用。而情感识别的研究起步于1997年，经过20余年的科学研究与技术探讨，研究人员仍在探索完美的情感替代模型。从语音学研究的技术角度进行考量，其发展主要分为以下几个阶段。

第一阶段为20世纪60年代至20世纪70年代。在此阶段，语音识别研究主要集中于小词汇量、孤立词的识别，而后开始大规模的语音识别；声纹识别在此阶段一直注重语音特征提取和模板匹配技术的发展。此时，语谱图、线性预测编码系数、自相关系数和倒谱系数等被广泛应用于特征提取，而模板匹配技术则采用动态时间规整和矢量量化的方法。

第二阶段为20世纪80年代至20世纪90年代，该阶段是语音识别和声纹识别成果丰收的时期。Davis提出了梅尔频率倒谱系统(Mel Frequency Cepstral Coefficents，MFCC)。MFCC作为语音学的核心传统特征，从人耳构造的机理出发，被广泛应用于这两个任务中。而模板匹配技术逐渐被隐马尔可夫模型(Hidden Markov Model，HMM)、高斯混合模型(Gaussian Mixture Model，GMM)和人工神经网络(Artificial Neural Network，ANN)模型等替代，此时模型规模在不断增大，表征力越来越强，其负面效应越发明显，需要大量的实际数据来训练通用的模型，从而提升其泛化性。基于此，在20世纪90年代后产生的高斯混合模型，以较高的灵活性和较强的鲁棒性获得了研究人员的青睐，成为了声纹识别的主流技术。而此时语音识别从音频特征单元和音素之间的关联角度出发，利用HMM构建声学模型和随机语言模型。

第三阶段为20世纪90年代末至2010年前后。此阶段仍采用统计学的方法来处理语音特征和优化模型，由于此时计算机硬件正飞速发展，语音识别和声纹识别的技术均由科研逐步转为实际应用。语音识别于2001年达到了约80%的识别率，然而在此阶段未有突破性的特征和模型促使语音识别的成功率有大幅度的提升。此时，声纹识别则基于高斯混合-通用背景模型(Gaussian Mixture Model-Universal Background Model，GMM-UBM)在个性化领域迅速发展，支持向量机(Support Vector Machine，SVM)、相关融合系统(如GMM-HMM，GMM-SVM，GSV-SVM)以及评分规整技术(如T-norm，Z-norm)开始被应用于声纹识别。在此基础上，为了进一步提高识别的鲁棒性，联合因子分析(Joint-Factor Analysis，JFA)[3]、本征信道分析(Eigen-Channel)[4]、本征音分析(Eigen-voice)[5]、扰动属性干扰算法[6]、i-vector等信道补偿技术被应用于声纹识别，其因具有灵活、简单、有效、鲁棒性较强的特点而成为了当时声纹识别研发的主流。同时，在此阶段提出了情感计算，标志着情感识别领域在语音学领域正式拉开了帷幕。由于声纹识别着重个性化声纹特征的提取和模型设计，而语音识别和情感识别任务主要针对通用化、非个性特征进行判断，因此语音识别和情感识别无法完全复用声纹识别的模型。此时，语音识别考虑了环境噪声、信道、用户口音等方面的影响，采用自适应技术有针对性地对系统进行调整，利用语言模型的概率统计方法来预测期望的音素。目前情感识别处于起步阶段，大多采用韵律特征和SVM模型来实现非线性的语音特征提取和情感预测。

此时的研究结果表明，在执行不同任务时，已经分别建立了跨领域的音频数据的模型，而且针对不同任务的音频数据的提取角度，其在各自领域有所侧重。受到这种双重效应的影响，语音识别、声纹识别和情感识别的系统性能分别在各自领域获得了不同程度的提升。

第四阶段起始于2010年前后，一直持续至今。在此期间，深度学习等人工智能方法的应用推动了语音学的发展。在起始阶段，声纹识别采用i-vector和PLDA(Parallel Latent Dirichlet Allocation)压缩声学特征，在融合深度学习之后，涌现出了d-vector，j-vector，s-vector，x-vector和Sequence2-Sequence等模型。文献[7]提出了一种深度信念网络(Deep Belief Network，DBN)和深度神经网络(Deep Neural Network，DNN)的通用自适应模型，当样本量较少时，利用背景i-vector构建全局模型通用DBN，为每个说话者提供平衡数据。Schmidt等[8]采用DBN从幅度谱中提取高级情感的特征表示，与传统声学特征相比，其表现出了更好的性能。

然而，随着深度学习模型的推广、其变换能力的不断加深、自我学习能力的不断增强，3种语音学任务均希望找到适用于自身性能需求的语音特征和相关模型。深度学习的灵活性使得上述3种语音学任务均可利用深度学习技术来完成性能的提升和功能的完善。

本文第2节和第3节分别描述了语音学任务使用的公共特征，以及在深度学习领域中每个任务所使用的相似模型之间的不同侧重点；第4节提出了一个适用于3种任务的特征提取综合模型，通过任务之间的相辅相成和相互促进，来提高特征提取的准确度和效率。

2 语音学公共特征提取

语音学特征的鲁棒性对分类和识别的效果有显著影响。

语音信号不是静止信号，在提取语音特征时，其对说话人和噪声等环境极其敏感，易受到环境的干扰，从而影响识别的性能。

由于在时域难以对语音的形状进行描述，因此一般采用频域分析的方法，将语音信号分解成多个单一频率的波形。目前流行的语音学特征提取方法主要有两类：1）采用传统特征，从原始音频文件中提取信号特征，捕获最原始的不同类型的声学特征，从而判定该特征所属的语音学任务类型；2）将传统特征与深度学习模型相融合，在交叉领域中突出特征的重点，由于不同任务的侧重点不同，融合的方式体现出了多样化、个性化的特点。

本节主要针对第 1 类语音学特征进行整理，即提取常用的公共特征。由于第 2 类语音学特征与深度学习联系紧密，本文将在第 3 节中进行详细介绍。

通过对传统特征进行长期研究发现，语音声学特征蕴藏在一些不同层次的特征中，如韵律、音质、谱特征等。下面将针对这些特征的特点和适用任务进行总结。

2.1 韵律特征和音质特征

表 1 列出了常见的语音特征，包括韵律特征和音质特征的特征信息。

表 1 常见的声学特征

Table 1 Common acousticfeatures

特征	特点	具体特征
韵律特征	动态表达	能量、共振峰、音高、时长、发音、基音频率、过零率
音质特征	固有属性	相位、频率微扰、声门参数

作为非个性化语音特征的重要组成，常见的韵律特征主要由能量、共振峰、音高、时长和发音等组成。

在语音中，大部分情感内容会影响连续的韵律特征。Wang 等[9]对不同情感的同一样本语句进行能量及共振峰分析，以确定此类特征对于高兴、愤怒、悲伤、平静等情感的区别度，其中能量特征在情感识别过程中的区分度较高，即效价高时声音振幅较大，其平均能量值较大，反之则能量值相对较低。

共振峰对许多形式的情绪和精神状态敏感，常被用于认知负荷分类和抑郁识别与评估，可以提高情感识别的表现力。一般认为共振峰值的趋势与情绪的效价有关。在语音识别方面，共振峰在提取过程中的误差信息和多个峰值之间的重叠均会对识别准确度产生影响，而且共振峰对于连续语音的识别准确度较差，在自然环境下其鲁棒性较差。

文献[10]指出基音频率（Pitch Frequency）在效价较高的情感中体现相似的特性，主要用于区分情感中的效价度。

文献[11]使用 OpenSMILE 工具包从音频信号中提取短时特征，用于识别语音中的副语言信息，以获得较高的情绪识别率。短时特征包括多个低级描述符：强度、响度、MFCC、音调、发声概率、F0、线谱频率和零交叉率等。文献[12]提出了一种基于音高轮廓的平滑样条，近似从语音中提取韵律特征的新方法，以提高语音识别的准确度。

音质特征主要体现为呼吸声、喉化音和明亮度等多种形式。目前，音质特征与情感联系的方式存在争议性，普遍认为感知的情感和语音质量的关系体现在语音级别、语音音调、时间和特征边界结构等方面。

频谱类特征基于频谱分析的特征被视为语音信号的短时表示。频谱特征主要由线性预测系数和倒谱分析等特征构成。常用的频谱特征有 MFCC、线性预测倒谱系数（Linear Prediction Cepstrum Coefficient，LPCC）和梅尔刻度滤波器组过滤（logMel）等。频谱特征的提取方法简单，并且具有少量维度。在频谱分析之前往往设置预处理环节，将语音信号分解成具有固定帧特点的样本，有助于独立地分析信号。在分帧时，每帧的大小与特征提取方法有关，基于所使用的特征提取方法来选择帧的大小。允许通过帧与帧之间的重叠来消除帧之间的差异产生的影响。

MFCC 作为经典的倒谱分析方法，根据人耳听觉系统的非线性响应原理对其特征参数进行设置。MFCC 的设计思想起源于人类心理声学的研究，可以感知不同的频带，其证明了人耳对语音信号的感知与其频率变化有很大的关联。MFCC 由于鲁棒性强、识别率高，目前被广泛用于自动语音和说话人识别。然而，MFCC 的缺陷也非常明显，由于语音是动态变化的，传统的 MFCC 未涉及相邻固定帧之间的关联和帧内部参数之间的关联。基于此，文献[13]提出了一种新的一阶差分和二阶差分的 MFCC 参数组。

在分帧和加窗的基础上，对每一帧信号做离散傅里叶变换，计算对数幅度频谱，然后将其输入等带宽的梅尔滤波器组进行滤波，通过离散余弦变换最终得到 MFCC 特征。

MFCC 的计算方法如式（1）所示：

$$MFCC(t,i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg[E_{mei}(t,j)]\cos[i(j-0.5)\frac{\pi}{N}] \tag{1}$$

其中，N 为滤波器的数量，$E_{mei}(t,j)$ 是第 t 个时刻第 j 个滤波器的输出。通过式（1）可获得第 t 个时刻的 MFCC 参数。

针对情感识别，语音话语的频谱能量分布取决于其情感内容。唤醒度在较高频率下具有较高的能量，但具有低觉醒情绪，如悲伤的语音在相似范围内具有较少的能量。然而，单独使用 MFCC 无法提取到与特征标签完全吻合的特征。因此，文献[14]将 Teager 能量算子（TEO）和梅尔频率倒谱系数融合形成新特征，并将其作为 Teager-MFCC（T-MFCC）特征提取技术，用于识别来自语音信号的情绪。

2.2 语谱图

语音频谱图是通过处理接收的时域信号而得到的频谱图，更确切地说是频谱分析视图。对原始信号进行分帧加窗后，对每一帧做快速傅里叶变换，把时域信号转为频域信号，频域信号在时间上堆叠后就可以得到频谱图。图 1 给出了语谱图的生成流程。

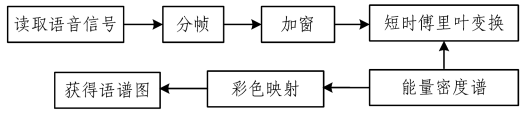

图 1 语谱图的生成流程

Fig. 1 Generation process of spectrogram

语谱图采用二维的图片形式来表达三维的坐标信息，为

了区分频域变换方式，常用的声谱图可分为短时傅里叶变换(STFT)声谱图和 CQT(constant-Q transform)声谱图，这两者对于噪声的鲁棒性较强。

针对声纹识别，文献[15]考查了连续语音中呼吸音的说话人识别潜力，利用 CQT 频谱图形成生物签名，结合卷积神经网络(Convolutional Neural Networks，CNN)和长短期记忆(Long Short Term Memory，LSTM)网络来判断声道的共振信息，从而高效地获得说话者的身份信息。

在语音识别中，梅尔倒谱表示对梅尔频谱进行一次频谱分析，MFCC 可减小识别误差[16]。文献[17]证明了梅尔倒谱可对说话人进行识别。文献[18-19]公布了现有技术的鲁棒语音识别系统，它是基于线性间隔光谱图实现的。因此，对于语音识别，消除部分音调信息的梅尔倒谱可以产生良好的效果。

Satt 等[20]将 CNN 与 LSTM 相结合，对线性间隔光谱图的情绪进行分类。谱图分为等长的段，或者以 0 填充到固定大小，以满足 CNN 的要求，此时语音连续性可能被分割破坏，可能导致系统难以捕捉整个情绪的变化过程。文献[21]应用频谱图有效地处理背景信息而非语音信号，使用谐波建模从频谱图中去除非语音的分量，从而使情绪识别的准确度得到显著改善。文献[22]提出了一种新颖的视觉词袋方法，用于表示声学事件的灰度谱图，常用于声学事件分类(Acoustic Event Classification，AEC)，它评估视觉特征直方图之间的卡方距离，对噪声更具鲁棒性，并实现了更高的识别准确度。

3 深度学习模型的特征提取在语音任务上的应用

深度学习方法可以从不同层次的输入中学习有效的语音信号的非线性表现形式，目前已经被广泛应用于声纹识别、语音识别和情感识别。目前常见的深度学习模型可以分为有监督和无监督两种，下面将针对解决不同语音特征提取任务的 DNN，CNN 等有监督模型，以及编码器等无监督模型进行介绍。为了突出不同任务的信号特征，本文分别介绍了瓶颈特征的提取和多通道注意力机制融合等方法的设计思路。使用深度学习模型在 3 种不同语音任务上进行特征提取的方法如图 2 所示。

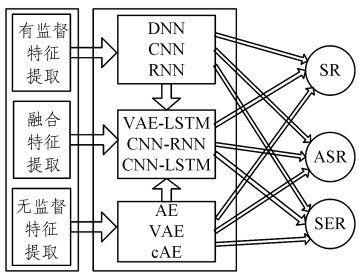

图 2 基于深度学习的语音特征提取方法

Fig. 2 Speech feature extraction method based on deep learning

3.1 DNN

2006 年，Hinton 利用预训练方法缓解了局部最优解问题。DNN 模型的参数较多，模型复杂，尺寸较大，训练时间较长，不能利用历史信息来辅助当前任务，但它能利用帧的上下文信息来学习深层非线性特征变换，在语音任务上得到了一定的应用。

在声纹识别中，DNN 的应用非常广泛。在国外，文献[3]提出使用 DNN 方法来研究有限语音数据在说话人验证中的影响，期望能够缩短所需语音数据的长度。针对短语音(30s 以内)，对说话者、会话、语音变化进行鲁棒建模，可使说话人的识别效果更佳。文献[3]介绍了两种说话者短语音的识别任务：基于音素后验 DNN 的 i-vector 系统和依赖于音素的基于子区域的 GMM-UBM 系统。针对混响导致的语音信号损坏的情形，文献[4]建立了多个 DNN 模型，每个 DNN 对应于设置步骤中的不同混响时间，通过对每个模型进行进一步联合训练来提升其在混响环境中的识别性能。在国内，针对中文特定的发音，文献[23]在训练时使用基于 i-vector 的距离度量，对在声学上彼此相近的说话者进行分层聚类。针对每个话者群，自适应地训练话者相关层的 DNN。该方法在话者簇数量适中时，对说话者的聚类匹配是有效的，并且充足的测试数据可用于提取可靠的 i-vector 特征，反之则准确度不高。针对更短时的语音，如咳嗽、笑和中文“喂”等短时语音，文献[24]设计了一个从大量原始数据中学习帧级说话者特征的模型 CT-DNN，其对帧级声纹特征进行计算得到 d-vector，经过 PLDA 评分发现 d-vector 系统明显优于 i-vector，利用此模型可以在非常短的语音段获得良好的识别准确度。在半文本独立的说话者验证任务中，文献[25]使用 DNN 模型来提取说话者的特征 d-vector，基于 DNN 的特征学习在与文本无关的任务上运行良好。然而，在半文本独立任务中，这种盲目学习往往是困难的，因为原始数据中涉及了太多与说话者无关的有效信息。

DNN 同样可以应用于语音识别，文献[26]提出了一种用于语音识别的结构化 HDNN(Highway Deep Neural Network)架构，能够训练更小的模型。它还设计了具有序列判别训练标准和话者自适应技术的无监督体系结构，使用交叉熵准则训练模型提高了语音识别的准确性。此外，绑定在所有隐藏层上的两个门能够控制整个网络上的信息流，在序列训练和适应性实验中更新这些门的功能可实现相当大的改进。

DNN 也被用于情感识别的研究，在国内，Han 等[27]使用具有最高能量的段来训练 DNN 模型，从而学习短期声学特征；使用传统的统计函数来构建话语级特征，以提取有效的情绪信息。DNN 使用个性化特征作为输入，因为个性化功能会受到各种说话风格、语音内容和环境的影响。一般而言，个性化特征携带大量的个人情感信息，能反映说话者的特征，并且不包含不同说话者的共同情感信息、内容和环境[28]。目前，在国内关于语音情感识别的大多数研究都是基于个性化的情感特征，并且已经获得了良好的识别性能，特别是对于特定的发言者。尽管改进的 SER 算法可用于获得更好的性能，但它阻碍了 SER 技术在独立于说话者的真实环境中的实际应用。因此，减小不同说话者和说话风格的个性化特征的数值差异，对于语音情感研究具有重要意义。文献[29]比较了时延网络模型 TDNN-Statistics Pooling，TDNN-LSTM，TDNN-LSTM-Attention，LSTM 和 LSTM-Attention 模型的性能，发现基于

注意力机制的 TDNN-LSTM 在语音情感识别方面可获得最好的结果。

3.2 CNN

CNN 由 Lecun 等提出，是第一个真正的深层结构学习算法，它通过卷积的权值共享及池化操作来降低网络参数的数量级。

针对语音学任务，CNN 通常采用的特征是低级特征表示或语谱图，它对噪声具有较强的鲁棒性[30]。此外，文献[31]表明，由于参数共享机制，CNN 更适用于小内存的关键字定位。

国外的研究主要针对大量的说话者信息，如文献[32]使用 CNN 来识别演讲中的笑声，将标记化的文本转换为具有形状的 2D 张量，通过对比发现，CNN 在面对新数据集时的学习效率比随机森林更高。Chen 等[33]发现二维卷积优于有限数据的一维卷积，时域卷积与频域卷积一样重要。如果 CNN 模型使用个性化特征作为输入，则不同说话者和各种说话风格的 SER 差异很大。

由于语音时频谱具有结构化的特点，因此在国内的语音识别领域，往往采用 CNN 来进行声学建模，这主要是因为 CNN 在时间和空间上的平移不变性，研究者希望利用这种特性来降低汉语语音信号本身的多样性。Huang 等[34]借鉴图像处理的方式，结合语音信号特征，建立了局部有限权重共享的卷积神经网络学习方式，提高了语音识别的效果与计算性能。

在情感识别中，国内的 Mao 等[28]使用 CNN 来学习 SVM，以寻找分类的显著特征。而国外的 Aldeneh 等[35]指出 CNN 也可直接应用于低级声学特征，以识别情绪显著区域，话语级别的统计函数由于可能会产生混淆信息，而鲜有定义或应用，其替代品可以是 Mel 滤波器组。文献[36]提出了卷积-池化(Conv-Pool)网络，用于识别可变长度话语内的情绪显著区域，提取出的特征具有较高的情感识别准确性。

针对 CNN 易产生过拟合的现象，文献[37]提出了用于语音情感识别的 Shake-Shake 正则化残差网络(ResNet)，其通过使多个流数据参与特征空间的不同部分来对抗噪声，证明了正则化残差网络可以改善未加权平均(Unweighted Accuracy，UA)并减少过拟合，有利于情感计算中的学习。

针对 CNN 中池化层信息容易丢失的问题，Aldeneh 等提出采用最大池化层，因为其对情绪识别效果最佳，而且存在平移不变性。

CNN 采用手工提取的特征，常忽略信号之间关联性的问题，而深度卷积神经网络(Deep Convolutional Neural Network，DCNN)可以弥补此不足。文献[38]提出了一种基于 DCNN 的有效情感识别系统来计算对数谱图，并且使用主成分分析(Principal Components Analysis，PCA)技术来降维并抑制干扰，从具有标记的训练语音片段中学习情绪信息。文献[39]为 SER 提出了深度视网膜卷积神经网络，用于从语谱图中学习高级特征，它在情绪识别准确性方面均优于以往的研究。

3.3 循环神经网络(Recurrent Neural Network，RNN)

CNN 网络模型融合的特征往往局限于语谱图或低级描述符，此类特征均忽视了语音信号的重要特性，即语音信号是具有时间序列的单元集合。RNN 作为语音学任务中流行的新型架构[40]，往往结合频谱类特征和韵律学特征，同时在其中添加与时间结点有关的自我连接形式，增强了其对时间序列的建模能力。

在自我循环的过程中，传统的 RNN 易产生梯度消失或梯度爆炸问题，而在其基础上提出的 LSTM 模型，引入了长时间信息有效性的机制，这些信息有选择性地被控制并保存下来，从而解决了梯度的问题。

在声纹识别过程中，国内的研究人员针对 LSTM 总结了两个重要特性：1)结构深度(多层)提取面向任务的特征；2)时间深度(循环连接)积累动态证据[41-42]。文献[43]指出，在 LSTM 中，存储器单元状态可随时间进行重复更新。但是，LSTM 只能记住顺序信息，在学习过程中不能持续关注显著性区域。

在语音识别任务中，国外研究人员使用 LSTM 解决了 DNN-HMM 无法固定窗口长度且无法有效利用上下文的问题，提升了语音识别的效果。文献[44]将 i-vector 与 LSTM 进行融合，既体现了递归神经网络的优势，又弥补了与话者无关的语音特征在识别准确率上的不足。针对许多时间序列任务在未来时刻存在反向相关性等问题，可利用双向 LSTM 模型[45]进行序列评估，文献[46]利用再训练的双向 LSTM 模型执行自发语音分类任务，基于先前的判断，结合 HMM 显著提升了对音素判断的能力。文献[47-48]提出了 2D 时频(TF)LSTM 和 Grid-LSTM 模型，其随时间和频率的变化来模拟大规模语音识别任务。但是，复杂的模型体系结构容易在小规模数据集上过拟合。此外，简单的平均池化或最大池化可能不足以导出需要分析更高阶统计量的复杂情绪的有效表示。

传统的 LSTM 结构在多个维度上缺乏对上下文依赖关系的理解，因此卷积循环神经网络(Convolutional Recurrent Neural Network，CRNN)应运而生，这是一种端到端的、可结合变长序列的模型，常与 CTC (Connectionist Temporal Classification)结合使用。文献[51]利用 CRNN 学习从时域语音信号到情感连续值分类的映射，来高效地预测连续的激活与效价空间。文献[52]比较了 CNN，CRNN 和 ResNet 3 种模型，在进行数据平衡之后发现 CRNN 的效果最好，它可以用较低的推理成本来建模复杂的空间依赖关系，占用更少的存储空间。

在情感识别方面，国外 Lee 等[49]应用 RNN 来学习情感识别的时空关系。文献[21]尝试向 LSTM 层添加一维注意力机制，以找到与情绪识别相关的语音信号的重要时间段。国内 Han 等[50]采用多种方法来解决 LSTM 的输出长度和标签长度不匹配的问题，采用 CTC 机制提升学习效率。

目前，常见的 RNN 相关模型有深度 RNN 结合多层感知机、双向 RNN、循环卷积神经网络、多维循环神经网络、LSTM、门控循环单元和记忆网络等，它们虽然变换成多种形态，但均以长时期的历史信号和未来信号的处理为主，因此，其输入往往与频谱类特征相关，与 CNN 截然不同。RNN 和 CNN 模型虽然应用的领域不尽相同，但是均属于有监督类别

的模型。目前，多级别联合等流行模型采用多通道的形式（见3.6节）将RNN和CNN完美地结合起来。

3.4 编码器模型

自动编码器（AutoEncoder）是一种典型的前向网络，其输入层和输出层的结点数相同，它的训练过程旨在重构输入数据，即让输入值尽可能等于输出值，利用反向传播算法来更新网络的权重。由于训练数据不需要任何标签，自动编码器的训练是无监督的。

自动编码器模型在话者识别中有一定的应用。对于训练好的自动编码器，在最后一个隐层后添加一个输出层，此层中的每个结点对应一个类别，通过输出结果完成分类任务，此时整个模型可以视作一个分类器。目前，最常用和最有效的生成模型是生成式对抗网络（Generative Adversarial Networks，GAN）和变分自动编码器（Variational Autoencoder，VAE）。GAN针对生成任务进行了优化。VAE是概率图形模型，针对潜在建模进行了优化。VAE学习隐空间中的输入概率分布的参数，通过使隐分布尽可能接近隐变量的“先验”，来提升模型的有效性。相比自动编码器，VAE的主要优点是先验数据允许注入领域知识，能够估计预测中的不确定性。文献[53]提出了基于变分自动编码器（β-variational AutoEncoder，β-VAE）的i-vector说话人识别系统，可以在大型未标记数据集上以完全无监督的方式进行训练，利用β调节其训练模式为“hard”或“soft”，从而得到接近于完全协方差VAE的模型。文献[54]采用收缩自编码器（correspondence AutoEncoder，cAE）来识别噪声词对，为训练深度神经网络提供了弱监督，利用此种帧级别学习方法得到的特征优于MFCC特征的训练结果。

自动编码器模型在语音识别上也得到了一定的应用。在国外，文献[55]提出了Speech2Vec模型，用于学习从语音语料库中切除音频片段的固定长度矢量表示，它基于RNN编码器-解码器框架，编码器RNN是单层双向LSTM，解码器RNN是另一个单层单向LSTM。编码器-解码器框架还采用了注意力机制，并借用了词袋进行训练，从语音中学习单词嵌入，使Speech2Vec能够利用语音所携带的语义信息。

目前，在语音识别方面流行的模型为Sequence-to-sequence（Seq2seq），此模型是从序列到序列的过程，一般由多层RNN（RNN，LSTM，GRU等）构成，其输出可以为可变长度的向量，在语音识别任务中，该模型常与注意力机制结合使用，以提升目标句中文字间的关联性。

国内研究人员将自动编码器应用于语音情感识别。文献[56]将VAE应用于解决语音情感分类问题，对IEMOCAP数据集的评估表明，VAE学习的特征有利于语音情感的分类。其利用生成模型的能力来学习数据的真实分布，从而自动创建强大的特征。文献[57]提出了一种多声道自动编码器（Multi-Channel Auto-Encoder，MTC-AE），用于从声学信息中识别情感。MTC-AE包含基于不同低级描述符的多个本地DNN，将每个本地DNN的瓶颈层连接在一起作为训练全局分类器的总表示；同时，训练全局分类器和本地分类器不仅考虑了不同低级描述符的相关性和独立性，而且减轻了由高维输入和数量限制的训练数据引起的过拟合。在国外研究中，文献[58]提出了条件变分自动编码器，用于学习语音情感的潜在表现。为了客观地测量这种潜在表示形式，其使用LSTM训练潜在表示并将结果作为特征，来对语音情感进行识别，并提供了新的VAE-LSTM特征提取框架。在特征选择上，高维特征不一定产生最佳的性能，需要根据经验确定合适数量的隐特征，以避免选择的特征维度过小或过大。

3.5 瓶颈特征（Bottleneck）

瓶颈特征是在多层感知（Multiplayer Perceptron，MLP）中的瓶颈层产生的特征，经过层层非线性模型分离出前后扩展的语音特征中有利于输出分类的特征信息，起初模型中的神经元个数较少，早期往往使用深度信念网络模型。随着深度学习模型的多样化发展，瓶颈特征开始应用于语音任务的相关模型，实现了对系统性能的提升与简化。

针对声纹识别，其全局特征将整个话语编码为固定长度的矢量。而帧级特征经常用于训练短语音素、说话者身份或其组合任务[59-60]，也可用于提取传统的倒谱特征。文献[61]使用精准的音素单元训练高性能的7层DNN模型，其中第6层得到的Bottleneck特征是DNN后验信息的抽象表示，优于SR或LR特征的输出后验信息，降低了系统的性能。在短语条件下，文献[47]从说话者识别DNN中提取Bottleneck特征，在训练过程中消除了说话者的词汇多样性，然而在测试过程中，由于Bottleneck提取的特征信息的复杂度不够，导致模型没有达到良好的性能。

针对语音识别，Bottleneck特征可以充分挖掘相邻帧之间的关联，从而寻求最重要的显著性特征。文献[62]将Bottleneck特征与DBN融合，形成BN-DBN方法，其对输入的要求不严格，而且对话者、外界噪声等干扰信息的鲁棒性很强，从而提升了语音识别的表征能力。

针对情感识别，Bhargava等[63]使用堆叠Bottleneck结合DNN模型来训练窗口语音波形，获得的结果与MFCC相似，同时提升了系统性能。文献[64]提出了一种深度谱特征，它将相邻帧的语谱图特征串联之后，再与DBN进行组合，从中间层获取Bottleneck特征，从而降低了有效信息的丢失率，提升了识别率。

3.6 多通道模型融合

以上3种有监督模型和编码器相关的无监督模型各有其特点和擅长的领域，而且输入特征不尽相同，侧重点也不同。近年来，越来越多的研究人员致力于将这几种模型有效地组合在一起，试图发挥全部模型的特点，提升识别的准确度。

针对话者识别，文献[65]引入TDNN提取语音的嵌入，然后基于PLDA后端对嵌入进行相似度打分。该模型采用端到端的方法，可用于区分来自可变长度语音段的说话者。

DeepSpeech作为最流行的语音识别系统，共经历了3代，均采用多通道融合的端对端深度学习框架。其第2代将DCNN应用于语音识别声学建模中，并将其与基于LSTM和CTC的端对端技术相结合，大幅提升了语音识别产品的性能。最新的DeepSpeech 3进一步简化了模型，使用带有冷聚变的Seq2Seq模型能更好地运用语言信息，以获得更好的泛化效果和更快的收敛速度。

针对情感识别，国外的Keren等[66]将CNN与LSTM相

结合，以改进基于梅尔滤波器组或原始信号的语音情感识别。文献[67]将 DNN 应用于频谱图，针对可变长度语音段提出了一种情感识别方法。该方法从频谱图中提取情绪特征，并将 CNN 与 RNN 相结合来完成情绪识别任务。与传统方法相比，此方法可以解决语音分割过程中产生的精度降低的问题。文献[68]研究了两种基于 CRNN 的联合表征学习结构，旨在从语音中捕获更丰富的情感信息。其通过 CRNN 与协作手工制作的高级统计特征(High level Statistics Functions，HSF)，开发了双通道 SER 系统(HSF-CRNN)，以共同学习具有更好辨别性的情感相关特征。此外，文献[68]还提出了另一种双通道 SER 系统，从不同的时间尺度的谱图片段中获取特征，用于联合表示学习。CRNN 输入的一部分数据是声谱图，结合 HSF 手工特征，可以更好地学习情感特征。文献[21]介绍了一种两步预测的方法，并利用此方法提升了未加权平均(Unweighted Accuracy，UA)，由于丰富情感语音中的大部分信息是中立的信息，而带有情感性质的内容只占很少的部分，因此对于平静等类别的情感，需要进行进一步判定。

目前，对于语音情感识别，最流行的处理方法是将多通道模型与注意力机制相融合，在提升效率的基础上，增加对显著性区域的探索，从而用最有效的特征和最短时间来实现高效的识别。文献[43]指出，注意力可以被描述为用于分配有限信息处理能力的“选择机制”，因此它有助于快速分析信息，并将所有计算能力置于重大任务上，以自下而上的方式从特征中获得显著区域，关注数据中的重要信息。文献[70]提出了一种基于三维注意力的卷积递归神经网络来学习 SER 的判别特征，该方法不仅可以保留有效的情感信息，还可以减小情感无关因素的影响，从而减少错误分类。文献[70]提取了 743 维特征向量，在对其进行 PCA 白化的基础上，将其馈入基于注意力机制的双向 LSTM 和全卷积网络中，以获得更准确的情绪预测。

所谓多通道的融合，是将不同神经网络的优势串联在一起，从而提升不同任务的识别效率。由于每个模型采用的输入特征不尽相同，利用多通道融合可以完善局部特征和全局特征的基本要素，在多通道的基础上，模型结合注意力机制、编码器和瓶颈特征，反向获取输入特征中的显著性区域，反向促进了识别效率的提高。

4 个性化和通用网络模型的特征提取在语音任务上的应用

在语音任务解决方案中，研究人员针对的语音任务是各自独立的，甚至任务之间的交叉会产生负面影响。然而，在实际生产环境下，3 种语音任务是无法完全剥离的。基于此，需要寻求一种完善的网络模型，来同时解决声纹识别、语音识别和情感识别任务问题，使三者形成相辅相成的关系，而非对立关系。这类模型实现的前提是具备简单的结构和复杂的细节，否则将无法同时应用于 3 种不同的任务之中。

目前，现有的研究成果主要围绕其中两种任务的结合，例如，文献[72]将从音频、视频和文本中提取的特征量化为词袋表示形式，将每个低级特征分配给来自码本的单词来量化低级特征，将 MFCC 用于视频的局部图像描述符或文本的图谱，通过这种方式提升语音识别的效率。文献[5]设计了一种循环 LSTM 模型来构建 ASR 组件和 SRE 组件，可以同时完成语音和声纹识别。该模型基于统一的神经网络，其中一个任务的输出被馈送到另一个任务的输入，从而实现多任务循环网络，此方法实现了 ASR 和 SRE 任务的改进。文献[72]提出了统一的模型来同时执行语音识别和声纹识别。该模型基于多任务循环神经网络来产生协作学习框架，可以在任务之间共享信息。由于组件信息引起了说话者归一化效应，因此其对语音识别的改进尤其显著。

目前，鲜有将情感识别与另外两种任务相融合的实例[73-74]，这是由于情感识别的成功率仍处于瓶颈期，研究人员在情感识别的特征提取和模型设计环节发挥了重要的作用，但是并未辅助其他的语音任务。事实上，单纯的语音信号并非是情感识别的唯一数据来源，可以利用声纹识别和语音识别的信息进行情感的再次判定，从而提升情感识别的成功率，同时声纹识别也为语音识别提供了一定的准确度。

针对不同的实际任务，3 种语音任务的融合形式也不尽相同。本文在综述流行技术的基础上，针对不同的背景提出两种不同的融合形式，即个性化模型和通用模型。

个性化模型主要针对声学信号的众多个性化特征。本文设计了一个基于多任务的 Tandem 模型，如图 3 所示。此模型由多层组成，第 1 层为声纹识别，通过此层得到话者的基本信息和显著性特征；将其与语音识别的特征一同馈入第 2 层，即语音识别层，基于此获得了话者信息和文本信息；在提取关键字的基础上，将其与低级描述符一同送入情感识别层，利用个性化的语音特征和情绪特征进行情感识别，从而提升情感的识别率。

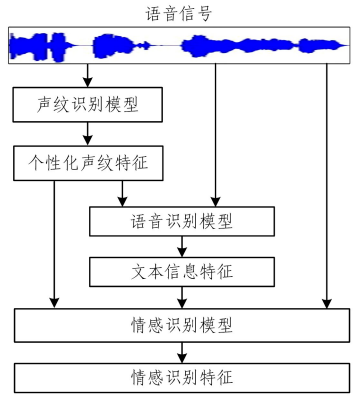

图 3 基于多任务的 Tandem 模型

Fig. 3 Multi-task model based on Tandem

针对用户需求的不同，亦可将 Tandem 模型转换为 2 层结构，以有针对性地进行语音识别和情感识别。

通用模型主要用于同时解决 3 个任务，它需要各个任务协作进行特征提取。针对通用任务，可以利用前文综述的公共特征和任务的公共模型，来设计一个多通道的网络模型，每种任务可以自主选择若干条通道，协作完成特征提取，从而实现一个输入经历多条通路来解决多个任务。多通道网络模型较为灵活，可以采取多种模型的组合方式，图 4 列举了其中的一种可行方案。

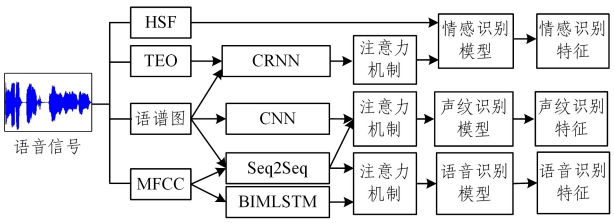

图 4 多通道网络模型

Fig. 4 Multi-channel network model

此种多通道网络模型主要由 CNN，BIMLSTM，CRNN 和 Seq2Seq 组成，此网络模型的输入为语谱图、MFCC、TEO 和低级描述符，数据源则为同一音频文件。针对声纹识别，可采用语谱图作为输入，结合 CNN 和 Seq2Seq 通道进行输出；针对语音识别，可采用 MFCC 作为输入，结合 BIMLSTM 和 Seq2Seq 通道进行输出；针对语音情感识别，可采用语谱图和 TEO 作为 CRNN 的输入，结合低级描述符和高级统计函数(HSF)通道进行输出。此外，可以在 3 个不同的语音特征提取任务中，增加注意力机制来获得显著性特征区域，从而提取关键特征信息，利用反向传播的原理提升关键特征的准确度。

虽然语音识别、声纹识别和情感识别均隶属于声学识别任务，但研究人员往往认为三者之间存在相互制约的关系，如果设计符合 3 种任务的合理的特征集合和模型通道，它们之间亦可实现相互促进、提升识别率的效果。因此，基于多任务的网络模型将是未来语音任务研究的一个新的有利方向。

结束语 本文针对语音学领域的 3 个重要研究方向——语音识别、声纹识别和情感识别来展开研究，通过总结得出，韵律特征、音质特征、频谱特征、图谱特征对 3 个任务的特征描述起到了重要的作用。近年来，随着深度神经网络技术的发展，深度学习模型在语音任务上也得到了广泛应用，DNN、CNN、RNN、编码器模型、瓶颈特征、多通道模型及注意力机制在 3 个任务上也发挥了很大的作用。针对不同的应用和设计场景，例如多任务、多通道、个性化、通用化等一种或多种组合，本文分别提出了可行的解决方案。

在未来的研究过程中，我们将进一步简化个性化和通用的网络模型结构，寻求一种通用的特征集合来同步提升 3 个语音任务的效率。

参 考 文 献

[1] ZHANG S, ZHANG S, HUANG T, et al. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching[J]. IEEE Transactions on Multimedia, 2017, 20(6): 1576-1590.

[2] RICHARDSON F, REYNOLDS D, DEHAK N. A Unified Deep Neural Network for Speaker and Language Recognition[J]. arXiv: 1504.00923.

[3] KANAGASUNDARAM A, DEAN D, SRIDHARAN S, et al. DNN based Speaker Recognition on Short Utterances[J]. arXiv: 1610.03190.

[4] LEE J, LEE M, CHANG J H. Ensemble of Jointly Trained Deep Neural Network-Based Acoustic Models for Reverberant Speech Recognition[J]. arXiv: 1608.04983.

[5] TANG Z, LI L, WANG D. Multi-task Recurrent Model for Speech and Speaker Recognition[C]// 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2016.

[6] CHU W, CHEN R. Speaker Cluster-Based Speaker Adaptive Training for Deep Neural Network Acoustic Modeling[C]// ICASSP 2016. IEEE, 2016.

[7] GHAHABI O, HERNANDO J. Deep Learning for Single and Multi-Session i-Vector Speaker Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(4).

[8] JIN Q, CHEN S Z, LI X R, et al. Speech emotion recognition based on acoustic characteristics [J]. Computer Science, 2015, 42(9): 24-28.

[9] WANG W, YANG L P, WEI L, et al. Extraction and Analysis of Speech Emotion Characteristics[J]. Research and Exploration in Laboratory, 2013, 32(7): 91-94, 191.

[10] YANG M H, TAO J H, LI H, et al. Nature Multimodal Human-Computer-Interaction Dialog System [J]. Computer Science, 2014, 41(10): 12-18, 35.

[11] RAMANARAYANAN V, PUGH R, QIAN Y, et al. Automatic Turn-Level Language Identification for Code-Switched Spanish-English Dialog[C]// 9th International Workshop on Spoken Dialogue System Technology. 2019: 51-61.

[12] DELLAERT F, POLZIN T, WAIBEL A. Recognizing emotion in speech[C]// International Conference on Spoken Language. 1996.

[13] AHMAD J, FIAZ M, KWON S I, et al. Gender Identification using MFCC for Telephone Applications- A Comparative Study [C]// International Journal of Computer Science and Electronics Engineering 3.5. 2015: 351-355.

[14] BANDELA S R, KUMAR T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC [C]// International Conference on Computing. IEEE Computer Society, 2017.

[15] ZHAO W, GAO Y, SINGH R, et al. Speaker identification from the sound of the human breath[J]. arXiv: 1712.00171v2.

[16] DENG L. A tutorial survey of architectures, algorithms, and applications for deep learning[J]. Apsipa Transactions on Signal & Information Processing, 2014, 3.

[17] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[18] AWNI H, CASE C, CASPER J, et al. Deep Speech: Scaling up end-to-end speech recognition[J]. arXiv: 1412.5567.

[19] AMODEI D, ANUBHAI R, BATTENBERG E, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[J]. arXiv: 1712.00171.

[20] SATT A, ROZENBERG S, HOORY R. Efficient emotion recognition from speech using deep learning on spectrograms[C]// Proc. Interspeech 2017. 2017: 1089-1093.

[21] EYBEN F, SCHERER K R, TRUONG K P, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Re-

search and Affective Computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[22] MULIMANI M, KOOLAGUDI S. Robust Acoustic Event Classification using Bag-of-Visual-Words[C]// Proc. Interspeech. 2018: 3319-3322.

[23] LI L, WANG D, ZHENG T F. System Combination for Short Utterance Speaker Recognition[C]// Signal & Information Processing Association Summit & Conference. IEEE, 2016.

[24] ZHANG M, CHEN Y, LI L, et al. Speaker Recognition with Cough, Laugh and "Wei"[J]. arXiv: 1706.07860.

[25] LI L, WANG D, ZHANG Z, et al. Deep Speaker Vectors for Semi Text-independent Speaker Verification[J]. arXiv: 1505.06427.

[26] LU L. Sequence Training and Adaptation of Highway Deep Neural Networks[C]// 2016 IEEE Spoken Language Technology Workshop (SLT). 2016.

[27] HAN K, YU D, TASHEV I. Speech emotion recognition using deep neural network and extreme learning machine[C]// Fifteenth Annual Conference of the International Speech Communication Association. 2014: 223-227.

[28] MAO Q, MING D, HUANG Z, et al. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks[J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213.

[29] SARMA M, GHAHREMANI P, POVEY D. Emotion Identification from raw speech signals using DNNs[C]// Interspeech. 2018: 3097-3101.

[30] PALAZ D, COLLOBERT R, et al. Analysis of cnn-based speech recognition system using raw speech as input[C]// Proceedings of Interspeech. 2015: 11-15.

[31] SAINATH T, PARADA C. Convolutional neural networks for small-footprint keyword spotting[C]// Proceedings of Interspeech. 2015: 1478-1482.

[32] CHEN L, LEE C M. Predicting Audience's Laughter Using Convolutional Neural Network [J]. arXiv: 1702.02584.

[33] CHAN W, LANE I. Deep convolutional neural networks for acoustic modeling in low resource languages[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. 2015: 2056-2060.

[34] HUANG Y L, LUO X X, LIU D R. Local Finite Weight Sharing of MFSC Coefficients Based CNN Speech Recognition[J]. Control Engineering of China, 2017, 24(7): 1507-1513.

[35] ALDENEH Z, PROVOST E M. Using regional saliency for speech emotion recognition[C]// IEEE International Conference on Acoustics. IEEE, 2017.

[36] KHORRAM S, JAISWAL M, GIDEON J, et al. The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild [C]// Interspeech 2018. 2018: 1903-1907.

[37] HUANG C W, NARAYANAN S. Shaking acoustic spectral sub-bands can better regularize learning in affective computing [C]// ICASSP 2018- 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[38] ZHENG W Q, YU J S, ZOU Y X. An experimental study of speech emotion recognition based on deep convolutional neural networks[C]// 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE Computer Society, 2015.

[39] NIU Y, ZOU D, NIU Y, et al. A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks[J]. arXiv: 1707.09917.

[40] SWIETOJANSKI P, RENALS S. Differentiable Pooling for Unsupervised Acoustic Model Adaptation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(10): 1773-1784.

[41] WANG D, ZHENG T F. Fransfer learning for speech and language processing[C]// Proceedings of APSIPA Annual Summit and Conference. APSIPA, 2015.

[42] HUANG J T, LI J, YU D, et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013: 7304-7308.

[43] ZHONG G, LIN X, CHEN K. Long Short-Term Attention[J]. arXiv: 1810.12752.

[44] GUPTA V, KENNY P, OUELLET P, et al. I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2014: 6334-6338.

[45] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional lstm networks[C]// International Joint Conference on Neural Networks. 2005.

[46] BERINGER N, GRAVES A, SCHIEL F, et al. Classifying Unprompted Speech by Retraining LSTM Nets[J]. Lecture Notes in Computer Science, 2005, 58(1956): 575-581.

[47] LI J, MOHAMED A, ZWEIG G, et al. Exploring multidimensional lstms for large vocabulary ASR[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016: 4940-4944.

[48] LI B, SAINATH T N, NARAYANAN A, et al. Acoustic modeling for Google home[C]// Proc. of INTERSPEECH. 2017: 399-403.

[49] LEE J, TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C]// Interspeech. 2015.

[50] HAN W J, RUAN H B, CHEN X M. Towards Temporal Modelling of Categorical Speech Emotion Recognition[J]. arXiv: 10.21437/Interspeech, 2018.

[51] TRIGEORGIS G, RINGEVAL F, BRÜCKNER R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]// IEEE International Conference on Acoustics. IEEE, 2016.

[52] TANG D, ZENG J, LI M. An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals [C]// Proc. Interspeech. 2018: 162-166.

[53] PEKHOVSKY T, KORENEVSKY M. Investigation of Using VAE for i-Vector Speaker Verification[J]. arXiv: 1705.09185.

[54] KAMPER H, JANSEN A, GOLDWATER S. A segmental framework for fully-unsupervised large-vocabulary speech recognition[J]. Computer Speech & Language, 2017, 46: 154-174.

[55] CHUNG Y A, GLASS J. Speech2vec: A sequence-to-sequence

framework for learning word embeddings from speech[C]// INTERSPEECH. 2018:811-815.

[56] LATIF S,RANA R,QADIR J. Variational Autoencodersfor Learning Latent Representations of Speech Emotion:A Preliminary Study[C]// Interspeech 2018. 2018:3107-3111.

[57] ZONG Z F,LI H,WANG Q. Multi-Channel Auto-Encoder for Speech Emotion Recognition[J]. arXiv:1810. 10662v1.

[58] LATIF S,RANA R,YOUNIS S,et al. Transfer Learning for Improving Speech Emotion Classification Accuracy[C]// INTERSPEECH. 2018:257-261.

[59] LI C,MA X,JIANG B,et al. Deep Speaker:an End-to-End Neural Speaker Embedding System[J]. arXiv:1705. 02304.

[60] DUMPALA S H,PANDA A,KOPPARAPU S K. Improved I-vector-based Speaker Recognition for Utterances with Speaker Generated Non-speech sounds[J]. arXiv:1705. 09289.

[61] YI L,LIANG H,YAO T,et al. Comparison of Multiple Features and Modeling Methods for Text-dependent Speaker Verification[C]// 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE,2017.

[62] LI J H,YANG J A,WANG Y. New Feature Extraction Method Based on Bottleneck Deep Belief Networks and its Application in Language Recognition[J]. Computer Science,2014,41(3):263-266.

[63] BHARGAVA M,ROSE R. Architectures for deep neural network based acoustic models defined directly over windowed speech waveforms[C]// INTERSPEECH. 2015:6-10.

[64] LI S,XU L T. Research on Emotion Recognition Algorithm Based on Spectrogram Feature Extraction of Bottleneck Feature [J]. Computer Technology and Development,2017,27(5):82-86.

[65] SNYDER D,GARCIA-ROMERO D,POVEY D. Deep neural network embeddings for text-independent speaker verification [J]. arXiv:10. 21437/Interspeech. 2017.

[66] KEREN G,SCHULLER,BJÖRN. Convolutional RNN:an Enhanced Model for Extracting Features from Sequential Data [C]// 2016 International Joint Conference on Neural Networks (IJCNN). 2016.

[67] MA X,WU Z,JIA J,et al. Study on Feature Subspace of Archetypal Emotions for Speech Emotion Recognition[C]// ICASSP-2017. 2016.

[68] Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms[C]// Interspeech 2018. 2018:3683-3687.

[69] LUO D Q,ZOU Y X,HUANG D Y. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition[C]// 2018 Conference of the International Speech Communication Association (INTERSPEECH 2018). 2018:152-156.

[70] MINGYI C,XUANJI H,JING Y,et al. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition[J]. IEEE Signal Processing Letters,2018:1.

[71] SAKR M,ANDRIENKO G,BEHR T,et al. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems[C]// Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2011:505-508.

[72] NICHOLAS C,SHAHIN A,SANDRA O. Multimodal Bag-of-Words for Cross Domains Sentiment Analysis[C]// IEEE International Conference on Acoustics,Speech,and Signal Processing. IEEE,2018.

[73] LI L,TANG Z,DONG W,et al. Collaborative Learning for Language and Speaker Recognition[C]// ICASSP 2017. 2017.

[74] LI Y,WEI Z H,XU K. Hybrid Feature Selection Method of Chinese Emotional Characteristics Based on Lasso Algorithm [J]. Computer Science,2018,45(1):39-46.

ZHENG Chun-jun, born in 1976, master, associate professor, is a member of China Computer Federation. His main research interests include speech emotion recognition, deep learning and big data analysis.