浙江理工大学学报，2020，43(6)：815-822
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2020.06.011

基于注意力机制与LSTM的语音情绪识别

陈巧红，于泽源，孙　麒，贾宇波
（浙江理工大学信息学院，杭州 310018）

摘　要：针对现有语音情绪识别方法特征提取完整性和准确率较差的问题，将注意力机制和长短时记忆网络(Long short-term memory，LSTM)相结合，提出了一种语音情绪识别模型。该模型首先采用语音信号的梅尔频率倒谱系数(Mel frequency cepstrum coefficient，MFCC)作为LSTM的输入，借助LSTM对频谱序列进行建模，并在LSTM的遗忘门和输入门中做窥孔连接，将单元状态也作为输入数据加入门限层中；然后将LSTM得到的情感特征输入注意力层，计算每一帧语音信号的权重；最后使用权重较高的语音特征来区分不同情绪，完成对语音信号的情绪识别。结果表明：该模型与基础LSTM模型相比，在EMO-DB、CASIA和RAVDESS三种数据集上准确率分别提高2.96%、2.66%和7.06%，召回率和F1值也均有提高。这表明提出的模型语音分类识别性能较强，有效提升了语音情绪识别的准确率。

关键词：语音情感识别；梅尔频率倒谱系数；长短时记忆网络；注意力机制
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2020) 11-0815-08

Speech emotion recognition based on attention mechanism and LSTM

CHEN Qiaohong, *YU Zeyuan*, *SUN Qi*, *JIA Yubo*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Aiming at the problems of poor feature extraction integrity and accuracy of the existing speech emotion recognition method, a speech emotion recognition model based on the attention mechanism and long short-term memory network (LSTM) was proposed. The model firstly used the Mel frequency cepstrum coefficient (MFCC) of speech signal as the input of LSTM, and the spectrum sequence was modeled with the help of LSTM. Besides, peephole connection was carried out in the forgotten door and the input door of LSTM, the unit state was also added to the threshold layer as input data. Then, the emotional characteristics obtained from LSTM were input into the attention layer to calculate the weight of each frame of speech signal. Finally, the speech features with higher weight were used to distinguish different emotions so as to complete the emotion recognition of speech signals. Experimental results show that, compared with the basic LSTM model, the accuracy of EMO-DB, CASIA and RAVDESS data sets was improved by 2.96%, 2.66% and 7.06% respectively, and the recall rate and F1 value also increased, showing that the proposed model has a good speech recognition ability and effectively improves the accuracy of speech emotion recognition.

Key words: speech emotion recognition; Mel frequency cepstrum coefficient (MFCC); long short-term memory (LSTM); attention mechanism

收稿日期：2020－03－02　　网络出版日期：2020－06－03
基金项目：国家自然科学基金项目(51775513)
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习技术方面的研究。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

0 引 言

随着科技的进步与发展,人机交互在当今社会中得到了广泛的应用,语音识别作为人机交互的媒介之一,也渐渐成了实现自然人机交互的关键[1]。传统的语音信息处理系统,包括语音理解和语音会话模型,侧重点在于语音信号中表达词汇提取的正确性,以及语音信号中生成文本的可读性。但是语音信号不止包含着沟通的信息、表达的词汇,还隐含了发声人的情感状态。计算机的语音情绪识别,通过提取语音的声学特征来反映对应的人类情感,是实现更加和谐、高效的人机交互的基础,具有很重要的研究意义和应用价值[2]。

传统的语音情绪识别方法主要包括语音信号预处理、语音情感特征提取和语音情绪分类识别3个步骤。其中情感特征的提取和情绪识别的模型是语音信号处理的关键,直接影响语音情绪识别的精度。Nwe等[3]提出了一种基于隐马尔科夫模型(Hidden Markov model,HMM)的语音情绪识别的方法,该方法对每一种情感状态都建立隐马尔科夫模型,并使用混合高斯分布计算输出的概率函数,可以较准确地保持语音信号的整体非平稳性和局部平稳性。但是由于语音中包含多种情感状态,该方法需要建立多种隐马尔科夫模型,从而导致训练计算量较大,情感状态分类困难,整体识别率较低。Hervé等[4]为解决隐马尔可夫模型后验概率计算量大且准确率不够的问题,提出了一种使用人工神经网络(Artificial neural network,ANN)估计后验概率的方法,该方法利用转换的局部条件后验概率来估计整体后验概率,提高了后验概率的准确性。ANN可以直接学习语音信号,处理并行语音信号的效果较好,识别能力较强。Li等[5]提出一种基于深度神经网络(Deep neural networks, DNN)的语音情绪识别方法,该方法通过在输入端进行扩帧,从而能够利用上下文信息,提升模型的标注能力。同时为了避免生成性预训练单一性的缺点,该方法使用判别性预训练,即在训练中判别当前的语音标记是否被语音模型替换过,从而提高预训练的效率,得到较高的准确率。但是DNN的扩帧是有限的,所以它能够利用的上下文信息也是有限的。Kipyatkova[6]通过实验提出的长短时记忆网络(Long short-term memory, LSTM)对于大型声学建模更加有效,针对语音序列的长期依赖性特点,在每一层网络中对语音序列进行建模,相比DNN可以更快速地收敛,对于上下文信息的提取更全面,整体识别性能较高。Zhang等[7]提出了一种基于注意力机制与卷积神经网络(Convolution neural network,CNN)的语音情感识别方法,该方法通过注意力机制来计算语音信号中各个时域与情感特征的相关性权重,然后比较语音信号中不同时域的相关性权重,从中选取权重较大的时域信号进行识别,使网络能够更专注于语音的情感突出,确保关键信息不会丢失。该方法的识别精度优于基础卷积神经网络,其缺点是通过频谱图来计算情感特征需要大规模的训练,并且进行预处理时参数选取的难度较大。

针对上述方法在语音情绪识别中准确率较低且提取的语音信号特征存在信息缺失等问题,本文提出一种基于注意力机制与LSTM的语音情绪识别模型。该模型使用梅尔频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)作为情感特征,利用LSTM来学习语音序列在时间上的关联性,并通过注意力机制计算每一帧语音信号在情感特征中的权重。由于该模型融合了注意力机制和LSTM的优点,语音情绪识别的准确率和性能将得到提升。

1 基于注意力机制与LSTM的语音情绪识别模型

本文提出的基于注意力机制与LSTM的语音情绪识别模型主要包括:语音数据预处理及MFCC提取、LSTM情感特征提取、注意力机制权重计算和情绪分类输出。首先对输入的语音信号做语音数据预处理,然后使用opensmile工具进行MFCC的提取,将MFCC提取后的特征输入LSTM模型,通过添加了窥孔链接的LSTM,得到完整的语音序列。然后将LSTM中输出的矩阵作为注意力机制层的输入。在注意力机制层中,通过相似度计算,学习语音的每一帧信号相对于识别目标的注意力权重,将学习到的权重值与输入的矩阵相乘得到最后的权值。最后将获得的信息通过全连接层进行分类,由softmax层实现语音情绪的输出。整体模型如图1所示,其中:h_i, $i=1,2,\cdots,t$, 为通过LSTM后得到的每一帧语音序列;t 为语音信号的帧数。

该模型结合了LSTM和注意力机制的特性。通过LSTM提取更全面的信息,并与自注意力模型对所有输出节点的计算呼应,从而获取整段语音的完整信息。此外,通过对每一帧语音信号权重的计算,将权重较高的语音序列输入全连接层和softmax层,进行分类识别,从而完成语音情绪识别的过程。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

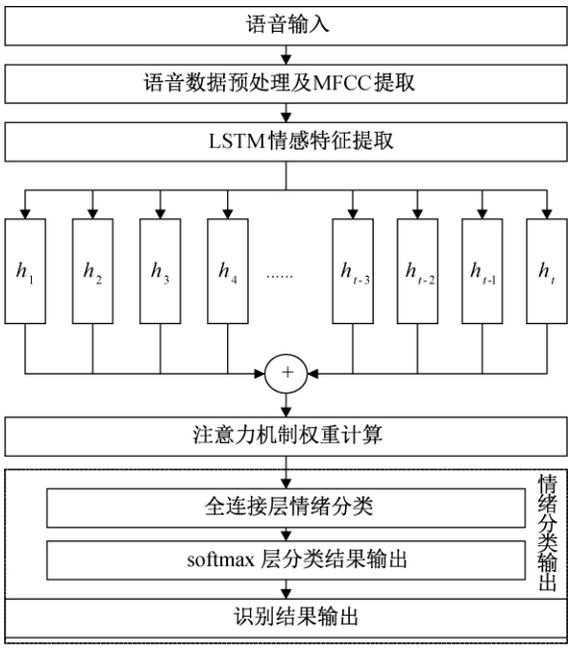

图 1　注意力机制与 LSTM 的语音情绪识别模型

1.1　MFCC 提取

传统的语音特征被分为声学特征、韵律特征和音质特征。声学特征主要分为语音中的元音、辅音；韵律特征分为能量、语速和共振峰等；音质特征代表着语音信号的清晰度[8]。由于语音情感的复杂性，很多情感仅凭韵律、音质难以有效识别，导致单一的原始特征区分效果不好，这就需要将不同的语音特征共同融合以用于语音识别。而且语音信号是非平稳的随机过程，具有很强的时变性，因此为了增加特征参数的实用性，并减少特征提取的复杂度，本文选取 MFCC[9] 作为语音情感特征。

MFCC 是一种语音情感特征参数，该参数是在梅尔标度频率域提取出来的倒谱系数，是一种在自动语音和说话人识别中广泛使用的特征。梅尔标度对于人耳频率的非线性特征描述十分准确，它与频率的计算关系可用式(1)表示：

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

其中：M 表示梅尔频率函数，f 表示线性频率。式(1)表示了梅尔频率和线性频率之间的关系。在梅尔频域内，人对音频的感知度为线性关系，而 MFCC 系数就是通过模拟人耳特性，通过人的听觉特征去构建特征参数。

1.2　LSTM 情感特征提取

循环神经网络(Recurrent neural network，RNN)作为被广泛使用的一种模型，在自然语言处理(Natural language processing，NLP)的很多任务中得到了较高的准确率[10]，但是在 RNN 训练过程中会存在梯度消失和梯度膨胀的问题。对于梯度膨胀问题，可以在训练过程中加一些约束条件[11]，当梯度超过一定值后设定一个固定值；针对梯度消失问题，比较有效的解决方法是将里面的 RNN 单元变成 LSTM 模型。因为传统的 RNN 无法很好处理权重指数级爆炸或消失的问题，所以很难处理长时间的相互关联。LSTM 作为 RNN 的改进方法之一，解决这种问题具有更强的优势。LSTM 是一种时间 RNN，和普通时间 RNN 不同，该网络的前馈神经网络接受较特定结构的输入，LSTM 在自身网络中循环传递状态，因此时间序列结构可以接受的输入范围更广泛，有描述动态时间行为的功能。LSTM 的关键就是单元状态。单元状态的传递作用在整个链结构上，在传递的过程中存在少量的线性关系相互作用，从而较容易得到信息。LSTM 通过三种门在细胞状态中去除或者增加信息。门是一种让信息有选择性通过的结构，一个 LSTM 单元由遗忘门、输入门和输出门构成[12]。

LSTM 网络中多个单元形成了一个串联结构，LSTM 的产生是为了避免 RNN 受长时间之前的状态影响而出现的长期依赖问题，但在三个门控单元中，每一单元的计算会丢失上个序列中处理过的信息，所以为了保证信息的完整性和下一时刻输出的准确性，本文提出了一种改进 LSTM 算法用于语音情绪识别，该算法在基础的 LSTM 网络中添加窥孔链接，模型结构如图 2 所示。图 2 中的虚线为添加的窥孔链接，在原有的 LSTM 结构中将上个单元状态 $c^{<t-1>}$ 与遗忘门和输入门做一个连接，将上一个单元的状态加入遗忘门和输入门的计算中，让当前状态不会损失上层状态已经得到的信息，使下一时刻的输出更加具体、完整。

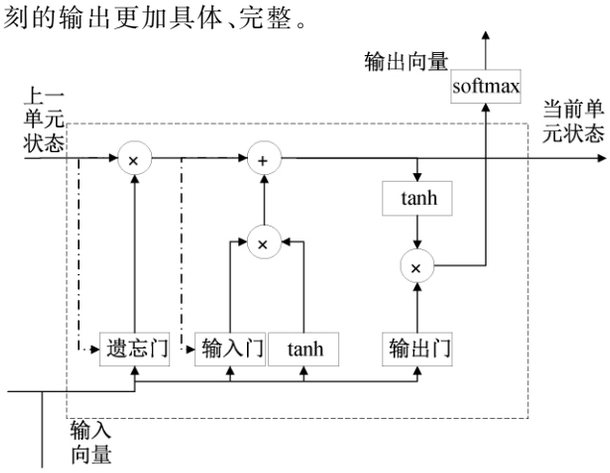

图 2　改进 LSTM 模型整体架构

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

遗忘门自身功能是从细胞状态中选择丢弃什么信息,该门会读取输出向量和记忆单元的输入,通过一个 sigmoid 函数,输出一个区间在[0,1]之间的值,区间两头的值代表完全舍弃和完全保留。本文对 LSTM 进行了改进,在原有的遗忘门基础上,本文将上一单元的状态也加入 sigmoid 函数中,在遗忘门计算时考虑到上一单元状态,使遗忘层的选择性更加准确,避免一部分有效信息被选择遗忘。在 t 时刻遗忘门的公式更新为:

$$F_t=\sigma(\boldsymbol{W}_F[c^{\langle t-1\rangle},\boldsymbol{a}^{\langle t-1\rangle},\boldsymbol{x}^{\langle t\rangle}]+\boldsymbol{b}_F) \tag{2}$$

其中:F_t 表示 t 时刻的遗忘门,c、$\boldsymbol{a}$ 和 $\boldsymbol{x}$ 分别表示记忆单元状态、记忆单元的输出向量和输入向量,$\boldsymbol{W}_F$ 和 $\boldsymbol{b}_F$ 表示遗忘门单元中的权重矩阵和偏置向量。输入门根据上一次得到的输出和当前的输入值进行计算,通过更新程度来控制新状态信息的添加。本文在输入门中同样加入的上一层的单元状态,确保添加到单元状态的信息都是重要的,且不是冗余的,使更新效果更加准确,充分选择哪些数据被更新输入,在 t 时刻输入门和当前状态的公式改进为:

$$I_t=\sigma(\boldsymbol{W}_1[c^{\langle t-1\rangle},\boldsymbol{a}^{\langle t-1\rangle},\boldsymbol{x}^{\langle t\rangle}]+\boldsymbol{b}_1) \tag{3}$$

$$c^{\langle t\rangle}=I_t\times\overline{c}^{\langle t\rangle}+F_t\times c^{\langle t-1\rangle} \tag{4}$$

$$\overline{c}^{\langle t\rangle}=\tanh(\boldsymbol{W}_c[\boldsymbol{a}^{\langle t-1\rangle},\boldsymbol{x}^{\langle t\rangle}]+\boldsymbol{b}_c) \tag{5}$$

其中:I_t 表示 t 时刻的输入门,$\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 表示输入门单元中的权重矩阵和偏置向量。遗忘门和输入门添加上一单元状态后,当前单元状态会随着两个门限层的改良而更加准确,遗忘门和输入门与更新单元同时控制着当前单元状态,从而使当前单元状态的准确性提高,最后单元状态点乘输出门数据所得到的输出结果更加全面。为了复杂度和最后结果考虑,在输出门中不添加窥孔链接。输出门在 t 时刻的公式为:

$$O_t=\sigma(\boldsymbol{W}_o[\boldsymbol{a}^{\langle t-1\rangle},\boldsymbol{x}^{\langle t\rangle}]+\boldsymbol{b}_o) \tag{6}$$

其中:O_t 表示 t 时刻的输出门,$\boldsymbol{W}_o$ 和 $\boldsymbol{b}_o$ 表示输出门单元中的权重矩阵和偏置向量。输出门控制长期记忆对当前输出的影响。LSTM 网络最终的输出由输入门、遗忘门、输出门、上一记忆单元状态和当前记忆单元状态共同决定。

1.3 注意力机制权重计算

注意力机制是根据某一种事物的不同部分的重要程度来计算的一种算法,即为事物的关键部分分配更多的注意力,通过注意力概率分布的计算,对某一关键部分分配更大的权重[13]。

自注意力机制(self-attention)的本质思想可以假设为一种键与值的映射关系,键值查询包含三个基本元素:查询、键和值。通过计算每一个查询项和各个键的相关性得到每个键的对应值的权重系数,然后将权重与对应的键值进行加权求和,在语音信号由 LSTM 输入后,计算每一帧的权重数值,注意力机制的具体计算过程可由图 3 表示。图 3 中:Key、$Query$ 和 $Value$ 分别表示输入特征的关键字、查询值和当前关键字的权重数值;F、Sim 和 a 分别表示计算权重系数的函数、通过计算权重系数得出的相似性和对应键值的权重系数。

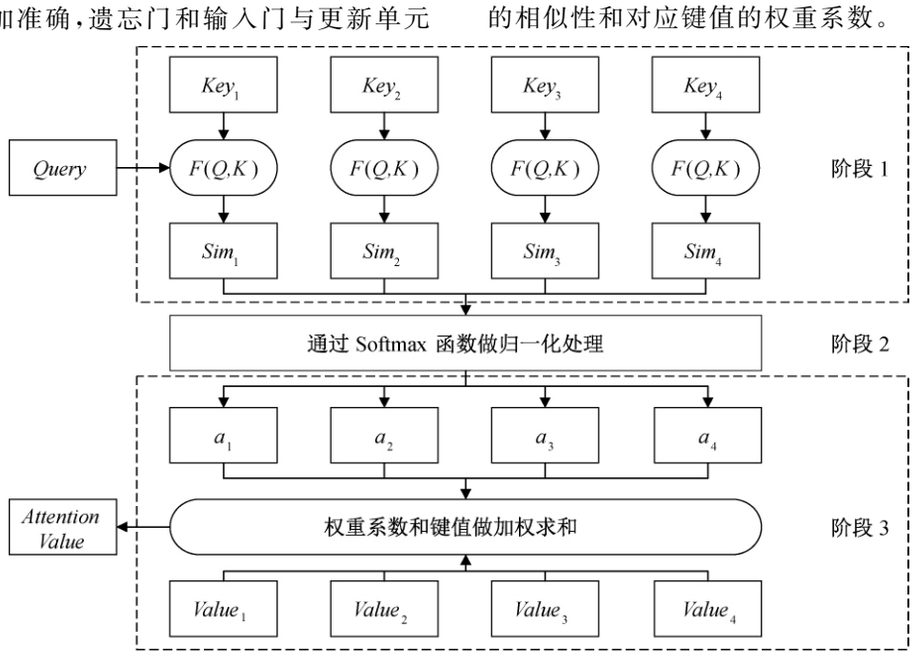

图 3 注意力机制三阶段计算过程示意图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第一阶段通过计算每一个查询值和各个关键字的相关性来得到每个关键字对应数值的权重系数，包括三种计算方式：向量点积法；余弦函数法和引入额外的神经网络求值法，分别由式(7)—(9)所示：

$$Similarity(Query,Keyi)=Query\cdot Keyi \tag{7}$$

$$Similarity(Query,Keyi)=\frac{Query\cdot Keyi}{\|Query\|\cdot\|Keyi\|} \tag{8}$$

$$Similarity(Query,Keyi)=MLP(Query,Keyi) \tag{9}$$

第二阶段即使用类似 Softmax 函数将权值进行归一化处理，如式(10)所示。

$$a_i=\mathrm{softmax}(Sim_i)=\frac{e^{Sim_i}}{\sum_{j=1}^{L_x}e^{Sim_j}} \tag{10}$$

其中：L_x 表示对应的数据源长度。第三阶段，将权重系数和相应的键值做加权求和，从而得到最后的注意力数值，如式(11)所示：

$$Attention(Query,Source)=\sum_{i=1}^{L_x}a_i\cdot Value_i \tag{11}$$

注意力权重计算后，由于语音分割帧数，假设得到 T 帧，每一帧的纬度为 LSTM 网络的神经元个数，所以经过 LSTM 网络后得到的矩阵，形状为 $n*T$，n 为语音每一帧的维度。通过式(12)的计算得到注意力层的编码输出。

$$A=\mathrm{softmax}(g(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}_1)\boldsymbol{W}_2) \tag{12}$$

其中：$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 表示实验中人为调试的最合适的参数矩阵；$\boldsymbol{H}$ 表示由 LSTM 提取出的输入矩阵。最终将注意力机制计算得出的权重值与输入矩阵 $\boldsymbol{H}$ 相乘的结果输入到全连接层，进行分类输出。

2　实验与结果分析

2.1　语音情感数据集

语音情绪识别的准确率与语音情感数据库的质量紧密相关。本文为了有效地验证基于改进 LSTM 的语音情绪识别模型的有效性，选用了 EMO-DB 柏林德语数据库、CASIA 中文数据集和 RAVDESS 北美情感数据集进行实验。

EMO-DB 数据库是由德国柏林工业大学录制的德语语音情感数据库，共 535 句情感语句，包括中性、生气、害怕、高兴、悲伤、厌恶和无聊 7 种情感，语料文本的选取遵从语义中性、无情感倾向的原则，且为日常口语化风格，无过多的书面语修饰[14]，采样率为 48 kHz，采用 16 bit 量化级数据[15]，即使用 16 位的二进制数进行采样。CASIA 中文数据集由中国科学院自动化研究所录制，本文使用了其中 1200 条语句，包括生气、害怕、高兴、中性、悲伤和惊讶 6 种情感，平均一种情感 200 条语句[16]。RAVDESS 北美情感数据集由 24 位专业演员录制，包括中性、无聊、快乐、悲伤、生气、恐惧、惊奇和厌恶 8 种情绪，共 1440 条语句，48 kHz 采样率，采用 16 bit 量化级数据，是情感分类较多的情感语料库[17]。

2.2　数据预处理

本文使用 opensmile 工具进行 MFCC 系数提取，MFCC 提取流程如图 4 所示。

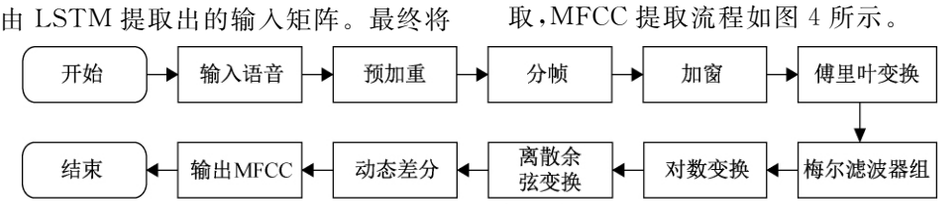

图 4　MFCC 提取流程

首先，将语音信号通过传递函数一阶高通滤波器做预加重处理，增强信号高频成分，对损失的语音信号进行补偿；然后，为保证语音信号的短时平稳性，要进行分帧处理，把语音信号分为若干段，每一个短段称为一帧，以此将帧左右两端的连续性进行提高。帧长设为 2048，设置帧移为 512，将一个时间序列转化为元素部分重叠的帧序列；之后进行加窗处理，基础 MFCC 系数的加窗操作是加海明窗处理，本文由普通的海明窗改为四阶新型海明窗，并进行对比实验。本文增加了余弦的二次谐波分量和四次谐波分量，通过式(13)进行计算：

$$w(n)=\left(a+b\cos\frac{2\pi n}{N-1}+c\cos\frac{4\pi n}{N-1}+d\cos\frac{8\pi n}{N-1}\right)\boldsymbol{R}_N(n) \tag{13}$$

其中：a、b、c、d 为窗函数的定向系数，N 为窗口长度，$\boldsymbol{R}_N$ 为当前窗口长度的矩形窗时域序列。加窗后的语音中的数据逐帧进行快速傅里叶变换(Fast Fourier transform，FFT)，取绝对值的平方后，可以得到每一帧语音信号的频谱；对取得的频谱进行取模运算和平方运算，生成语音信号的功率谱；将能量谱通过一组梅尔尺度的三角形滤波，通常一个滤波器组含有 22～26 个滤波器，来消除谐波的作用，凸

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

显原本语音的共振峰,并降低运算量设置。本文设置梅尔滤波器的个数为128,与能量谱进行点积运算;计算每个滤波器组输出的对数能量后,将计算结果带入离散余弦变换中,改变数据分布将冗余数据分开,低频区集中大部分信号数据,取前20个数据作为实验;然后加上一帧的对数能量,做动态差分参数的提取,计算不同帧下的信号差;最后将全局的语音信号倒谱参数作为静态特征,静态特征的差分谱作为动态特征,通过差分参数的计算方式结合静态特征和动态特征,提高系统的识别性能。

2.3 实验设置与结果分析

本文实验在TensorFlow框架上完成。在特征提取时将梅尔滤波器个数设为128,统一转换语音信号为16 kHz,使用16 bit量化语音信号。LSTM网络隐藏层单元个数设为128,初始学习率设为0.01。为了证明本文的有效性,对以下两组方案进行对比试验。第一组:

a)采用原始完整的MFCC提取方法提取MFCC语音情感特征,预加重后做分帧操作,最后做二阶差分。然后首先用基础LSTM模型进行情感分类识别,LSTM网络后接两个全连接层和softmax层,此语音情绪识别方法为现有语音情绪识别方法,以这个实验作为基准,围绕基础模型进行对比。

b)使用改进LSTM模型进行实验,同样由softmax分类器输出情感类别,区别在于窥孔链接的添加,实验目的是证明增加LSTM各单元的时刻状态是否可以有效提高精度,即改进LSTM模型是否有效。

c)使用LSTM混合自注意力模型,通过注意力机制保留输入序列的长度信息,通过权重进行输出。

d)使用改进LSTM-self-attention做情绪识别对比试验,这一组实验使用改进LSTM模型的同时对每一帧语音的重要程度进行学习,以权重方式进行加权输出。

在对比实验中,本文采用准确率、召回率和F1值作为评价指标,进行不同模型的实验效果评估。第一组方案的实验结果准确率如表1所示,召回率与F1值如表2所示。

表1 方案一不同数据集下准确率实验结果

方案	数据集		
	EMO-DB	CASIA	RAVDESS
LSTM	0.8356	0.8532	0.7930
改进LSTM	0.8429	0.8575	0.8031
LSTM+self-attention	0.8603	0.8729	0.8468
改进LSTM+self-attention	0.8652	0.8798	0.8636

表2 方案一召回率及F1值实验结果

方案	召回率			F1值		
	EMO-DB	CASIA	RAVDESS	EMO-DB	CASIA	RAVDESS
LSTM	0.7645	0.7892	0.7379	0.7984	0.8200	0.7645
改进LSTM	0.7794	0.8003	0.7585	0.8099	0.8279	0.7802
LSTM+self-attention	0.7960	0.8256	0.8027	0.8269	0.8486	0.8242
改进LSTM+self-attention	0.8118	0.8321	0.8186	0.8376	0.8553	0.8405

由表1的对比实验得出,上述的改进LSTM语音情感识别方法优于原始LSTM语音情感识别方法,因为在输入门和遗忘门中加入了窥孔链接,使信息接收更加完整准确,对状态的提取更加充分。在每一单元层的计算中,上一层的单元状态输入输入门和遗忘门,语音信号接收更加完整,从而提高了性能。LSTM与注意力机制结合的模型结果优于单一LSTM模型,每一帧语音信号权重的计算对提高语音情绪分类的准确性有更大的帮助。改进LSTM与自注意力机制结合的模型最终识别率最高,由于本身改进LSTM对信息的充分提取,再通过帧数分割后输出到注意力机制中,作为整个自注意力层的输入,在注意力层中学习到每一帧注意力信号的重要程度,并对输出进行加权计算,由于改进LSTM输出信息的更加具体,所以每一帧语音的权重更加准确。在三种数据集中,RAVDESS数据库的识别率最低,CASIA情感库的识别率最高,证明在多种类语音情绪识别中,相差较近的情感状态很难区分,越多种情绪分类其相近的语音信号特征就越多,语音情绪的多样性和语音情绪间的差异性在一定程度上影响着语音情感识别的性能。RAVDESS数据集整体准确率提高较大,表明改进LSTM与自注意力机制在每一帧权重上的计算对多种情绪分类有着更好的效果,通过语音信号的每一帧权重计算,可以更好地反映多种情绪间的差异,在多情绪计算上有更好的分辨率。

由表2的对比实验得出,本文提出的语音情绪识别方法对召回率有一定的提高,由于改进LSTM模型对于上下层信息的学习更加全面,并且自注意力机制可以加强语音信号中关键部分的影响。使召

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

回率和 F1 值优于基础 LSTM 模型，这表明本文提出模型可行、有效。

为了研究不同特征提取方法的改变对实验结果的影响，证明本文提出的四阶海明窗 MFCC 系数的有效性，设计第二组对比实验：

a）采用基础 MFCC 系数提取方法做数据处理，然后使用改进的 LSTM-自注意力模型进行测试训练。

b）采用四阶海明窗改进后的 MFCC 系数提取方法提取 MFCC，使用改进的 LSTM-自注意力模型进行训练测试。

第二组实验中采用准确率、召回率和 F1 值作为评价指标。结果如表 3 所示。

表 3　方案二准确率、召回率及 F1 值实验结果

方案	准确率			召回率			F1 值		
	EMO-DB	CASIA	RAVDESS	EMO-DB	CASIA	RAVDESS	EMO-DB	CASIA	RAVDESS
基础 MFCC 系数	0.8652	0.8798	0.8636	0.8118	0.8321	0.8186	0.8376	0.8553	0.8405
四阶海明窗 MFCC 系数	0.8867	0.8875	0.8758	0.8287	0.8392	0.8302	0.8567	0.8626	0.8524

由表 3 的对比实验得出，上述的四阶海明窗加窗方法得到的 MFCC 系数，对语音情绪识别的准确率、召回率及 F1 值有一定的提高。通过加窗函数的改变，使语音信号在加窗操作后主要信号频率更加突出。四阶海明窗处理后的信号经过 FFT 和滤波器组，得到的 MFCC 系数更加准确。将特征提取的倒谱向量输入分类器模型，由于主频信号突出，通过 LSTM 模型的语音特征更加准确，更利于注意力机制学习到重要程度较高的语音信号，语音信号通过分类识别，最终得到的实验结果整体略高于使用基础 MFCC 系数的实验结果。实验证明，特征提取的改进对语音情绪识别的准确率、召回率和 F1 值有一定的影响。

在第二组实验中，MFCC 提取过程中加海明窗的语音信号与四阶海明窗的语音信号幅度对比如图 5 所示。由图 5 所示，由于新型四阶海明窗的信号主瓣宽度增加，衰减更稳定，阻带效果衰减也优于海明窗，从而加强了主瓣的作用，提高了频谱幅值精度，更适合计算频段不同，且频段上的各频率成分的贡献值不同的问题，而且对于频率的分辨率不是很看重。由于新型加窗使主瓣信息更加完整，旁瓣性能更好，从而经过三角滤波后效果良好，通过加窗使中间的数据完全体现，在窗口移动时，两侧的一帧或两帧信号重新得到体现，同时改善频率泄漏的问题，倒谱向量提取更准确。

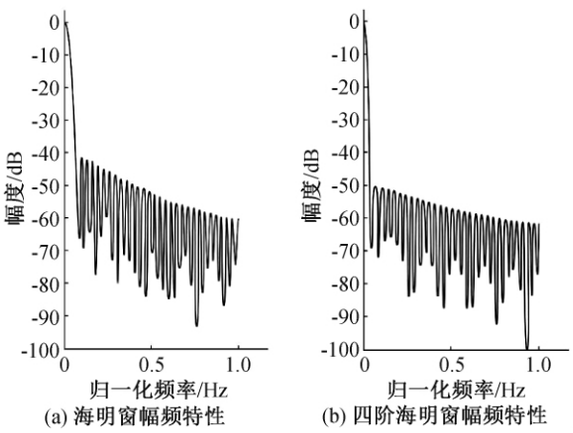

图 5　海明窗与四阶海明窗加窗后的幅频特性曲线

3　结束语

本文提出了一种基于注意力机制与 LSTM 的语音情绪识别方法，使用窥孔链接将上一单元状态信息加入遗忘门和输入门中，将整个 LSTM 的输出作为注意力层的输入，通过权重计算后输入到全连接层及 softmax 层，并在预处理过程中对 MFCC 系数提取进行了一定的改进。为了验证本文提出的方法在语音情绪识别中可以有效提升识别效果，分别从两个方面进行对比实验。首先在相同预处理条件下，以准确率、召回率和 F1 值作为评价标准，将本文模型与其余三种模型进行对比，从这三个评价标准的对比中可以看出，本文模型的语音情绪识别的准确率和召回率更高，分类更加准确；接着对比了本文提出的四阶海明窗 MFCC 预处理方法和基础 MFCC 预处理方法的实验结果，从三种评价标准的对比结果可以看出，本文提出的预处理方法情绪识别结果更加准确；并通过幅频特性曲线对比得出，本文提出的预处理方法得到的语音信号幅值精度更高，信号衰减更稳定，能够更好地提取出语音每一帧的倒谱向量。通过这两组对比实验能够看出本文提出的基于注意力机制与 LSTM 的模型在语音情绪识别上效果较好，能有效提升语音情绪识别的准确率。

参考文献：

[1] 韩文静，李海峰，阮华斌，等. 语音情感识别研究进展综述[J]. 软件学报，2014，25(1)：37-50.

[2] 余伶俐，蔡自兴，陈明义. 语音信号的情感特征分析与识别研究综述[J]. 电路与系统学报，2007，12(4)：76-

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

83.

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

[4] Bourlard H, König Y, Morgan N, et al. A new training algorithm for hybrid HMM/ANN speech recognition systems[C]//1996 8th European Signal Processing Conference. Trieste, Italy: IEEE,1996:1-4.

[5] Li L F, Zhao Y, Jiang D M, et al. Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition [C]//2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland: IEEE, 2013: 312-317.

[6] Kipyatkova I. LSTM-based language models for very large vocabulary continuous Russian speech recognition system[M]//Speech and Computer. Cham: Springer International Publishing, 2019: 219-226.

[7] Zhang Y Y, Du J, Wang Z R, et al. Attention based fully convolutional network for speech emotion recognition [C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775.

[8] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal Processing Magazine, 2001, 18(1): 32-80.

[9] 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. Chennai: IEEE, 2017: 2257-2260.

[10] Zhang X, Chen M H, Qin Y. NLP-QA Framework based on LSTM-RNN[C]//2018 2nd International Conference on Data Science and Business Analytics (ICDSBA). Changsha: IEEE, 2018: 307-311.

[11] 温正棋, 刘斌, 张大伟. 解读"智能交互"的核心技术[J]. 人工智能, 2018, 5(1): 60-75

[12] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10):2222-2232.

[13] 曾义夫, 蓝天, 吴祖峰, 等. 基于双记忆注意力的方面级别情感分类模型[J]. 计算机学报, 2019, 42(08): 1845-1857.

[14] 韩文静, 李海峰. 情感语音数据库综述[J]. 智能计算机与应用, 2013, 3(1): 5-7.

[15] Vlasenko B, Schuller B, Wendemuth A, et al. On the influence of phonetic content variation for acoustic emotion recognition.[M]//Lecture Notes in Computer Science. Heidelberg: Springer Berlin, 2008: 217-220.

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

[17] Livingstone S R, Russo F A, Joseph N. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English[J]. PLoS One, 2018, 13(5): e0196391.

(责任编辑:康 锋)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net