

一种面向对抗攻击的鲁棒性语音情感识别方法

陈 港^{1,2}, 陈 杰², 张石清², 赵小明^{1,2}

(1. 浙江理工大学机械与自动控制工程学院, 浙江 杭州 310018;

2. 台州学院智能信息处理研究所, 浙江 台州 318000)

✉904699855@qq.com; 1424179695@qq.com; tczsq@163.com; tzxyzxm@163.com



摘 要: 目前, 现有的语音情感识别研究主要考虑在实验环境下收集语音数据进行情感识别, 并没有考虑现实世界中存在各种噪声的影响。为此, 考虑到噪声的影响, 提出一种面向对抗攻击的鲁棒性语音情感识别方法, 用于实现带有噪声的情感语音的分类。首先采用快速梯度符号法生成对抗数据, 然后将真实数据和对抗数据进行混合, 再将混合数据输入防御模块中进行模型的对抗攻击训练。最后, 在IEMOCAP数据集上的实验结果表明, 该方法用于语音情感识别能有效提高深度学习模型的鲁棒性和识别准确率。

关键词: 语音情感识别; 鲁棒性; 对抗攻击

中图分类号: TP391 **文献标识码:** A

A Robust Speech Emotion Recognition Method for Confrontational Attacks

CHEN Gang^{1,2}, CHEN Jie², ZHANG Shiqing², ZHAO Xiaoming^{1,2}

(1. Faculty of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China;

2. Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China)

✉904699855@qq.com; 1424179695@qq.com; tczsq@163.com; tzxyzxm@163.com

Abstract: At present, the existing research on speech emotion recognition mainly considers speech data collection in an experimental environment for emotion recognition, without considering the influence of various noises in the real world. For this reason, considering the influence of noise, this paper proposes a robust speech emotion recognition method for confrontational attacks to realize classification of emotional speech with noise. Firstly, fast gradient sign method is used to generate confrontation data which is mixed with the real data. Then the mixed data is input into the defense module to conduct confrontation attack training of the model. Finally, experimental results on the IEMOCAP dataset show that the method used in speech emotion recognition can effectively improve the robustness and recognition accuracy of the deep learning model.

Keywords: speech emotion recognition; robustness; confrontational attack

1 引言(Introduction)

语音情感识别技术^[1]一直是人机交互领域中重要的研究热点, 但大量的语音情感识别方法主要考虑在实验环境下对语音进行情感识别, 并没有考虑现实世界中存在各种噪声的影响, 因此, 计算机在噪声的影响下准确地识别出人类的情感^[2], 有助于实现其与人类进行友好的交互。

为了推动语音情感识别技术的发展, 研究者在每年举办一届的INTERSPEECH会议^[3]中进行先进的学术交流和经验分享。最早的语音情感识别方法主要是根据语音信号设计

手工特征^[4-6], 例如音律特征、谱特征、音质特征及特别设计的声学特征集^[7-8], 再将其输入简单的分类器进行相应的情感分类。如NWE等人^[9]提出采用短时对数频率功率系数(Log Frequency Power Coefficients, LFPC)表示语音信号, 再使用离散隐马尔科夫模型(Hidden Markov Model, HMM)作为情感分类器; ALEX等人^[10]提出分别使用韵律特征和频谱特征表示语音信号, 然后使用注意力机制和特征选择方法增强模型的性能, 再将输出的特征进行分数级融合, 最后预测出语音样本的情感类别; DANESHFAR等人^[11]提出使用频谱-韵

律混合特征向量表示语音信号,再使用基于量子行为的粒子群优化(Quantum-behaved Particle Swarm Optimization, QPSO)算法对特征进行降维,最后利用高斯椭圆基函数(Gaussian Elliptical Basis Function, GEBF)型分类器对特征进行情感分类。虽然上述方法具有不错的识别性能,但是手工设计的特征属于低阶特征,表示情感信息的判别能力相对较低。

随着深度学习技术的快速发展,各种深度神经网络模型^[12-14](Deep Neural Networks, DNN)被提出,研究者们开始采用DNN学习语音信号中高阶的深度特征。例如, HUANG^[13]等人提出将语音信号转为频谱图,再利用CNN从频谱图中提取深度语音情感特征,然后使用softmax层进行最后的语音情感分类。ZHAO等人^[15]提出使用1-D CNN-LSTM和2-D CNN-LSTM分别从语音信号和对数梅尔频谱图(Log-Mel Spectrum)中学习局部和全局情感表示,再将局部和全局情感表示进行融合,最后获得情感语音样本的类别。现有的语音情感识别方法主要对在实验环境下采集的语音数据进行情感分类,没有考虑现实环境中噪声的影响。而且,在存在噪声的情况下,大部分训练好的模型性能也会显著下降。针对上述问题,本文提出一种面向对抗攻击的鲁棒性语音情感识别方法。首先采用快速梯度符号法(Fast Gradient Sign Method, FGSM)将真实数据转为对抗数据,该模块由模型中的攻击模块实现;然后将得到的对抗数据与真实数据一并输入防御模块中,并采用对抗训练的方式训练模型;最后对语音样本的情感类别进行预测分类。在IEMOCAP数据集^[16]中的实验结果表明,本文提出的方法能对抗噪声的干扰,并提高语音情感识别模型的准确率。

2 面向对抗攻击的语音情感识别模型(Speech emotion recognition model for confrontational attacks)

图1给出了面向对抗攻击的鲁棒性语音情感识别方法示意图,该模型由攻击模块和防御模块两部分组成。本文方法首先将真实数据输入攻击模块中生成对抗数据;然后将对抗数据和真实数据混合输入防御模块中,以训练出一个具有鲁棒性的语音情感识别模型;最后实现语音情感分类。

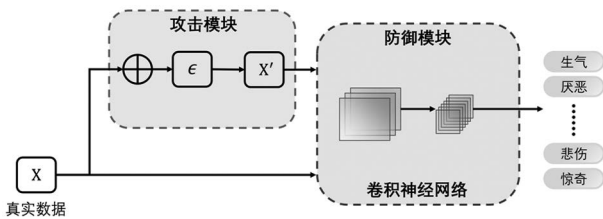


图1 一种面向对抗攻击的鲁棒性语音情感识别方法示意图

Fig.1 The schematic diagram of a robust speech emotion recognition method for confrontational attack

2.1 面向攻击模块的对抗数据生成

相比于简单的语音信号,本文采用Log-Mel频谱图作为攻击模块模型的输入数据,然后使用FGSM生成对抗数据。该方法将计算得到的梯度作为对抗噪声,其中损失函数定义为 $L(X, Y, \omega)$,相应计算的梯度为 $\nabla_x L(X, Y, \omega)$,对抗数据由如

下公式计算得到:

$$X' = X + \varepsilon \cdot \text{sign}(\nabla_x L(X, Y, \omega)) \quad (1)$$

$$X = \text{Chip}(X', X' - \beta, X + \beta) \quad (2)$$

其中, ε 表示一个干扰常量,用于控制噪声添加量; $\text{Chip}(\cdot)$ 表示一种幅值裁剪运算操作; β 表示一个常量,用于控制输入数据的取值范围; $\text{sign}(\cdot)$ 表示符号函数,生成的对抗数据被用于扩增训练数据集,并有助于提高模型的鲁棒性。图2给出了对抗数据生成示意图。例如,某个情感类别为厌恶的Log-Mel频谱图在添加对抗噪声后,输入模型中会被误判为高兴情感。由此可知,本文提出的对抗数据能影响模型的判别性能。

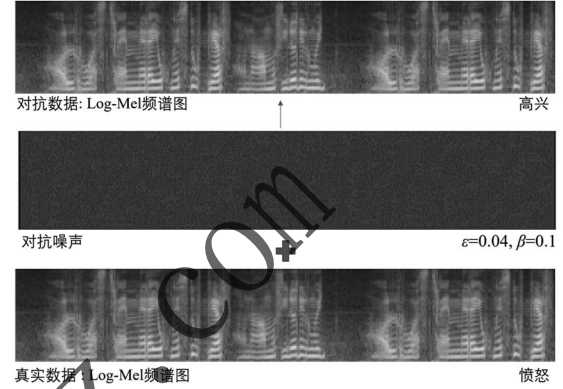


图2 对抗数据生成示意图

Fig.2 The generation diagram of confrontational data

2.2 面向防御模块的卷积神经网络模型

为了提取Log-Mel频谱图中高阶的语音情感特征,本文选取三种深度神经网络模型来验证对抗攻击模型的有效性,三种模型分别为普通卷积神经网络(CNN-5)、VGG-16和ResNet-50。表1给出详细的三种网络结构参数,其中Conv表示卷积层,MP和AP分别表示最大池化操作和平均池化操作, stride表示步长。

表1 CNN-5、VGG-16和ResNet-50的网络结构参数

Tab.1 The network structure parameters of CNN-5, VGG-16 and ResNet-50

CNN-5	VGG-16	ResNet-50
Conv: $[5 \times 5, 64] \times 1$	Conv: $[3 \times 3, 64] \times 3$	Conv: $[7 \times 7, 64] \times 1$
MP: $[2 \times 2]$, stride=2	MP: $[2 \times 2]$, stride=2	MP: $[3 \times 3]$, stride=2
Conv: $[5 \times 5, 128] \times 1$	Conv: $[3 \times 3, 128] \times 3$	Conv: $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
MP: $[2 \times 2]$, stride=2	MP: $[2 \times 2]$, stride=2	Conv: $\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv: $[5 \times 5, 256] \times 1$	Conv: $[3 \times 3, 256] \times 3$	Conv: $\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
MP: $[2 \times 2]$, stride=2	MP: $[2 \times 2]$, stride=2	Conv: $\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Conv: $[5 \times 5, 512] \times 1$	Conv: $[3 \times 3, 512] \times 3$	
MP: $[2 \times 2]$, stride=2	MP: $[2 \times 2]$, stride=2	
	MP	AP
	Fully Connected Layer	

2.3 对抗训练

对抗训练是将对抗数据和真实数据混合后作为训练模型的数据集，从而提高识别模型的通用鲁棒性。相比在真实数据训练下的损失函数，本文提出的对抗损失函数能同时考虑真实数据和对抗数据的损失值。对抗损失函数定义如下：

$$\hat{L}(X, Y, \rho) = \alpha \cdot L(X, Y, \rho) + (1 - \alpha) \cdot L(X', Y, \rho) \quad (3)$$

其中， α 表示真实数据和对抗数据之间的自适应权重； ρ 表示网络模型权重参数； Y 表示真实数据和对抗数据的标签。因此，训练模型的目标是最小化对抗损失值 $\hat{L}(X, Y, \rho)$ 。

2.4 情感分类

本文提出在模型的末端嵌入一个softmax层对输出的特征进行情感分类，softmax的计算原理如下所示：

$$Y' = \text{softmax}(x, W) \quad (4)$$

其中， Y' 表示模型的预测分类结果； $\text{softmax}(\cdot)$ 表示softmax函数； x 表示输入的特征； W 表示softmax层中的权重参数。

3 实验(Experiment)

3.1 数据集

本文采用的语音情感数据集为IEMOCAP，该数据集由5个会话(Sessions)组成，每个会话由一名男演员和一名女演员组成，并且每个会话都由不同的演员参演。该数据集所有音频文件的时长大约有12小时。本文实验仅使用数据集中高兴、悲伤、中性和愤怒4种情感类别样本，并且采用与人无关的实验方式，即使用1个会话作为测试集，其余4个会话作为训练集的5倍交叉验证方式进行实验。IEMOCAP数据集中4种情感类别样本数量分布状况如表2所示。

表2 IEMOCAP数据集中的情感语音样本数量分布
Tab.2 The number distribution of emotion speech samples in IEMOCAP dataset

会话	高兴	悲伤	中性	愤怒
1	278	194	384	220
2	327	197	362	137
3	286	305	320	240
4	303	143	258	327
5	442	245	384	170

3.2 数据预处理

相较于一般的语音信号，Log-Mel频谱图的表示方式更接近人类的听觉系统，因此本文将数据集中语音信号转为Log-Mel频谱图。由于数据集中的音频文件时长不一致，同时，模型的输入是固定的，因此将语音信号输入64阶梅尔滤波器组，并通过重采样和欠采样的方式生成维度大小为 373×64 的Log-Mel频谱图，并标注相应的情感类别标签，以训练识别模型。

3.3 评价指标

为了使模型快速收敛并保持模型的训练稳定性，初始学习率设置为0.001，每迭代100次后降低到当前的学习率的90%，迭代训练到10,000次结束模型训练。为了说明对抗数

据能降低模型的识别准确率，本文实验中的干扰常量 ε 依次设置为0.0、0.02、0.04、0.06、0.08和0.1，超参数 α 设置为0.5。本文中的模型性能评估标准为无权重平均召回率(Unweighted Average Recall, UAR)，由如下公式计算得到：

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$UAR = \frac{\sum_{i=1}^N Recall_i}{N} \quad (6)$$

其中， TP_i 和 FN_i 分别表示第 i 类的真正例和假反例，即预测该类正确和错误的数量； $Recall_i$ 表示第 i 类的召回率，即在所有该类的样本中被预测正确的比例； UAR 表示无权重平均召回率，即对所有类的召回率取算数平均。

3.4 实验结果及分析

为了验证FGSM的有效性，首先在真实数据上训练得到模型准确率的基线，再将训练好的模型在生成的对抗数据上进行测试。各个CNN模型在IEMOCAP数据集上的准确率随着对抗攻击的增强而持续下降，如图3所示。由图3可知，干扰常量 ε 的增加确实能降低模型的准确率，其中 $\varepsilon = 0.0$ 表示在真实数据上进行的测试。随着 ε 的增加，UAR也随之减少，如当 $\varepsilon = 0.06$ 时，各模型的准确率已下降到20%以下，因此，本文中的对抗数据对模型的攻击是有效的。

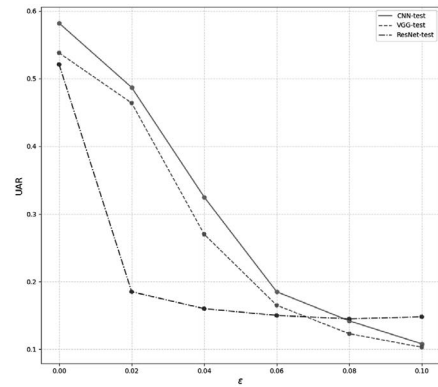


图3 添加对抗攻击时，通过单一训练获得的CNN模型性能表现

Fig.3 The performance of the CNN model obtained through a single training when adding confrontational attacks

表3列出了CNN-5、ResNet-50和VGG-16模型在IEMOCAP数据集的性能比较。由表3可知，在IEMOCAP数据集中，CNN-5在真实数据和对抗数据中性能表现最佳，而ResNet-50在真实数据和对抗数据中均表现最差。原因可能使ResNet的网络架构较为复杂，即相比其他两个CNN模型含有更多的卷积层，太多的卷积层可能会降低模型收敛速度。CNN-5在对抗训练中($\varepsilon = 0.04$)获得59.7%的准确率，相比在非对抗训练($\varepsilon = 0.0$)中，模型的性能得到了一定的提升。综上所述，本文提出的面向对抗攻击的鲁棒性语音情感识别方法能有效地提高模型的鲁棒性和识别准确率。

表3 CNN-5、ResNet-50和VGG-16模型在IEMOCAP数据集的性能比较

Tab.3 The performance comparison of CNN-5, ResNet-50 and VGG-16 models in the IEMOCAP dataset

训练方法	干扰常量 (ε)	无权重平均召回率(UAR)					
		CNN-5		ResNet-50		VGG-16	
		真实 测试	对抗 测试	真实 测试	对抗 测试	真实 测试	对抗 测试
Single	0.000	0.582	—	0.521	—	0.538	—
Vanilla	0.020	0.584	0.511	0.560	0.405	0.573	0.492
Vanilla	0.040	0.597	0.460	0.562	0.369	0.585	0.455
Vanilla	0.060	0.591	0.431	0.560	0.338	0.584	0.469
Vanilla	0.080	0.588	0.416	0.548	0.334	0.592	0.437
Vanilla	0.100	0.582	0.404	0.571	0.332	0.579	0.442

4 结论(Conclusion)

本文提出一种面向对抗攻击的鲁棒性语音情感识别方法实现过程：首先采用快速梯度符号法生成对抗数据，对模型进行攻击；然后将对抗数据和真实数据进行混合作为模型的训练数据，以训练出一个具有对抗攻击的网络模型，并在三个卷积网络模型中验证了训练方法的有效性。由于当前工作中提出的攻击方法较为单一，因此，在未来的工作中需增加更为多样化的对抗数据或攻击方式，以训练出更为健壮的语音情感识别模型。

参考文献(References)

- [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] REN Z, BAIRD A, HAN J, et al. Generating and protecting against adversarial attacks for deep speech-based emotion recognition models[C]// IEEE Signal Processing Society. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, New Jersey, United States: IEEE, 2020:7184-7188.
- [3] SCHULLER B, STEIDL S, BATLINER A, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism[C]// International Speech Communication Association. Proceedings of the INTERSPEECH 2013-14th Annual Conference of the International Speech Communication Association (ISCA). Amsterdam, Netherlands: ELSEVIER, 2013:148-152.
- [4] DEMIRCAN S, KAHRAMANLI H. Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech[J]. Neural Computing Applications, 2018, 29(8):59-66.
- [5] ZHAO X M, ZHANG S M. Spoken emotion recognition via locality-constrained kernel sparse representation[J]. Neural Computing Applications, 2015, 26(3):735-744.
- [6] ZHANG Z X, COUTINHO E, DENG J, et al. Cooperative learning and its application to emotion recognition from speech[J]. IEEE/ACM Transactions on Audio, Speech, Language Processing, 2014, 23(1):115-126.
- [7] KAYAOGU M, EROGLU ERDEM C. Affect recognition using key frame selection based on minimum sparse reconstruction[C]// Association for Computing Machinery. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. New York, United States: Association for Computing Machinery, 2015:519-524.
- [8] VALSTAR M, SCHULLER B, SMITH K, et al. Avec 2013: The continuous audio/visual emotion and depression recognition challenge[C]// Association for Computing Machinery. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. New York, United States: Association for Computing Machinery, 2013:3-10.
- [9] 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.
- [10] ALEX S B, MARY L, BABU B P. Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features[J]. Circuits, Systems, Signal Processing, 2020, 39:5681-5709.
- [11] DANESHFAR F, KABUDIAN S J, NEEKABADI A. Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier[J]. Applied Acoustics, 2020, 166:107360.
- [12] MIKOLOV T, KARAFIĆ T M, BURGET L, et al. Recurrent neural network based language model[C]// International Speech Communication Association. Eleventh Annual Conference of the International Speech Communication Association. Amsterdam, Netherlands: ELSEVIER, 2010:1045-1048.
- [13] HUANG Z W, DONG M, MAO Q R, et al. Speech emotion recognition using CNN[C]// Association for Computing Machinery. Proceedings of the 22nd ACM International Conference on Multimedia. New York, United States: Association for Computing Machinery, 2014:801-804.
- [14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.
- [15] ZHAO J F, MAO X, CHEN L J. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control, 2019, 47:312-323.
- [16] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. IEEE Transactions on Affective Computing, 2008, 42(4):335-359.

作者简介:

陈 港(1998-), 男, 硕士生.研究领域: 情感计算, 模式识别.
 陈 杰(1999-), 男, 本科生.研究领域: 情感计算.
 张石清(1980-), 男, 博士, 教授.研究领域: 情感计算, 模式识别.
 赵小明(1964-), 男, 硕士, 教授.研究领域: 情感计算, 模式识别.