doi: 10.12068/j.issn.1005-3026.2020.12.002

基于CGRU模型的语音情感识别研究与实现

郑 艳, 陈家楠, 吴 凡, 付 彬
(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 目前普遍使用深度神经网络用于语音情感特征的提取,但使用哪种神经网络模型、如何缓解模型过拟合问题还需进一步研究.针对这些问题,提出了一种结合一维卷积(CNN)以及门控循环单元(GRU)的CGRU模型,从原始语音信号的MFCC特征中提取语音的低阶以及高阶情感特征,并通过随机森林对其进行特征选择,在三种公用的情感语料库EMODB,SAVEE,RAVDESS上分别取得了79%,69%以及75%的识别精度.通过添加高斯噪声及改变速度等方法来增加样本量实现数据扩充,进一步提高了识别精度.通过在线识别系统验证了模型在实际环境中的可用性.

关 键 词: 语音情感识别;梅尔频率倒谱系数;CGRU模型;随机森林;数据扩充
中图分类号: TN 912.3 **文献标志码**: A **文章编号**: 1005-3026(2020)12-1680-06

Research and Implementation of Speech Emotion Recognition Based on CGRU Model

ZHENG Yan, *CHEN Jia-nan*, *WU Fan*, *FU Bin*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: CHEN Jia-nan, E-mail: ChenJianan86025@outlook.com)

Abstract: Speech emotion recognition is a very important research direction in emotion computing and human-computer interaction. At present, deep neural network is widely used to extract emotional features of speech, but further research is needed on which neural network model to use and how to alleviate the problem of model overfitting. To solve these problems, a CGRU model was proposed, which combined one dimensional convolutional neural networks (CNN) and gated circulation unit (GRU). The low-order and high-order emotional features of speech were extracted from the MFCC features of the original speech signal, and the features were selected through random forest, which achieved 79%, 69% and 75% recognition accuracy respectively on three common emotional corpus: EMODB, SAVEE, RAVDESS. By using the data augmentation technique, the sample size was increased by adding gaussian noise and changing the speed, which further improved the identification accuracy. The availability of the model in the real world was verified through the online identification system.

Key words: speech emotion recognition; Mel-frequency cepstral coefficients; CGRU model; random forest; data augmentation

MIT媒体与科学教授Rosalind Picard传递了这样一种思想:如果人类想让计算机拥有真正的智能,并且与之自然地互动,那么必须赋予计算机识别、理解,甚至表达情感的能力[1].语音情感识别正是赋予计算机智慧的重要组成部分,近年来,随着深度学习的不断发展,越来越多的深度神经网络结构被用于语音情感特征的提取.

Kim等使用深度信念网络用来提取语音情感的非线性特征[2].Deng等提出了一种使用自编码来重构训练数据的方法,提高了系统的鲁棒性[3].Lee等提出了一种基于高效学习算法训练的递归神经网络语音情感识别系统,为了提取情感在时间维度上的高层次表达,采用了双向长短期记忆网络[4].目前最常用的研究方法是卷积神

收稿日期: 2020-02-05
基金项目: 国家自然科学基金资助项目(61773108).
作者简介: 郑 艳(1963-),女,辽宁沈阳人,东北大学副教授,博士.

经网络(CNN)与循环神经网络(RNN). 先前的研究发现,这两种方法各有优缺点. 卷积能够从时域以及频域两个方向提取语音的局部情感特征,具有良好的泛化性能. 循环神经网络善于对序列数据建模. 但是基于语谱图和 CNN 进行语音情感识别的方式有两大缺陷:第一,这种研究方式相当于将问题转化成一个图片分类的问题,而语谱图和普通图片相比没有很好的区分度;第二,语音和图片数据有本质的区别, 对于语音这种时序数据会丢失很多时间维度上的信息. 使用 RNN 模型进行语音情感识别也有两个缺点:第一,输入的是低层次的特征,模型的识别率很大程度上取决于这些输入的特征,这违背了特征自动提取的原则;第二,语音情感识别有别于语音识别,尽管语音是序列数据,但是本文研究的任务是情感识别. 在前人研究的基础上,本文提出了一种新的网络结构,将一维 CNN 与 GRU 进行结合,简称为 CGRU 模型,用于提取语音中的高阶情感特征. 针对所提取的特征维度较大而造成过拟合的问题,使用随机森林对特征进行特征选择,减少特征的冗余,从而提高特征的识别率. 并引入数据扩充技术,通过添加高斯噪声以及改变速度等方法来增加样本量,进一步提高了识别精度.

1　基于 CGRU 模型的语音情感识别

基于 CGRU 模型的语音情感识别框架如图 1 所示,首先对语音样本进行预加重、分帧加窗等预处理,然后通过特征提取得到 40 维梅尔频率倒谱系数(MFCC),接着将 40 维 MFCC 特征输入一维 CNN[5-6] 中沿着时间轴方向进行特征提取,所提取出的局部情感特征沿着时间步输入 GRU 中,得到一段语音中的全局情感特征,最后通过全连接层实现情感的分类.

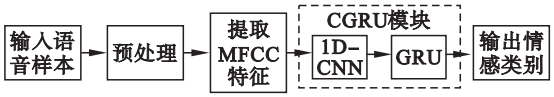

图 1　CGRU 模型语音情感识别框架

Fig. 1　Framework of CGRU speech emotion recognition

1.1　MFCC 特征提取

语音情感识别的关键一步是特征提取,特征提取是提取语音中对于分类有益的成分,丢弃其他成分. 其中对于分类有益的特征通常包含谱特征,也就是和频率相关的特征. MFCC 可以用来准确地反映声道的形状,而声音的产生和声道的形状有很大关联,因此 MFCC 特征是语音情感识别中最常用的特征[7].

1.2　基于随机森林的 CGRU 模型改进

数据经过一维卷积以及 GRU 所提取出的特征经过全连接层完成分类,全连接层的作用主要是将学到的"分布式特征表示"映射到样本标记空间,但是同时全连接层的参数占据了所有参数中的大部分,如果直接使用全连接层配合 SoftMax 激活函数进行分类,过多的参数会增加过拟合的风险. 因此提出一种设想:将 CGRU 模型作为特征提取工具,使用随机森林进行特征选择.

传统标准算法的 MFCC 只反映了语音参数的静态特性,而人耳对语音的动态特性更为敏感. 而随机森林具有准确率高、鲁棒性好、易于使用等优点. 具体的操作是,训练一个 CGRU 模型,然后在模型的 GRU 层的输出断开,重新输入样本,得到每个样本的预测值,这里的预测值就是每个样本的向量表示,向量即为语音样本的特征表示,再通过随机森林对特征向量进行特征选择,减少冗余的特征,最后通过支持向量机对降维后的特征矩阵进行训练得到一个模型,用支持向量机模型进行情感的识别. 相对于直接通过多个全连接层连接 SoftMax 的分类方式,支持向量机的预测往往更准确.

2　实验研究

2.1　情感语料库

合适的情感语料库是正确识别情感的前提条件,在评价一个语音情感识别系统时,其中一个重要标准是语料库的质量,语料库的质量直接关系着后续提取的特征,关系着整个系统的可靠性.

本文采用三种最常用的表演型情感语料库:EMODB[8], SAVEE[9] 和 RAVDESS[10] 语料库. EMODB 是柏林工业大学录制的德语情感语料库,一共有 535 个语音样本;SAVEE 是由 4 名男演员演绎的 7 种不同情感的录音,共计 480 句英式英语. RAVDESS 包含 7356 个文件,由 24 名专业演员以中性的北美口音说出包含中性、快乐、悲伤、愤怒、害怕、惊讶和厌恶等 7 种情感.

2.2　实验对比

1)基于 MFCC 特征的语音情感识别分析. 通过 Python 工具包 librosa 提取语音中的 MFCC 特征,默认的采样频率为 22 050 Hz,提取出 MFCC 特征可以用(维数,帧数)来表示,这里对帧数取平均,这样就得到了一个长度为 MFCC 维度的特

征向量,多个样本的特征向量组成一个特征数组,按照 9∶1 划分训练集和测试集,三个语料库分别有 472,427,1 296 个样本作为训练集,通过分类器对训练集的特征数组进行训练. 这里使用 sklearn 中的网格搜索得到分类器的最佳参数,同时得到最佳参数下 10 折交叉验证的平均准确率作为模型最终的准确率,10 折交叉的划分标准是完全随机划分. 为了确认 MFCC 维度对模型性能的影响,提取 13 维到 100 维的 MFCC 特征(每隔 10 维)做出如图 2 所示的折线图. 由图 2 可知,并不是维度越高识别效果越好,维度越高计算量就越大,而且容易造成过拟合,因此这里选择 MFCC 的维度为 50.

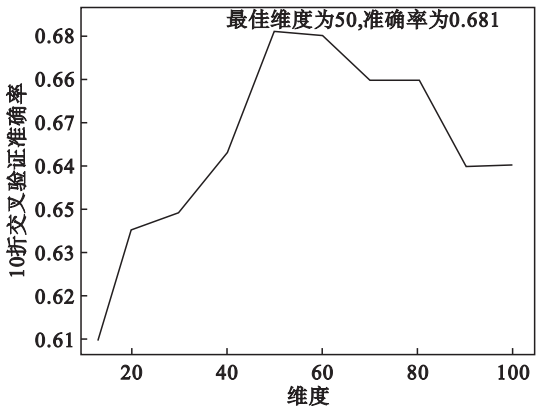

图 2 最佳 MFCC 维度

Fig. 2 Optimal MFCC dimension

这里使用的分类器是支持向量机,在网格搜索到最佳参数下的 10 折交叉验证平均准确率为 68.1%,模型在 EMODB 上的情感识别率可以用图 3 表示.

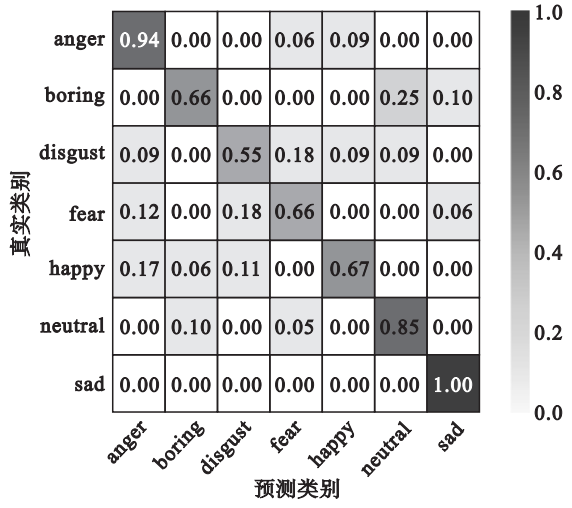

图 3 50 维 MFCC 特征识别结果混淆矩阵

Fig. 3 The 50-dimensional MFCC feature recognition confusion matrix

从混淆矩阵可以看出,以 50 维 MFCC 特征的模型对于“生气”、“伤心”和“中性”的识别率较高,对于“厌恶”这种情感识别率较低,容易与“害怕”这种情感混淆.

2)随机森林特征选择. 使用随机森林对特征进行筛选后再次进行模型训练,训练得到的模型识别结果混淆矩阵如图 4 所示. 使用随机森林对 180 维特征进行训练,训练后会根据袋外错误率得到每个特征的重要程度,对每个特征按照重要性排序,设置一个阈值,将阈值外的特征筛去,留下 41 维特征作为数据集新的特征,再使用支持向量对特征数据集重新进行训练,得到了 70.6% 的 10 折交叉验证平均识别率,比单纯的 50 维 MFCC 特征 68.1% 的识别率提高了 2.5%,这说明使用随机森林进行声学特征选择的方法有效地提高了 EMODB 中的情感识别率.

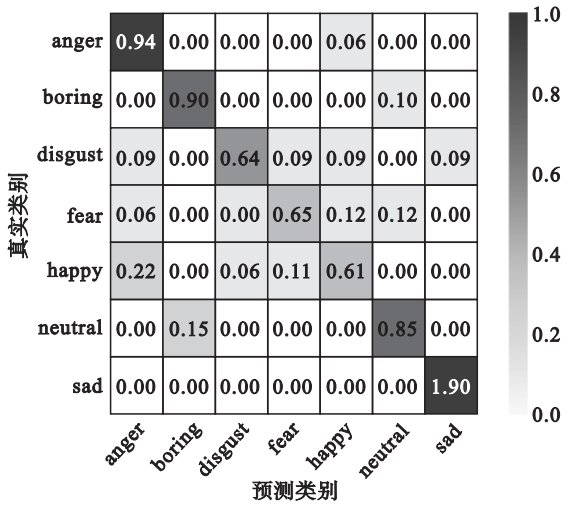

图 4 特征选择后模型识别结果混淆矩阵

Fig. 4 Recognition confusion matrix after feature selection

为了进一步验证这一方法的有效性,对另外两个数据库 SAVEE 和 RAVDESS 进行相同的实验流程,得到的实验结果如表 1 所示.

表 1 不同数据集上不同模型的 10 折交叉验证平均准确率

Table 1 Average accuracy of 10-fold cross validation for different models on different data sets

%

特征类型	EMODB	SAVEE	RAVDESS
MFCC	68.1	71.2	47.1
随机森林	70.6	73.6	50.3

2.3 实验结果及分析

1)CGRU 模型验证. 为了验证本文所提出的 CGRU 模型的实际效果,分别在 EMODB,SAVEE,RAVDESS 上进行训练. 通过留出法将

80% 的样本用于训练,20% 的样本用于验证. 模型的训练曲线如图 5 ~7 所示.

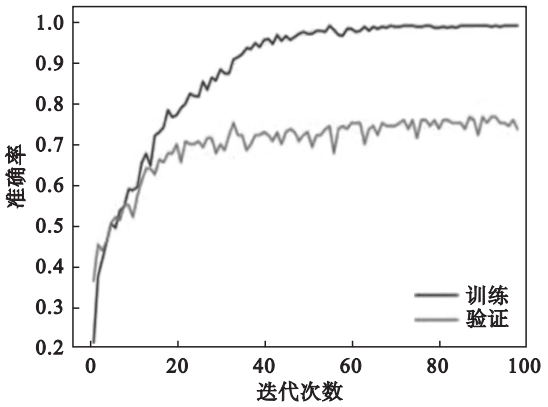

图 5　CGRU 模型在 EMODB 数据集上的训练曲线
Fig. 5　Training curve of CGRU model on EMODB

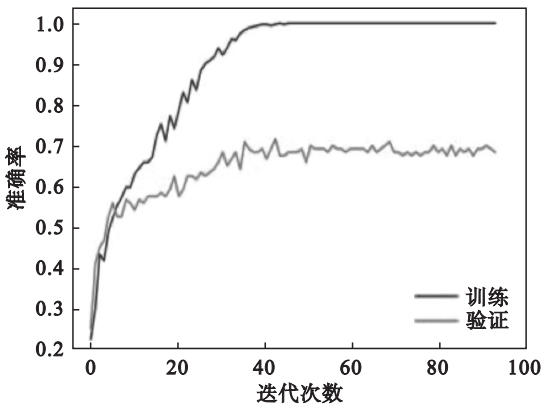

图 6　CGRU 模型在 SAVEE 上的训练曲线
Fig. 6　Training curve of CGRU model on SAVEE

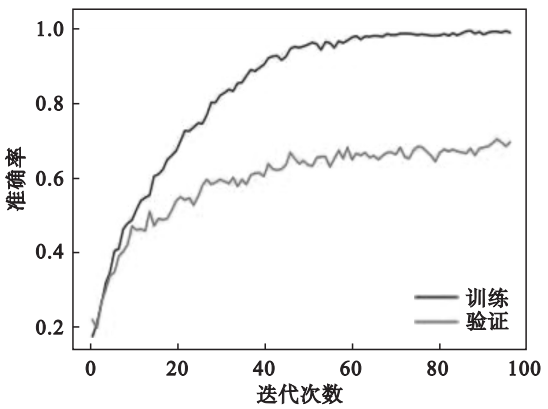

图 7　CGRU 模型在 RAVDESS 上的训练曲线
Fig. 7　Training curve of CGRU model on RAVDESS

从上述训练曲线可见,训练集精度在多次迭代后准确率接近 1,验证集精度与训练集精度存在一定的差值,说明 CGRU 模型存在一定程度的过拟合.

2)随机森林进行降维处理. 分析 CGRU 模型存在一定程度的过拟合原因可能是全连接层存在大量参数造成的,因此使用随机森林的方式对 CGRU 模型中 GRU 模块的输出特征进行特征选择(降维处理). 具体方法式是使用训练好的 CGRU 模型提取训练集以及测试集中的样本特征,然后通过随机森林对样本特征再次训练得到每一维特征的重要程度,根据设定的阈值,筛选出不重要的特征,留下的特征通过支持向量机训练,模型在三种数据集上的识别结果如表 2 所示.

表 2　CGRU 模型与 CGRU 结合随机森林对比结果
Table 2　Comparison between CGRU and CGRU combined with random forest　%

特征类型	EMODB	SAVEE	RAVDESS
MFCC	70	61	62
CGRU 模型	77	68	73
CGRU 模型 + 随机森林	79	69	75

由表 1 可知,使用随机森林进行降维后的 CGRU 模型相对于普通的 CGRU 模型的识别率平均提高了 1.7% . 这说明随机森林能够有效地筛选出冗余特征,提高模型的识别效果. 但是过拟合的情况仍未有效改善,为此进一步引入了数据扩充技术.

3)基于数据扩充的语音情感识别分析. 使用了数据扩充技术,通过添加高斯噪声、变速拉伸以及压缩等方法来增加样本量. 在全部数据集上进行扩充,数据集按照 holdout 方式,分为训练集和测试集,测试集占数据集的 30% ,在训练集上使用扩充. 使用的数据库仍然是 EMODB, SAVEE 和 RAVDESS 三种数据库. EMODB 数据库一共有 525 个样本,通过对每个样本分别单独使用添加噪声、改变速度等方式扩大一倍的样本量,这样数据就被扩充到 1 050 个,同理 SAVEE 被扩充到 950 个,RAVDESS 被扩充至 2 880 个,得到 735, 665 和 2 016 个样本作为训练集. 分别对扩充后的数据集进行三种模型的训练. 识别结果如表 3 所示.

表 3　加高斯白噪声扩充样本后识别结果
Table 3　Recognition result after adding white noise　%

特征类型	EMODB	SAVEE	RAVDESS
MFCC	79.1	70.6	76.6
CGRU 模型	84.3	74.3	80.1

表 3 为通过添加高斯白噪声的方式对样本量进行一倍扩充后的识别结果,通过和不使用样本扩充对比,使用添加噪声的方式在 EMODB 数据

集上的识别率平均提高了 9%,在 SAVEE 数据集上的识别率平均提高了 7%,在 RAVDESS 数据集上的识别率平均提高了 11%. 表 4 表示混合加噪声、变速、拉伸等三种方式扩充一倍样本后的识别结果,从中可以发现,混合多种方式样本扩充相比单一添加噪声的方式对模型识别效果又有了进一步提高.

表 4 多种方式混合扩充样本后识别结果

Table 4 Recognition result after mixing various techniques %

特征类型	EMODB	SAVEE	RAVDESS
MFCC	84.7	75.1	82.3
CGRU 模型	88.7	82.6	89.1

从表 3 和表 4 可以看出,经过数据扩充后,每一种模型的识别效果都有了很大提高,其中 CGRU 模型提升的幅度最大,这说明数据扩充后,CGRU 模型的性能得到了很大提高,过拟合的情况有了很大改善. 这一点可以从 CGRU 模型在三种数据集中的训练曲线看出,以 CGRU 模型在 RAVDESS 上训练曲线为例说明,如图 8 所示.

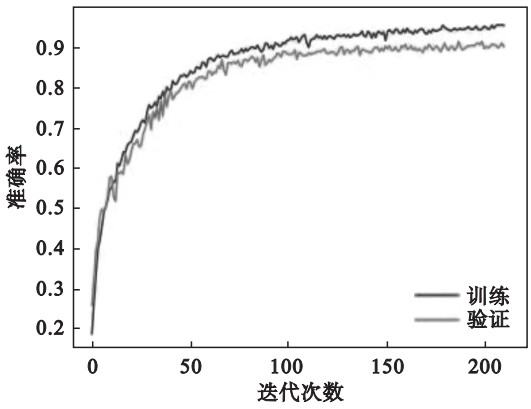

图 8 CGRU 模型在 RAVDESS 上训练曲线

Fig. 8 CGRU model's training curve on RAVDESS

由图 8 可知,CGRU 模型在扩充后的样本上的训练曲线相对于不使用样本扩充而言,训练集和测试集曲线之间的距离相差比较小. 模型的验证集精度比之前有了很大提高,这说明类似于 CGRU 这样的深度神经网络需要大量数据才能发挥网络的性能. 当缺乏数据时,模型会陷入“过度学习”,表现在训练集精度非常高,但是在验证集中的精度相对而言不是很高. 同时模型参数过多表示模型的搜索空间就越大,必须有足够的数据才能更好地刻画出模型在空间中的分布. 数据扩充可以有效地防止神经网络学习到不相关的模式,从根本上提高模型的性能. 在语音情感识别研究过程中,通过适当添加噪声、改变音速、拉伸音高等方式,在原有基础上,相当于重新获取了一批新样本,这对于深度神经网络搭建的模型提升效果显著.

从以上两组实验中可以得到结论:

1) 加噪声、变速、拉伸的数据扩充方式应用于基于 CGRU 模型的语音情感识别系统模型时,数据扩充带来的正面效益大于噪声等带来的负面影响,有效提高了系统的性能.

2) 加噪声、变速、拉伸的数据扩充方式相比于仅仅添加高斯白噪声的方式,能更好地提高识别率.

3) 加噪声、变速、拉伸的数据扩充方式并非增加的越多,使数据量越大越好. 最佳的添加比例随着数据集不同而变化,整体随着添加比例增加,模型性能先提高后降低. 原因是适度增加噪声能够增加数据量,但若增加噪声过量反而会使深度学习过程中将其作为情感分类的重要特征,使得在干净的训练集上表现不好,噪声的负面影响就会大于数据扩充的正面增益.

2.4 在线识别系统搭建

为了验证 CGRU 模型在实际环境中的表现,本文通过 python 中的 GUI 模块设计了一个简易的语音情感在线识别系统,如图 9 所示.

图 9 语音情感在线识别系统

Fig. 9 Speech emotion online recognition system

系统可以通过点击“在线录制音频”进行音频录制,然后通过点击不同的模型进行实时情感预测. 在这一简易的系统上进行说话人测试,内容包括同一语种测试以及不同语种测试,同一语种是指模型训练使用的语音和录制使用的语种一致. 这里使用了一位作者自己录制的音频,语言为中文,选取了 5 种具有代表性的情感:中性、高兴、

生气、伤心以及害怕，每一种 15 个样本，识别结果如表 5 和表 6 所示.

表 5　同一语种在线识别结果
Table 5　Online test results with same language

情感	中性	高兴	生气	伤心	害怕	识别率/%
中性	12	2	0	0	1	80
高兴	2	10	3	0	0	67
生气	0	4	9	0	1	60
伤心	0	0	0	9	6	60
害怕	1	0	0	4	10	67

表 6　不同语种在线识别结果
Table 6　Online test results with different language

情感	中性	高兴	生气	伤心	害怕	识别率/%
中性	7	2	0	3	3	46
高兴	2	6	4	1	2	40
生气	0	3	9	1	2	60
伤心	5	2	1	5	2	33
害怕	1	2	1	5	6	40

从表 5 及表 6 可以看出，在同一语种下进行测试，测试平均识别率达到了 67% ，在不同语种下进行测试，平均识别率下降到了 43% ，这说明在同一语种下模型有较好的性能，但同时模型还是会受到语种的影响. 相对于目前匮乏的情感识别系统，本文设计的 CGRU 情感在线识别系统是当前研究的一个很好补充. 文献[11]中使用的 CNN + BLSTM 的方法在 EMODB 上获得了最高 82. 35% 的识别率，而本文则达到了最高 89% 的识别率.

3　结　　论

本文以语音情感识别为研究背景，结合一维 CNN 与 GRU 提取语音中的局部和全局情感特征，并通过随机森林对提取后的特征进行特征选择. 实验结果表明 CGRU 模型提取的局部特征和全局特征在情感识别中更加接近真实的情感特征. 使用随机森林进行降维后的 CGRU 模型相对于普通 CGRU 模型识别率平均提高了 1. 7% . 说明随机森林能有效地筛选出冗余特征，提高模型的识别效果. 使用数据扩充技术进一步改善了过拟合的情况，使用添加噪声的方式在 3 个数据集上的识别率平均提高了 9% ，7% 和 11% ，因此，数据扩充能够显著提高语音情感识别效果.

参考文献：

[1] Picard R W. Affective computing[M]. Cambridge，MA：MIT Press，1997：14 – 16.

[2] Kim Y，Lee H，Provost E M. Deep learning for robust feature generation in audiovisual emotion recognition [C]// Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Vancouver，2013：3687 – 3691.

[3] Deng J，Zhang Z，Marchi E，et al. Sparse autoencoder-based feature transfer learning for speech emotion recognition [C]//2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva，2013：511 – 516.

[4] Lee J，Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[J]. *Interspeech*，2015，5(1)：10 – 13.

[5] LeCun Y，Bottou L，Bengio Y，et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*，1998，86(11)：2278 – 2324.

[6] LeCun Y，Bengio Y. Convolutional networks for images，speech，and time series[M]//Arbib M A ed. The handbook of brain theory and neural networks. Cambridge：MIT Press，1995：255 – 257.

[7] Likitha M S，Gupta S R R，Hasitha K，et al. Speech based human emotion recognition using MFCC [C]//2017 International Conference on Wireless Communications，Signal Processing and Networking (WiSPNET). Chennai，2017：2257 – 2260.

[8] Burkhardt F，Paeschke A，Rolfes M，et al. A database of German emotional speech[C]//Proceedings of Interspeech 2005. Lisbon：ISCA，2005：1517 – 1520.

[9] Jackson P，Haq S. Surrey audio-visual expressed emotion (SAVEE) database[EB/OL]. [2015 – 01 – 05]. http://kahlan. eps. surrey. ac. uk/savee/.

[10] Livingstone S R，Russo F A，Joseph N. The Ryerson audio-visual database of emotional speech and song：a dynamic，multimodal set of facial and vocal expressions in North American English[J]. *PLOS ONE*，2001，13(5)：15 – 19.

[11] Pandey S K，Shekhawat H S，Prasanna S R M. Deep learning techniques for speech emotion recognition：a review[C]// 29th IEEE International Conference Radioelektronika. Pardubice，2019：1 – 6.