杨琦, 杨芳艳, 袁野, 等. 基于多层次语音情绪识别网络的机器人表情控制[J]. 智能计算机与应用,2024,14(10):41-49.
DOI:10.20169/j.issn.2095-2163.241005

基于多层次语音情绪识别网络的机器人表情控制

杨 琦[1,2], 杨芳艳[2], 袁 野[1], 王佳琦[1,3]
(1 上海理工大学 机器智能研究院, 上海 200093; 2 上海理工大学 机械工程学院, 上海 200093;
3 上海理工大学 健康科学与工程学院, 上海 200093)

摘 要: 面部表情与头部姿态是仿人机器人表达情绪的重要途径,精准的情绪识别与流畅的表情动作对于提升人机交互体验非常关键。为了满足上述要求,本文首先提出了一种基于跨越注意力与多层次声学集成学习的语音情绪识别算法,然后在自研仿人机器人平台上部署该算法,实现了高仿真的人机交互。具体地,研究搭建了包含 16 个伺服位置舵机且拥有高仿真表情和多自由度头部姿态的仿人机器人,基于对关节角度的插值算法与轨迹规划,实现人机交互过程中的机器人面部表情的柔顺控制。此外,研究构建了基于跨越注意力与多层次声学集成学习的语音情绪模型,该模型首先使用深度卷积网络对多源音频信号进行特征提取,再将多种特征进行跨越注意力机制特征融合,解决了频域信息问题和其维度较高导致的维度含义不清晰的问题。实验结果表明,本文提出的方法比现有其他方法具有更好的性能,结合仿人机器人平台能够实现高仿真的人机情感交互。

关键词: 跨越注意力; 多层次声学; 语音情绪识别; 深度卷积网络; 插值算法
中图分类号: TP241 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0041-09

Robot facial expression control based on multi-level speech emotion recognition network

YANG Qi[1,2], YANG Fangyan[2], YUAN Ye[1], WANG Jiaqi[1,3]
(1 Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
3 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Facial expressions and head posture are important ways for humanoid robots to express emotions. Accurate emotion recognition and smooth facial expressions are crucial for improving the human-computer interaction experience. To meet the above requirements, this article firstly proposes a speech emotion recognition algorithm based on cross attention and multi-level acoustic ensemble learning, and then deploys the algorithm on a self-developed humanoid robot platform to achieve high simulation human-machine interaction. Specifically, the paper builds a humanoid robot that includes 16 servo position servos and has high simulation expressions and multi degree of freedom head posture. Based on interpolation algorithms for joint angles and trajectory planning, the paper achieves smooth control of robot facial expressions during human-machine interaction. In addition, the paper constructs a speech emotion model based on cross attention and multi-level acoustic ensemble learning. This model firstly uses deep convolutional networks to extract features from multi-source audio signals, and then fuses multiple features across attention mechanisms to solve the problem of frequency domain information and unclear dimensional meanings caused by high dimensionality. The experimental results show that the proposed method has better performance than other existing methods, and combined with a humanoid robot platform highly simulated human-machine emotional interaction could be achieved.

Key words: crossing attention; multi-level acoustics; speech emotion recognition; deep convolutional network; interpolation algorithm

作者简介: 杨 琦(1998-),男,硕士研究生,主要研究方向:语音情绪识别; 杨芳艳(1979-),女,博士,副教授,主要研究方向:电路与系统; 袁 野(1994-),男,博士,讲师,主要研究方向:类脑计算,计算神经科学。
通讯作者: 王佳琦(1999-),男,硕士研究生,主要研究方向:类别不平衡学习。Email:1014102811@qq.com

收稿日期: 2023-06-06

0 引 言

仿人机器人是集感知能力、学习思考能力、交互能力、运动能力等于一身的一种高度智能化的类人机器人。目前仿人机器人已经能应用在课程教学、导游指引、戏剧表演等场景,并随着科技的发展逐步地融入到人们的日常生活中。因此,人与机器人之间的自然互动已经成为类人机器人在社会环境和医疗保健中的关键问题。相关研究指出,机器人的非语言行为例如:面部表情和头部动作能提供更好的交互体验[1]。为增强这类非语言行为和语言行为之间的耦合关系,本文将语音情绪变化与仿人机器人的舵机柔顺控制相结合,以实现语音对实物仿人机器人控制系统。

近年来,已经有根据语音生成虚拟人的表情研究,利用深度学习模型根据语音生成数字人表情特征、嘴型变化和头部姿态[2]。然而,实物机器人和虚拟数字人之间有着巨大差别。由于虚拟数字人的动作表现不受机械结构的约束,能够更加自由地调节每个像素点的状态,并且在表情变换过程中不存在表情不连贯以及运动速度过快的问题。由于伺服控制的表情机器人内部空间有限,所以对面部区域进行像虚拟人物一样精确控制是不现实的。因此,对实物表情机器人而言仍然是一个具有挑战性的任务。

除了机械结构存在局限性外,语音特征提取情绪的准确性同样重要,情绪的准确读取决定了机器人表情的相关度。在多模态的语音直接驱动表情模型中,Zhou 等学者利用语音和视频进行表情驱动训练,模型生成的表情动作与训练视频中的人物面部表情有较强的耦合关系,并且无法表达语音本身的情绪变化[3]。在结合文本进行语音情绪识别的多模态模型中,会不可避免地出现由于文本内容识别错误而影响最后情绪识别的问题。因此,仅从音频信息提取情绪特征变化也成为一个值得探索的问题。

与以往的根据基础常规声学特征研究为输入,生成分段级情绪状态概率分布的方式相比,利用人工特征 MFCC 的语音数据处理所提取的特征,可以更充分地利用语音本身的特征信息进行情绪分类[4]。在频域方面,Mustaqee 等学者[5]采用基于 RBFN 的聚类方法选择一个关键序列,通过对音频进行快速傅里叶变换(STFT)后的复数求取绝对值获得的语谱图,经过卷积网络后对其进行归一化传入双向长短期记忆网络(BiLSTM)学习时序信息并预测最终情感标签。Wu 等学者[6]对语谱图的图像本身进行卷积以获得时域和频域特征,将特征拼接后通过卷积网络以及注意力池化技术在 IEMOCAP 情感分类准确率上获得提升。在 MFCC 特征提取方面,陈巧红等学者[7]提出利用单个 Mel 频率倒谱系数(MFCC)特征进行混合分布注意力机制与混合神经网络的语音情绪识别方法。Atila 等学者[8]与 Ahmed 等学者[9]都是利用卷积神经网络(CNN)、长短期记忆(LSTM)模型和门控制单元(GRU)对语音信号的全局长期特征进行补充,以提高系统的识别性能。与利用单个特征进行混合神经网络的方法相比,Sun 等学者[10]使用并行的交叉和自注意模块来显式地建模 2 种特征信息的模态间和模态内交互。Zou 等学者[11]通过实验证明,整合不同特征的声学信息,将语音情感问题转换成多层次的融合问题的方法,是充分利用音频信号的有效方式。本文结合多层次语音特征的融合,将语音信号分段处理来获取情绪系数,对表情变化进行基本规划。利用舵机关节角度的插值算法与轨迹规划,对舵机运动进行柔顺处理,实现机器人表情动作、嘴部闭合以及语音播报的同步。

1 相关工作

情感互动能力是仿人机器人在人机互动中的关键。Macdorman 等学者[12]提出具有拟人化外观和面部表情的机器人能够适应人类的惯性思维,人类可以从类人机器人的非语言行为中理解相应的状态。因此,仿人机器人自然的面部表情和头部动作是实现有效地人机交互的必要条件,对于机器人情绪表达行为,将这些非语言行为与语音中情绪相结合有助于实现人机交互的自然性和有效性。

本文提出基于交叉注意力的多层次声学集成学习的语音情绪识别方法,相较于 Co-attention[11]将语谱图经过 AlxNet 处理,以及 MFCC 的频谱经过 BiLSTM 特征提取后直接与 Wav2Vec2.0 进行特征提取。本文为保留语谱图和 MFCC 频谱在情绪识别中的频域特征,结合跨越注意力机制将时序特征融合。利用集成学习分别对语谱特征和 MFCC 频谱的特征进行卷积特征提取,能更好地解耦合,减少相互间的线性关系,保证特征独立性,最大化利用特征进行分类任务。对语音进行分段处理后,将上述模型提取情绪系数作为依据,在语音时间轴上进行表情组合,预设表情动作基本规划以及运行时间。为了

避免由于舵机启动和结束过快出现机器人表情动作僵硬，而带来恐怖谷效应，因此利用插值对舵机运动轨迹进行处理，使得动作变化达到平滑程度更符合人类动作。本文的主要工作总结如下：

（1）制作了一款拥有柔软皮肤、伺服控制技术和精密机械结构高仿真表情机器人，实现了人脸表情、嘴巴闭合和头部动作。通过伺服控制技术，机器人头部平台可以实现主要的情绪表达和基本的交互。

（2）提出了一种仅仅通过语音信号进行情绪识别的深度神经网络。该网络保证了从语音中的情绪变化反应到仿人机器人的表情行为的实时变化，对表情变化进行基本规划，实现了高度的重合性。

（3）提出了一种结合人脸运动特征的仿人机器人头部伺服运动的控制方法，能更好地反映面部表情变化的细节，提高表情动作柔顺性，避免了出现由于动作过程过于机械而引起不适的情况。

2　方法

本文以仿人机器人为平台，研究其依据语音信号情绪进行面部表情变化和头部动作的过程。图 1 为语音驱动表情机器人整体流程，显示了所提方法的基本流程框架。利用语音信号将其分割，经过深度学习模型获取分段后情绪规划。依据情绪动作和语音文字播报时刻，解析出舵机基本运动曲线，再利用插值算法进行轨迹规划，最终实现仿人机器人的随着语音播报产生相应情绪变化的行为。

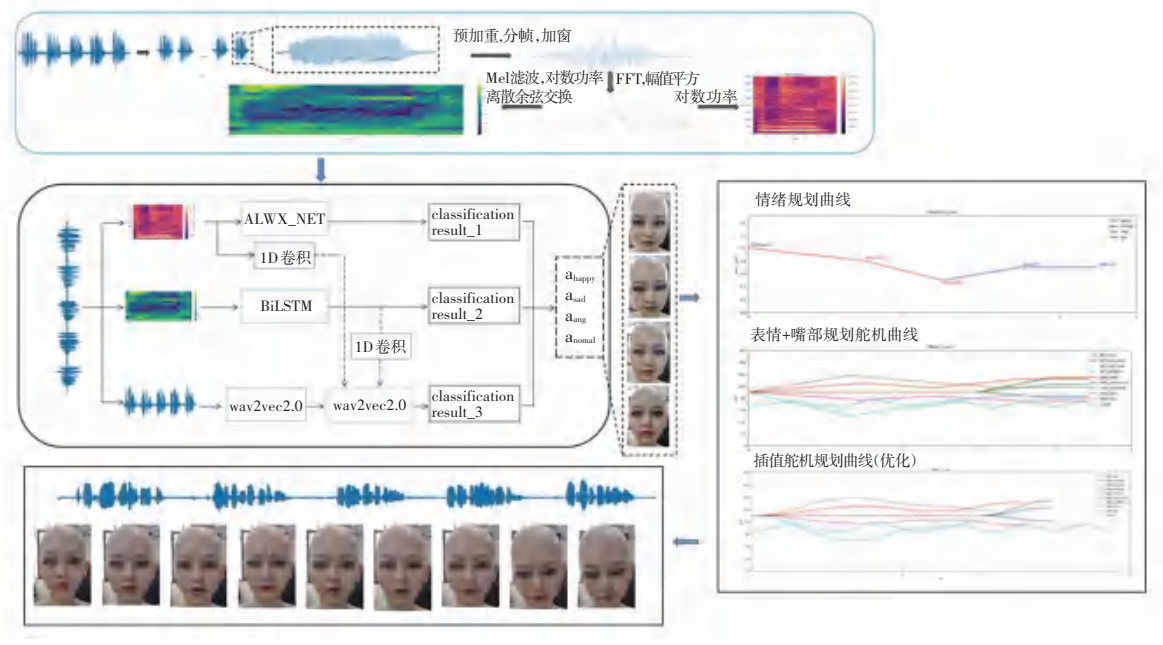


图 1　语音驱动表情机器人整体流程

Fig. 1　Overall process of voice driven facial robot

2.1　机器人头部平台的设计与控制

为了研究人与机器人之间的自然互动，设计开发了一个拟人化的机器人平台来实现随着语音播报产生相应情绪变化的行为。机器人头部平台利用其柔软的皮肤、微处理器、先进的伺服控制系统和精密的机械结构，可以模仿人类的面部肌肉动作和颈部姿态，显示多种面部表情和头部动作。根据语音信号提取语音中情绪特征点，计算出表情变化所需基本舵机的伺服位移，并将伺服位移通过串行接口发送给 STM32F103 微处理器。微处理器通过插值处理优化控制流程并且驱动相应的伺服系统，使机器人头部呈现出相应情绪变化所对应的行为的面部表情及与此相关的头部动作的行为。

2.1.1　仿人机器人平台设计

该表情机器人头部平台分为头部框架、内部舵机控制模块和颈部头部姿态模块。头部框架是根据真实的人脸 3d 打印后外部紧密贴合柔软硅胶皮肤，构成基本的人头外表面模型，头部内部舵机控制结构如图 2 所示，机械控制结构由尼龙绳结构和连杆机构组成。一对半球连杆机构用来控制眼睑的打开和关闭，上下眼睑相互远离的角度范围在 50° ~ 80°。眉毛紧缩和舒展、脸颊的凹陷和凸出，都是利用舵机通过拉动不同的尼龙线而变形的。颈部模块通过 4 个伺服的协同控制实现 6 个自由度的旋转，驱动头部位姿。

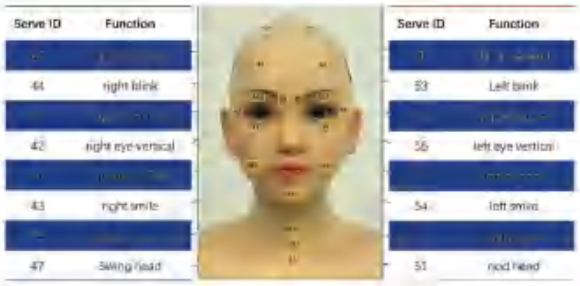


图 2 机器人头部平台机械结构

Fig. 2 Mechanical structure of robot head platform

2.1.2 仿人机器人平台控制

头部舵机的最终控制是通过一块 STM32F103 的芯片完成,利用语音模型输出基本表情规划路线,通过输入信号获取当前位置和目标位置的差异,再通过插值计算规划输出对应的 PWM 控制波形以实现舵机速度的加速、匀速、减速效果。表 1 给出了伺服的功能和自由度。机器人头部平台有 14 个自由度,包括控制眼球、嘴巴、脸颊、眼皮、眉毛和颈部的 3 个旋转角度。并且左右的控制单元都是可以单独运动或者同步运动,尽管所提出的仿人表情机器人的面部控制单元比人类少,但已可以表现出面部表情和头部各方向旋转的主要特征,并组合成稳定和丰富的情感表情动作,为模型实现提供可靠的实验平台。

表 1 16 伺服的功能和自由度

Table 1 16 functions and degrees of freedom of servo

舵机编号	动作单元	自由度
s1,s2	皱眉,挑眉	2
s3,s4	睁眼,闭眼	2
s5,s6	眼球上下	2
s7,s8	眼球左右	2
s8,s9	脸颊上提	2
s10	嘴部闭合	1
s11,s12,s13	摆头,摇头,点头	3

2.2 语音情绪识别深度学习模型

利用交互时播放语音的情绪变化获取表情系数,因此本文构建仅需要语音信号就可以获取准确的情绪及系数。该模型利用不同层次的音频特征信息,通过跨越注意力机制和集成学习将多层次特征信号进行整合获得不同音频的情绪。为更好地利用音频信号,本文在数据预处理和网络结构方面进行优化。

2.2.1 音频处理

本文采用数据集 IEMOCAP,这是一个广泛使用的情绪识别数据集,由 10 个不同的演员的 5 次会话组成,包括音频、视频和动作捕捉信息[13]。在训练过程中本文将把“快乐”和“兴奋”合并到“快乐”的类别中。因为这 2 种情绪在激活和价域上很接近,最终使用的数据集是由愤怒、悲伤、快乐和中性 4 种情绪组成的一共 5 531 个自由表达声学话语集,所使用的原始音频信号以 16 kHz 进行采样。研究中把每一段音频分成多个片段,长度为 3 s。当一个片段小于 3 s 时,将对该片段应用 0 的填充操作,以保持相同的长度。音频话语的最终预测结果将由来自该话语的所有分割片段决定。语音信号处理过程如图 3 所示。

2.2.2 语谱图和 MFCC

为了充分利用语音信息,本文在模型训练中需要用到音频信号的语谱图进行 AlexNet 处理,MFCC 梅尔频率倒谱系数用于双向长短期记忆网络特征提取[14]以及对原语音信号进行 Wav2Vec2.0 特征[15]提取后再做交叉注意力机制的特征加权。本文使用了 Librosa 库进行语谱图以及 MFCC 梅尔频率倒谱系数处理,提取步骤如下:

(1)对分割好的长度为 3 s 原始音频进行预加重,使用了 Librosa 库中的 *preemphasis* 函数,对输入信号语音信号进行预加重处理,需要用到的公式为:

$$x_{\text{pre}}[n] = x[n] - \alpha \times x[n-1] \tag{1}$$

其中,$x_{\text{pre}}[n]$ 表示输出信号;$x[n]$ 表示输入信号;α 表示预加重系数,取值 0.97。

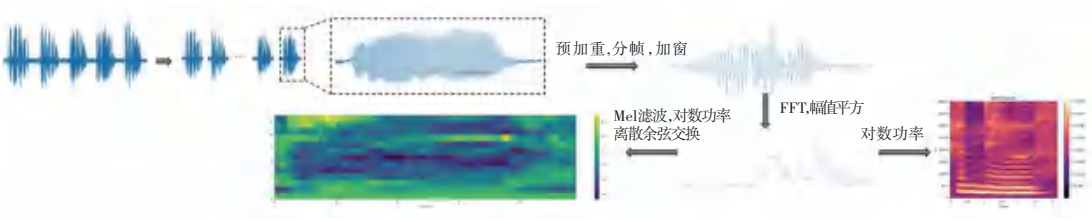


图 3 语音信号预处理过程

Fig. 3 Preprocessing process of speech signal

(2)再对音频进行分帧、加窗处理。应用窗长为 40 ms 汉明窗(Hamming)公式以帧移动 10 ms 得到帧序列,数学公式具体如下:

$$w[n] = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

$$x_{\text{frame}}[n] = x_{\text{pre}}[nR; nR + L - 1] \tag{3}$$

$$x_{\text{win}}[n] = x_{\text{pre}}[n] \times w[n] \tag{4}$$

其中, $w[n]$ 表示窗函数; n 表示窗口长度、取为 40; R 表示帧移,取值 10。

(3)对每个窗口中的加窗后的语音信号进行长度为 800 快速傅里叶变换,得到其频域信息。数学公式如下:

$$X[k,i] = \sum_{n=0}^{N-1} x_{\text{win}}[n] \times e^{-j2\pi\frac{kn}{N}} \tag{5}$$

其中, X 表示频域信息; n 表示时间序列上的采样点; k 表示频率通道索引; i 表示帧索引,即在频域中使用前 200 个频率点来表示信号的频谱。对于每个复数频谱值,进行变换处理转换为分贝,最终得到了每个音频片段大小为 300×200 的声谱图图像。数学公式可写为:

$$spec[k,i] = 20\log_{10}\frac{|X[k,i]|}{ref} \tag{6}$$

(4)将频谱图通过梅尔滤波器组转换成 Mel 频谱图、对数压缩和离散余弦(DCT)。计算公式如下:

$$M[m,i] = \sum_{i=1}^{N} H[m,k] \cdot |X[k,i]|^2 \tag{7}$$

$$M[m,i] = \log M[m,i] \tag{8}$$

$$C[m] = \sum_{n=1}^{N} \alpha(m) \cdot M[n] \cdot \cos[\frac{\pi}{n} \cdot m(n - \frac{1}{2})] \tag{9}$$

其中, $M[m,k]$ 表示经过对数压缩后的 Mel 频谱图的值; $H[m,k]$ 表示第 m 个 Mel 滤波器在第 k 个频率上的响应; $X[k,i]$ 表示频谱图的值; $C[m]$ 表示第 m 个 MFCC 系数; $M[n]$ 表示对数压缩后的 Mel 频谱图的值; $\alpha(m)$ 表示归一化系数; N 表示 DCT 变换的长度。通过上述步骤,得到模型所需的语谱图和 MFCC 特征,与语音通过 Wav2Vec 产生的特征进行交叉注意力的计算研究。

2.2.3　深度网络设计

本文对语音信号、梅尔频率的频谱特征和语谱图进行集成学习。相较于常见的直接将各个特征进行拼接后直接用于分类任务相比,将不同的特征放到分类器里,能更好地解耦合,减少特征之间的线性关系,保证特征独立性,从而最大化利用特征实现分类任务。研究中给出的多层次交叉注意机制语音情绪识别模型如图 4 所示。

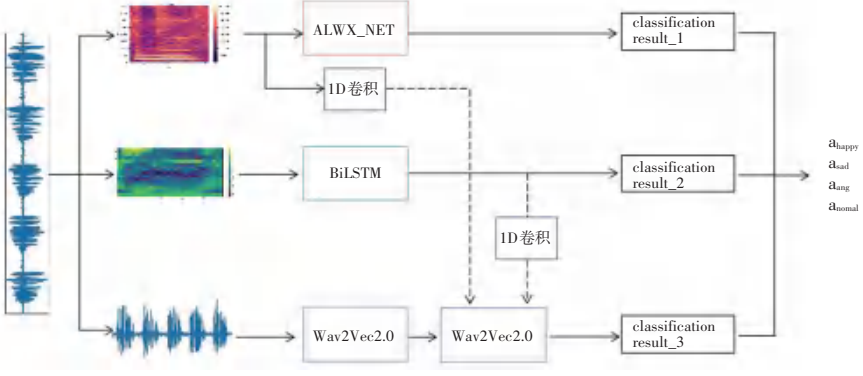


图 4　多层次交叉注意机制语音情绪识别模型

Fig. 4　A multi-level cross attention mechanism speech emotion recognition model

语音信号是典型的时间序列,循环神经网络(RNN)也常用于对语音信号在时间序列的特征提取。本文引入的双向长短期记忆网络(BiLSTM),是由 2 个能实现记忆长序列的长短期记忆网络(LSTM)组成[16]。该网络能充分考虑上下文信息,提高语音任务的性能。通过 MFCC 系数获取频谱特征向量,并在语音处理中获取(300,40)的特征张量。其中,MFCC 系数是基于人耳的听觉特性在 Mel 频率域提取的倒谱特征参数。该频域特征张量结合双向长短时记忆模型,能有效地提取语音情绪在语音信号 MFCC 频谱中的特征,生成(300,512)的特征用于情绪分类从而获得分类结果一。对于语谱图的特征提取引入 AlexNet[17],将语谱图分成 3 份进一步利用语谱图的图片信息经过 AlexNet 网络获取语谱图中的特征向量,获得情绪分类结果二。

除了上述对音频特征的频域信息进行特征提取以外,本文引入 Wav2Vec2.0 和交叉注意力机制,将语音信号放入 Wav2Vec2.0 中,利用交叉注意力机

制获取时序特征。Wav2Vec2.0 是一种新兴的语音特征提取方法,基于自监督学习,通过预训练一个自编码器,将语音信号转换为一组高维度的特征向量[18]。Wav2Vec2.0 可以有效地捕获语音信号的时域信息,其内部表示是高度抽象和复杂的,研究中模型的解释性相对有限。研究中引入跨越注意力机制加强模型的表征能力,通过对不同输入序列中的元素进行关联,获取更全局和综合的信息,同时通过为不同输入序列中的元素分配权重,交叉注意力机制可以自动选择和聚焦于最相关的元素,以利于更有效地传递和整合信息[19]。将 BiLSTM 后的 MFCC 系数特征经过一维卷积遍历,获得(128,128)的特征张量 K。MFCC 本身特征提取过程中借助梅尔滤波器模拟人耳对不同频率声音的感知,将原始的语音信号转换到梅尔频率上,但是抑制部分高频特征信号。为减少所忽略的部分频谱信息对结果的影响,可以利用对语谱图的特征提取来加以弥补。将 128 个一维卷积核大小为 9、步长为 9 的卷积核遍历大小为(256,1 152)的语谱图输出(128,128)的语谱图特征,该特征包含全局语音信号频率与时间的特征 Q。为此利用语谱的特征 Q 和 K 的相乘能够为 Wav2Vec2.0 捕获语音信号的时域 V 特征提供语音在频率上的特征权重。最后,将特征用于情绪分类获得情绪分类结果三。

在训练过程中,将上述结果进行集成学习,能更好地综合多个特征的优势,提高模型的准确性。并且在训练音频输入中添加高斯噪声,能够增强模型的鲁棒性和抗干扰能力,而通过扩充训练数据的多样性则能提高模型的泛化能力,就能更好地处理不同的情绪样本。

2.3 情绪表情映射

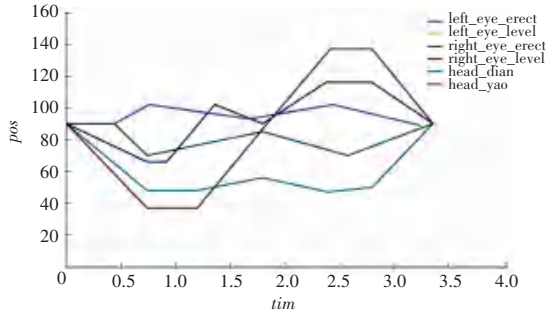
将语音信号进行分段预处理、并通过上述模型获得每段的情绪以及相应系数后,研究将利用预设好的情绪表情舵机位置,对其在时间序列上进行基本的运动规划。为使得表情运动过程更加自然,本文对轨迹进行二次插值样条。计算公式如下:

$$y_n = a_A \times M_A \tag{10}$$

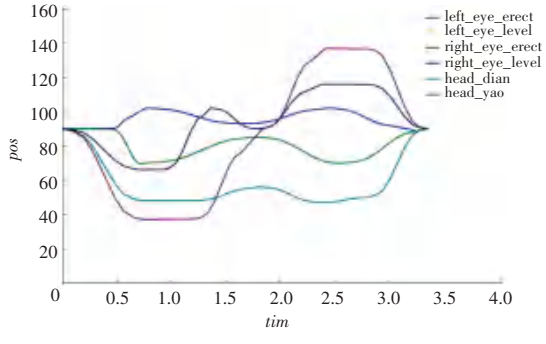
$$x_n = T_M \tag{11}$$

其中,A 表示对应的表情种类;a_A 表示对应表情种类的系数;M_A 表示对应表情种类舵机位置;T_M 表示舵机运动时刻。

这样一来,让运动、表情、动作速度呈现快慢快的运动趋势,使得表情动作柔顺且自然。舵机运动插值的控制曲线如图 5 所示。



(a)插值前舵机控制图



(b)插值后舵机控制图

图 5 舵机控制曲线图

Fig. 5 Control curve diagram of the servo motor

获取一组数据点 (x_i, y_i),其中 $i = 0,1,2,\cdots, n$。本文在每个相邻数据点之间进行三次样条插值。依据自然边界条件 $S''(x_0) = S''(x_n) = 0$,边界条件 $S'(x_0) = f'(x_0), S'(x_n) = f'(x_n)$ 和内部节点处的边界条件 $S''(x_i) \cdot h_i = 3\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right)$,计算样条曲线段 $S_i(x)$,计算公式如下:

$$h_i = x_{i+1} - x_i \tag{12}$$

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{13}$$

$$a_i = y_i \tag{14}$$

$$b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{3}(2c_i + c_{i+1}) \tag{15}$$

$$d_i = \frac{c_{i+1} - c_i}{3h_i} \tag{16}$$

3 实验结果

3.1 评估深度学习方法

本文系统在 PyTorch 中实现。实验使用交叉熵

准则作为训练的损失函数，优化方法是随机梯度下降（Stochastic Gradient Descent，SGD）算法和自适应矩估（Adaptive Moment Estimation，Adam）算法，优化器是 AdamW，学习率为 1e-5。训练批次大小为 64。本文使用五折交叉验证方法，实验使用了 2 个常见的指标来评估模型的性能：

（1）加权准确率（Weighted Accuracy，*WA*）。测试集中的每个句子具有相同的权重，直接计算得到的准确率。

（2）不加权准确率（Unweighted Accuracy，*UA*）：首先计算每种情感的准确率，然后进行平均得到的准确率[20]。

上述评估指标的计算方式如下：

$$WA = \frac{\sum_{i=1}^{L} TP_i}{\sum_{i=1}^{L} (TP_i + FN_i)} \tag{17}$$

$$UA = \frac{1}{L} \sum_{i=1}^{L} \frac{TP_i}{(TP_i + FN_i)} \tag{18}$$

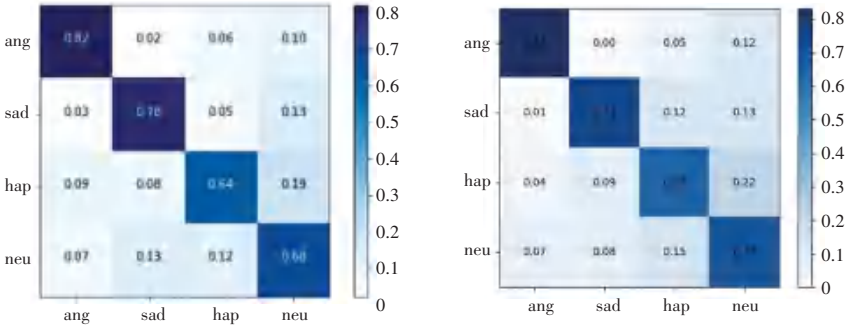
其中，L 表示不同情绪类别的数量，TP_i 表示对 i 类样本预测错误的样本和数量。

表 2 显示了具有不同声学信息组合的模型性能的消融研究，前 3 行是只有一个类型的声学信息的情绪识别结果：MFCC、声谱图和 W2E。分析可知，W2E 在最终的情绪识别方面提供了比其他模型更好的性能。而利用跨越注意力机制进行情绪识别的准确率相较单个特征有较为明显的提升，WA 提升了 6.33%和 UA 提升了 4.09%。本模型是相较 Co-attention 论文中的直接利用数据拼接的方式，对 MFCC 和声谱图特征在经过 1D 卷积后对 Wav2Vec2.0 获取的时序特征引入交叉注意力机制后在 WA 上的性能有所提升。从第 6、第 7 行显示的利用集成学习对多层语音特征进行训练的结果表明，相较于只是简单地将多层语音情绪特征识别结果进行相加融合相比，在训练的过程中将每一层的特征能更好地解耦合。为进一步提升识别率，并改进模型的鲁棒性，本文在输入的音频数据添加高斯噪声，从表 2 的第 9、第 10、第 11 行实验得到信噪比为-6 的高斯噪声，使得最终的模型在准确率上有部分提升。情感识别的归一化混合矩阵如图 6 所示。图 6 中，图 6（a）表示共同注意力机制；图 6（b）表示交叉注意力机制。从最终的归一化混淆矩阵来看，具有交叉注意力机制的模型的最终分类结果比共同注意力机制的模式更好。

表 2　语音情绪识别模型评估结果

Table 2　Evaluation results of speech emotion recognition model

方法	WA	UA
MFCC	57.600 0	58.900
Spectrogram	62.130 0	62.250
W2E	64.030 0	65.670
Cross_attention	70.360 0	70.760
Co-attention	69.800 0	71.050
Ensem+Co-attention	70.544 5	71.172
Ensem+Cross_attention	70.364 9	71.186
Noise+Cross_attention	69.270 0	70.810
Noise(20)+Ensem+Cross_attention	69.792 0	70.126
Noise(-6)+Ensem+Cross_attention	71.058 0	71.561
Noise(-20)+Ensem+Cross_attention	65.630 0	65.580



（a）共同注意力机制　　（b）交叉注意力机制

图 6　情感识别的归一化混合矩阵

Fig. 6　Normalized hybrid matrix for emotional recognition

3.2　驱动框架实现效果

根据上述模型获得语音的情绪参数进行基本的情绪表达规划，并且根据语音说话强度获取嘴部张开闭合的基本运动。为使得仿人机器人表情运动流

畅,再对舵机运动进行插值运动平滑处理,最终实现仿人机器人具有丰富表情的语音交互功能,整体实现效果如图 7 所示。

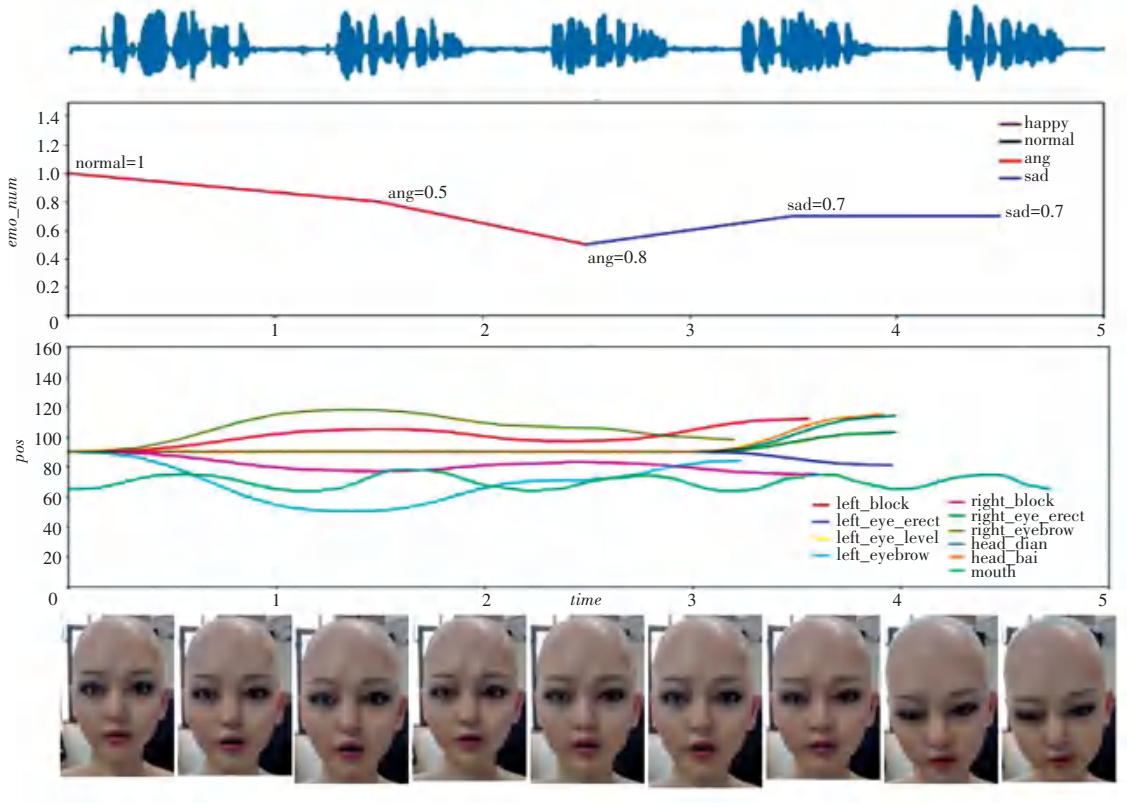


图 7 仿人机器人实物演示

Fig. 7 Physical demonstration of humanoid robots

4 讨论

结合语音进行情绪变化的仿人机器人交互系统,对于人机交互非常重要。然而,仅通过语音信号对实体仿人机器人进行面部控制的研究却仍不多见。部分研究以多模态的语音信息训练模型的主要参数,以及仿真平台为载体,无需考虑硬件控制中表情的变化问题。本文提出一种仿人机器人面部语音控制的方法。致力于探索机器人根据语音情绪进行情感传递,进一步提高机器人的人性化拟人化程度。

本文根据对语音信号多层次的特征提取,充分结合时序和频域特征,相较于其他单层次模型有着更好的准确性。并结合跨越注意力机制和对训练数据的噪声添加,使得模型有更好的鲁棒性能。此外,为使得仿人机器人在人机交互的过程中有更好的流畅性,本文对原本的舵机控制加以改进,利用插值的方法,为机器人的表情切换提供细滑控制指令。本文的研究也存在一些局限性。例如,文中的机器人头部平台还比较落后,有很多改进的余地,特别是在嘴部机械结构方面,多种嘴型的切换配合语音会有更好的效果。此外,仿人机器人语音驱动表情的实时性和自然性对人与机器人自然交互过程中的主观感知极其重要。虽然实验结果表明,仿人机器人在根据语音进行面部表情和头部动作的转换过程中表现良好,但目前的研究仍缺乏对其在实际人机交互场景中的有效性和自然性的评价。

5 结束语

本文提出了基于多层次语音情绪识别模型进行驱动表情行为方法。首先,开发了一个 16 自由度的仿人机器人头部平台,实现了语音驱动面部表情和头部动作的行为;此外,使用一种基于跨越注意力与多层次声学集成学习的语音情绪识别的深度学习框架,提取到不同时刻的情绪特征点。然后,通过对情绪特征点进行运动舵机的规划和优化,建立与最优伺服控制关系,实现语音驱动表情的行为。评价结果表明,该多层次模型能够在精度较高的前提下实现情绪系数识别。未来,将会结合视觉反馈、脑电波

信号反馈等能够反映交互人的情绪变化的信息，将其添加到仿人机器人表情变换的判断中，实现情绪的多模态表达。

参考文献

[1] XIAO Qian, XING Xiaosong, WANG Chengdong. Comprehensive experimental design of interactive facial expression robot[J]. Experiment Science and Technology, 2021, 19(1): 14-19.

[2] ZHOU Hang, SUN Yasheng, WU Wayne, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021: 4174-4184.

[3] LU Yuanxun, CHAI Jinxiang, CAO Xun. Live speech portraits: real-time photorealistic talking-head animation[J]. ACM Transactions on Graphics, 2021, 40(6): 1-17.

[4] JAGADEESHWAR K, SREENIVASARAO T, PULICHERLA P, et al. ASERNet: Automatic speech emotion recognition system using MFCC-based LPC approach with deep learning CNN[J]. International Journal of Modeling, Simulation, and Scientific Computing, 2023, 14(4): 2341029.

[5] MUSTAQEE M, SAJJAD M, KWON S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM[J]. IEEE Access, 2020, 8: 79861-79875.

[6] WU Xixin, LIU Songxiang, CAO Yuewen, et al. Speech emotion recognition using capsule networks[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6695-6699.

[7] 陈巧红，于泽源，贾宇波. 基于混合分布注意力机制与混合神经网络的语音情绪识别方法[J]. 计算机应用与软件，2022，44(12)：2246-2254.

[8] ATILA O, SENGUR A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition[J]. Applied Acoustics, 2021, 182(1): 108260.

[9] AHMED M R, ISLAM S, ISIAM A K M, et al. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition[J]. Expert System with Applications, 2021, 218: 119633-119633.

[10] SUN Licai, LIU Bin, TAO Jianhua, et al. Multimodal cross and self-attention network for speech emotion recognition[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE, 2021: 4275-4279.

[11] ZOU Hengqing, SI Yuke, CHEN Chen, et al. Speech emotion recognition with co-attention based multi-level acoustic information[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7367-7371.

[12] MACDORMAN K F, ISHIGURO H. The uncanny advantage of using androids in cognitive and social science research[J]. Interaction Studies, 2006, 7(3): 297-337.

[13] QI Zhong, FEN Yaqin, WANG Wei. Comparison of speech emotion recognition in cross language corpus[J]. Journal of Nanjing University, 2019, 55(5): 765-773.

[14] KUMBHAR H S, BHANDARI S U. Speech emotion recognition using MFCC features and LSTM network[C]// 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA). Pune, India: IEEE, 2019: 1-3.

[15] ZHAO Zihan, WANG Yanfeng, WANG Yu. Multi-level fusion of Wav2vec 2.0 and BERT for multimodal emotion recognition[J]. arXiv preprint arXiv: 2207.04697v1, 2022.

[16] SUN Bo, WEI Qinglan, LI Liandong, et al. LSTM for dynamic emotion and group emotion recognition in the wild[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16). Tokyo, Japan: ACM, 2016: 451-457.

[17] STOLAR M N, LECH M, BOLIA R S, et al. Real time speech emotion recognition using RGB image classification and transfer learning[C]// International Conference on Signal Processing & Communication Systems. Surfers Paradise, Australia: IEEE, 2017: 1-8.

[18] 邱智乾，陈霏，郎标. 基于循环神经网络的双麦克风语音增强算法[J]. 传感技术学报，2024，37(3)：430-438.

[19] RAJAN V, BRUTTI A, CAVALLARO A. Is cross-attention preferable to self-attention for multi-modal emotion recognition?[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 4693-4697.

[20] AUDHKHASI K, NARAYANAN S S. Emotion classification from speech using evaluator reliability-weighted combination of ranked lists[C]// IEEE International Conference on Acoustics. Prague, Czech Republic: IEEE, 2011: 4956-4959.