残差网络在婴幼儿哭声识别中的应用

谢 湘* 张立强 王 晶
(北京理工大学信息与电子学院 北京 100081)

摘 要：该文使用语谱图结合残差网络的深度学习模型进行婴幼儿哭声的识别，使用婴幼儿哭声与非哭声样本比例均衡的语料库，经过五折交叉验证，与支持向量机(SVM)，卷积神经网络(CNN)，基于Gammatone滤波器的听觉谱残差网络(GT-Resnet)3种模型相比，基于语谱图的残差网络取得了最优结果，F1-score达到0.9965，满足实时性要求，证明了语谱图在婴幼儿哭声识别任务中能直观地反映声学特征，基于语谱图的残差网络是解决婴幼儿哭声识别任务的优秀方法。

关键词：婴儿哭声识别；深度学习；残差网络；语谱图
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2019)01-0233-07
DOI: 10.11999/JEIT180276

Application of Residual Network to Infant Crying Recognition

XIE Xiang ZHANG Liqiang WANG Jing
(*School of Information and Electronics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: The deep learning model based on the residual network and the spectrogram is used to recognize infant crying. The corpus has balanced proportion of infant crying and non-crying samples. Finally, through the 5-fold cross validation, compared with three models of Support Vector Machine (SVM), Convolutional Neural Network (CNN) and the cochleagram residual network based on Gammatone filters (GT-Resnet), the spectrogram based residual network gets the best F1-score of 0.9965 and satisfies requirements of real time. It is proved that the spectrogram can react acoustics features intuitively and comprehensively in the recognition of infant crying. The residual network based on spectrogram is a good solution to infant crying recognition problem.

Key words: Infant crying recognition; Deep learning; Residual network; Spectrogram

1 引言

哭声是婴幼儿表达自己的主要方式，婴幼儿哭声的自动检测在婴儿看护，智能家居等领域具有重要作用，是智能家居中对婴幼儿看护的重要一环，能有效减少看护家长的负担。近年来机器学习以及深度学习极大地推动了人工智能的发展，在语音识别，图像识别等领域取得了巨大成就，婴幼儿哭声检测作为语音识别领域中的应用，过去的研究做了很多关于特征和模型选取以及婴幼儿哭声发声机理的工作[1–13]。

文献[1,2]研究了婴儿哭声的种类及其基音频率、共振峰、发音连续性等声学特征和发音机理，为特征选取提供了启发。Abdulaziz等人[3]采用了共轭梯度算法和前馈神经网络对梅尔频率倒谱系数(MFCC)和线性预测倒谱系数(LPCC)特征的识别进行了比较，发现在婴儿哭声检测方面MFCC的效果优于LPCC。Cohen等人[4]提取了基音频率、MFCC和短时能量，运用K-NN分类算法进行分类，在有噪声的环境下测试也取得了较好的效果。Lavner等人[5]，Torres等人[6]和Chang等人[7]均采用CNN模型对语谱图(spectrogram)进行分类，与传统的逻辑回归和SVM相比取得了较好的效果。文献[8—10]使用基于Gammatone听觉滤波器的变换域特征和听觉谱(cochleagram)，在语音识别和语音分离等任务中取得了优于基于短时傅里叶变换时频分解方法的结果。

之前的研究中，传统的机器学习方法如SVM普遍依赖特征的选取，识别结果很大程度取决于特征选取的好坏，所选特征难以全面反映婴幼儿哭声的特点，且对于不同的应用场景需要精心设计特征

收稿日期：2018-03-23；改回日期：2018-09-04；网络出版：2018-09-11
*通信作者： 谢湘 xiexiang@bit.edu.cn
基金项目：国家自然科学基金(61473041, 11590772, 61571044)
Foundation Items: The National Natural Science Foundation of China (61473041, 11590772, 61571044)

才能取得较好的效果。而卷积神经网络(CNN)模型虽然直接从较为直观的语谱图中学习特征，但由于层数加深使得网络训练困难，浅层的CNN模型用在婴幼儿哭声检测上的结果仍有较大提升空间。近年来对于婴幼儿哭声识别的研究较少，且深度学习中涌现出以残差网络为代表的适合深层网络训练的算法，本文将深度学习算法残差网络应用于婴幼儿哭声识别，并将语谱图与基于Gammatone听觉滤波器的听觉谱作对比，比较SVM, CNN和残差网络3种模型在婴幼儿哭声识别问题上的性能。

本文结构安排如下，第1节介绍研究现状和本文研究内容，第2节分析婴幼儿哭声和婴幼儿哭声识别任务的建模方法，第3节为SVM, CNN和残差网络实验及分析，第4节为结论及未来工作计划。

2 婴幼儿哭声分析与建模

2.1 婴幼儿哭声分析

图1为婴幼儿哭声，成人说话声和铃声语谱图对比，根据文献[1,2]对婴儿哭声机理的研究，与成人话音相比，婴幼儿哭声在基音频率、发音周期性以及频谱包络等声学特征上具有突出的特点，婴幼儿的哭声的基音频率在200～500 Hz之间，且普遍大于400 Hz，成人话音的基音频率在60～450 Hz之间，普遍在350 Hz以下，表现在语谱图上则是“横纹”的纵坐标较高，且共振峰之间，也就是“横纹”之间间隔较大。在发声周期性上，婴幼儿在啼哭的时候会将存于胸腔中的气呼出，随后进行吸气来准备下一声啼哭，这一过程被称为呼气单元，在

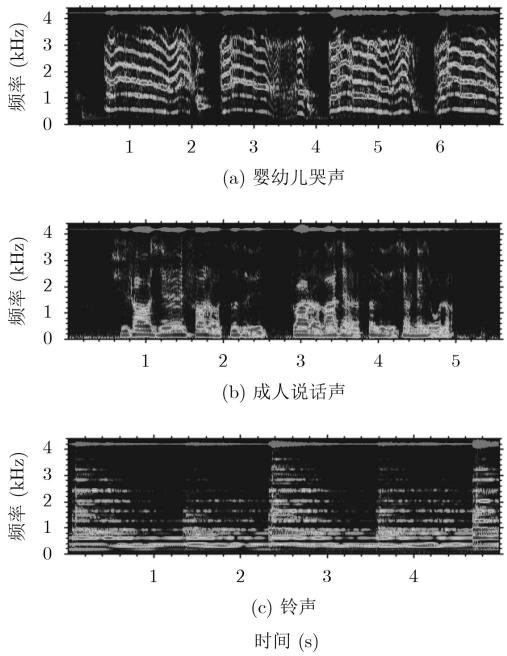

图 1 婴幼儿哭声，成人说话声和铃声语谱图对比

语谱图中表现为在连续的啼哭声中，不同呼吸单元的频域包络相似度很高，呈现周期性，而成人话音一般不会重复相同的内容。呼吸单元由吸气音和呼气音两段短时语音组成，吸气音持续时间较短，约为0.3 s，呼气音持续时间较长，约为1.7 s。基频与共振峰分别代表了声门脉冲激励信号的频率和声道的谐振频率，在呼气音的持续时间不断发生波动，表现在语谱图上即为频域包络的起伏，该特征也是与各种频域包络平稳的环境噪声区分的重要特征之一。

2.2 常用声学特征

对于传统机器学习方法，梅尔频率倒谱系数(MFCC)、基音频率、短时能量等是常用的声学特征，MFCC是基于人耳的听觉特性的一组参数，在语音信号处理的各个领域有广泛应用，基音频率和短时能量是能突出婴幼儿哭声特点的声学特征。

短时傅里叶变换与Gammatone听觉滤波模型[10]是两种常用的时频分解技术，与上述声学特征相比，时频分解后得到更为接近原始语料的声学特征，能较为全面地反映原始语料的特点，尤其是基于短时傅里叶变换的语谱图在深度学习模型中得到了广泛使用。语谱图是基于短时傅里叶变换的时频分解方法，能够较为全面反映语音信号时频特性，由2.1节可见共振峰、基频、能量等特征均可从语谱图中体现，有效区分婴幼儿哭声和成人话音以及环境噪声。

Gammatone听觉滤波器组的冲击响应为

$$g(t)=\begin{cases} t^{l-1}\exp(-2\pi bt)\cos(2\pi ft), & t\geq 0 \\ 0, & \text{其它} \end{cases} \tag{1}$$

其中，$l=4$为滤波器阶数，b约为等效矩形带宽(Equivalent Rectangle Bandwidth, ERB)，f为滤波器中心频率，沿对数频率轴等间隔分布在[80 Hz, 5 kHz]，对于4阶的Gammatone滤波器，带宽计算公式为

$$\left.\begin{aligned} &\mathrm{ERB}(f)=24.7(0.0043f+1.0) \\ &b=1.093\mathrm{ERB}(f) \end{aligned}\right\} \tag{2}$$

对每个滤波通道的频率响应进行分帧加窗，求出每个时频单元的能量即得到听觉谱，可见随着中心频率增加，滤波器的带宽加宽，基于Gammatone滤波器组的听觉谱符合人耳听觉特点，维度一般低于语谱图，能提高模型的训练效率。

2.3 婴幼儿哭声识别建模方法

支持向量机(SVM)是传统的机器学习算法，通过寻找特征空间中的划分超平面来进行分类，引入软间隔后可调整惩罚系数来降低过拟合的程度，引入核函数后可将非线性可分的特征映射到高维空

间，从而解决非线性可分问题。SVM依赖特征向量及核函数的选取，需要针对具体的分类任务设计特征，是常用的婴幼儿哭声检测模型。

卷积神经网络[14](CNN)是深度学习的重要模型之一，最初用在图像处理中，同时具有卷积结构和全连接结构，具有权值共享和稀疏连接等特点，增加网络深度能使网络提取到更加“高级”的特征，理论上网络层数加深分类效果会好，但是由于梯度消失等问题使得深层网络的训练变得困难，使得简单增加层数反而会使网络出现退化。

深度残差网络(deep Residual Network，简称Resnet)是2015年由He等人[15]提出，在卷积神经网络的基础上改进了卷积层的结构，是用来解决当网络层数加深时网络的学习能力反而退化的问题。假设存在一个恒等映射$H(\boldsymbol{x}, \boldsymbol{W})$，使得在浅层网络上叠加恒等映射可以使网络保持原状，而一般的多层非线性神经网络较难学到恒等映射$H(\boldsymbol{x}, \boldsymbol{W})$，引入残差模块后，残差网络可以轻易实现恒等映射，保证了网络层数加深不会导致学习能力退化。

图2为典型的残差模块[15]，设残差表示为

$$F(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^2 \cdot g(\boldsymbol{W}^1 \boldsymbol{x} + \boldsymbol{b}^1) + \boldsymbol{b}^2 \tag{3}$$

其中，$\boldsymbol{W}^1$和$\boldsymbol{W}^2$分别表示第1层和第2层卷积核的权重矩阵，$\boldsymbol{b}^1$和$\boldsymbol{b}^2$分别表示第1层和第2层卷积核的偏置矩阵，激活函数$g(\cdot)$一般为Relu函数，则待求的恒等映射表示为

$$H(\boldsymbol{x}, \boldsymbol{W}) = g(F(\boldsymbol{x}, \boldsymbol{W}) + \boldsymbol{x}) \tag{4}$$

其中，$F(\boldsymbol{x}, \boldsymbol{W})$表示对恒等映射的扰动值，当激活函数为Relu函数时，取$F(\boldsymbol{x}, \boldsymbol{W})$为0即实现恒等映射，当输入$\boldsymbol{x}$与$F(\boldsymbol{x}, \boldsymbol{W})$维度不同时，常用方法是对输入$\boldsymbol{x}$进行补零。相比一般的多层非线性神经网络，学习一个扰动值远比学习一个完整映射简单，这就保证了引入残差模块后网络层数加深不会引起网络退化。在网络不退化的基础上，深层的残差网络更易学到合适的残差映射$F(\boldsymbol{x}, \boldsymbol{W})$来提取更抽象的特征，从而提升卷积神经网络的性能。

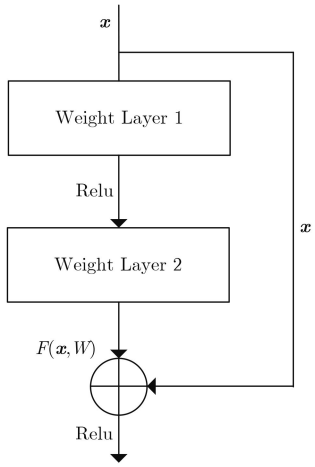

图 2 残差模块

残差网络在设计中常结合批标准化(batch normalization)，Adam梯度下降算法，Dropout，正则化和学习率衰减等技巧来加速网络的训练，防止过拟合。残差模块是残差网络与卷积神经网络的最关键区别，因此残差网络继承了卷积神经网络权值共享，稀疏连接和对空间结构敏感的特性，还能使得卷积神经网络的层数大幅加深。

将残差网络应用于婴幼儿哭声检测任务中时，首先要将1维音频信号进行时频分解为2维信号，2.1节分析语谱图能直观表现出婴幼儿哭声的基频、共振峰、周期性等声学特征，基于Gammatone滤波器的听觉谱更适合人耳听觉特性，在低频部分分辨率较强，因此将原始语料转化为语谱图或听觉谱作为输入信号是将残差网络应用于婴幼儿哭声识别任务的切入点。

3　实验与分析

3.1　语料库

本文所用的语料库来自Github上分享的“Donate a cry”[16]项目，均为用户手机录制上传的婴幼儿哭声语料，经过筛选得到450条清晰的、每条时长约为5 s的婴幼儿哭声语料。在哭声语料之外还选择了与婴儿真实环境相关联的声音作为非婴儿哭声的数据集，包括铃声(钟声)、鸟叫声、车笛声、猫叫、狗叫、玻璃碎裂声、警报声、鼾声和雷声各40条，每条时长约为3 s，还有100条普通话朗读语音，每条时长约为5 s。450条哭声及460条非哭声语料共同构成了本文实验的语料库，均为wav格式，采样率一致下采样为8 kHz。

本文实验环境为Intel Core i7-7700@4.20 GHz型号CPU, 16 G内存，双GTX 1070显卡。实验中所用的3种模型均使用五折交叉验证作为评估方法，使用分层采样将数据集划分为5个子集，每个子集保证哭声语料和非哭声语料的比例均衡。如2.1节分析，2 s的长度包含了婴幼儿哭声的完整周期，足以体现其特点，为了增加训练数据集的样本，本文将五折分层采样后的数据集以2 s的长度进行切分，并仿照语音分帧中的“帧移”建立切分窗口，每次切分向后移动1 s，若切分窗口移动到语音末端不足2 s，则从该段语音的末端向前移动2 s再切分。以切分完成后的5个子集进行交叉验证，其中训练集和测试集的平均规模如表1所示。

3.2　评价指标

查准率(precision)与召回率(recall)是分类性能

表 1 五折交叉验证数据集平均规模(条)

	婴幼儿哭声	非哭声	总计
训练集规模	1243	1148	2391
测试集规模	310	286	596

度量的常用指标，本文实验采用两者的调和平均值$F1$-score[17]为分类性能度量标准，$F1$-score定义为

$$F1=\frac{2PR}{P+R} \tag{5}$$

其中，P为查准率，R为召回率，$F1$-score兼顾了两指标，相比常用的准确率(accuracy)指标，能反映更多的分类性能信息。

本文实验以平均测试时间作为考察算法和模型在应用中实时性的重要指标，平均测试时间由特征提取时间和模型预测时间两部分组成。特征提取时间为从每条原始语音转化到特征向量、语谱图和听觉谱等模型输入所需时间，模型预测时间为保存好的模型对每个输入预测所需时间。

模型大小是评价模型复杂度的重要指标之一，本文实验将训练完成的模型保存到本地，SVM实验使用Libsvm工具保存，CNN及残差网络实验使用Tensorflow工具保存。

3.3 基线系统设置

SVM和CNN作为婴幼儿哭声识别的常用算法，具有各自的优势，本文实验将SVM和CNN均作为基线系统，实验语料库如3.1节所述。

3.3.1 SVM哭声识别模型

使用MATLAB中Libsvm[18]工具包实现SVM模型，首先进行预处理和特征提取[19]，预处理包括去静默段，加窗分帧和幅值归一化。每条语料提取了79维特征向量并进行归一化，SVM实验提取的特征向量是每段语料所有帧的统计平均特征，具体方法如表2。

在核函数选择方面，本文实验用了4种核函数，分别是线性核函数、高斯核函数、多项式核函数和sigmod核函数，实验结果及参数选择如表3，可见高斯核函数对于非线性可分问题具有良好的表现，$F1$-score达到0.9458，将其作为SVM基线模型。

3.3.2 CNN哭声识别模型

文献[5,7]中分别使用了两种不同的4层CNN网络结构，文献[5]以Mel尺度语谱图作为输入信号，文献[7]以227×227尺寸的语谱图作为输入信号，本文实验在两种文献所用结构的基础上进行了参数调优，将效果最好的CNN网络作为基线系统模型，结果如表4。

表 2 SVM实验特征提取

提取特征类型	统计处理方法	维数
MFCC及其1阶2阶差分	均值、方差	72
短时能量	均值、方差	2
基音频率	均值、方差、最大值、最小值、极差	5

表 3 SVM不同核函数性能比较

核函数类型	$F1$-score	参数
线性核函数	0.8717	c=0.68
多项式核函数	0.9316	c=0.30, g=0.35, r=–0.20, d=3.00
高斯核函数	**0.9458**	c=0.98, g=1.71
Sigmod核函数	0.8874	c=5.00, g=0.04, r=1.80

表 4 不同层数CNN性能对比

CNN模型	输入特征	$F1$-score
CNN-4-MEL	40×128Mel语谱图	0.9184
CNN-4-227	227×227语谱图	0.9233
CNN-4	128×128语谱图	0.9229
CNN-5-227	227×227语谱图	0.9482
CNN-5	128×128语谱图	**0.9489**
CNN-6	128×128语谱图	0.9365
CNN-7	128×128语谱图	0.9398

经过多组实验验证，227×227尺寸的语谱图与128×128尺寸语谱图得到的性能非常接近，考虑到计算复杂度，本文之后输入特征为语谱图的网络均采用128×128尺寸。由表4得，CNN-5取得了最优结果，$F1$-score平均为0.9489，结构与参数如图3所示，总计5层网络。其中第1层卷积层采用8×8大小的卷积核，步长为2，提取特征数为64，卷积层后使用批标准化和步长为2的2×2最大化池化层，其他卷积层与全连接层的参数如图3中所示。

CNN模型随着层数的加深会提取到更加抽象有效的特征，表4中5层CNN结果优于SVM，但第6层、第7层CNN反而出现了退化，实验表明直接增加CNN层数会出现训练困难，泛化能力退化的问题，本文提出使用残差网络应用于婴幼儿哭声识别中，以解决CNN随着网络层数加深出现的退化问题。

3.4 残差网络哭声识别模型

在CNN中引入残差模块，搭建残差网络加深网络层数，以提取更为抽象有效的特征，语料库如3.1节所述。本文实验将128×128尺寸的语谱图与基于Gammatone听觉滤波器的64×128尺寸的听觉谱用于不同结构的残差网络中作对比，经过参数调优，实验结果汇总于表5，3种模型与3种层数残差网络的测试集收敛过程分别绘制于图4和图5。

由3.5节分析，Resnet19在婴幼儿哭声识别应

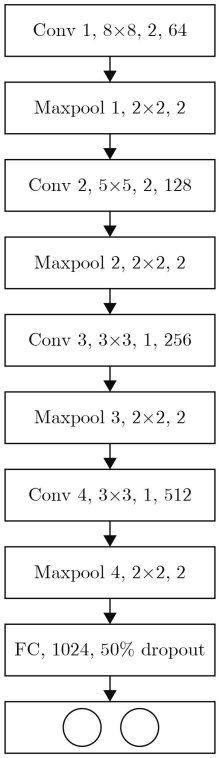

图 3 CNN-5模型结构

用中综合性能最优，平均F1-score达到0.9965，满足实时性的要求，其网络结构和参数如图6所示，虚线矩形框为一个ResBlock结构，随后重复3次，总计19层网络。每个ResBlock的features参数相同，4个不同的ResBlock分别设为64, 128, 256, 512，其它参数含义与图3相同，激活函数为Relu函数，每层卷积层后都有批标准化。学习率初始值为5×10^{-4}，每5个训练周期衰减5%，梯度寻优算法为Adam算法，批次大小设置为128，共训练100个周期，损失函数为全连接层和输出层的L2正则化项加交叉熵构成。

3.5 实验结果分析

结合表5及图4、图5，进行SVM, CNN，语谱图残差网络(Resnet)和听觉谱残差网络(GT-Resnet)4种婴幼儿哭声识别模型性能分析，各评价指标如3.2节所述。

(1)分类性能：以F1-score为标准，基于语谱图的Resnet模型普遍高于其他3种模型，F1-score达到0.9965，比基线系统SVM和CNN分别高出0.0507和0.0476，取得较大突破。GT-Resnet的F1-score达到0.9803，不如以语谱图作为输入的结果，说明基于Gammatone滤波器的听觉谱为突出人耳听觉特征而丢失的信息影响了残差网络对婴幼儿哭声特征的提取，但仍取得了较好的分类性能。因此基于残差网络的婴幼儿哭声识别模型学习和泛化能力强，能从原始语料时频分解后的语谱图中提取出有效特征，满足婴幼儿哭声检测任务的需求。

表 5　模型性能对比

模型	网络结构	输入特征	生成模型大小(MB)	平均测试时间(s)	F1-score
SVM	单层网络	统计特征	0.7	**0.0910+0.0001**	0.9458
CNN-5	4conv+1fc	语谱图	10	0.1251+0.0093	0.9489
Resnet15	3resblock+1fc	语谱图	48	0.1251+0.0281	0.9836
Resnet19	4resblock+1fc	语谱图	87	0.1251+0.0315	**0.9965**
Resnet27	6resblock+1fc	语谱图	171	0.1251+0.0355	0.9965
GT-Resnet15	3resblock+1fc	听觉谱	48	0.1933+0.0218	0.9803
GT-Resnet19	4resblock+1fc	听觉谱	87	0.1933+0.0237	0.9782
GT-Resnet27	6resblock+1fc	听觉谱	171	0.1933+0.0285	0.9719

注：平均测试时间=特征提取时间+模型预测时间

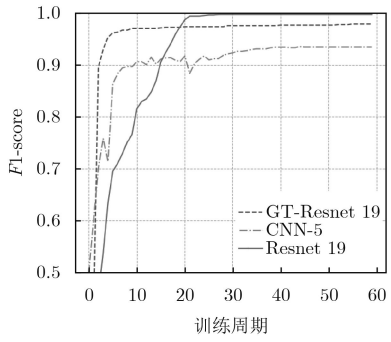

图 4 3种模型测试集F1-score对比

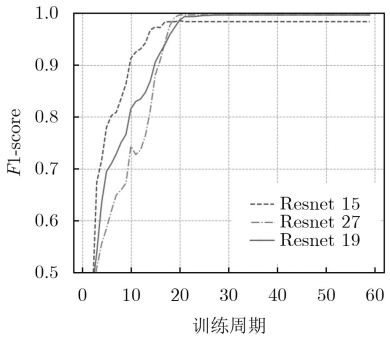

图 5 3种层数残差网络测试集F1-score对比

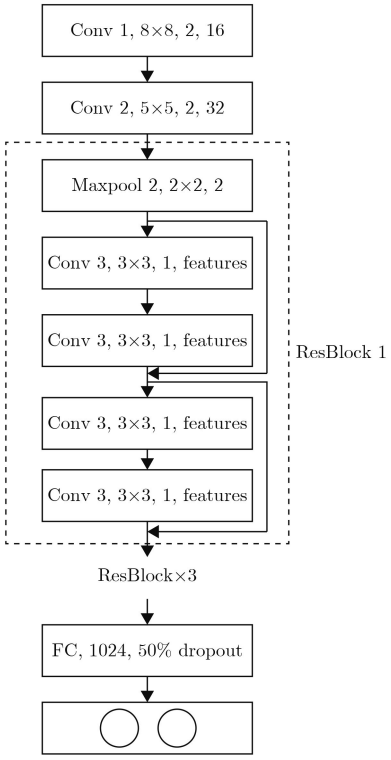

图 6 残差网络模型

(2)实时性：以每条语料的平均测试时间为标准，SVM模型速度最快，仅需要0.09 s，基于语谱图的Resnet小于0.16 s，相比长为2 s的输入语料，满足实际应用对实时性的要求。实验发现对原始语料进行特征提取等预处理的时间远比各种模型预测时间长，如Resnet19中约0.15 s的测试时间有80%用来特征提取，因此若对实时性进一步优化，需要着重提高特征提取等预处理工作的处理效率。语谱图的提取需要用到短时傅里叶变换，快速傅里叶变换(FFT)较为成熟的硬件实现方案有助于在实际应用中对基于语谱图的残差网络模型进一步提速。

(3)模型复杂度：SVM的模型大小仅为0.7 MB，模型复杂度最低，其实时性也最强。由图4所示，因为基于Gammatone滤波器的听觉谱比语谱图尺寸小，突出了基于人耳听觉的声学特征，使GT-Resnet网络训练效率高，收敛速度快，但不会降低模型复杂度，分类性能和特征提取时间均不如基于语谱图的Resnet。Resnet19的模型大小为87 MB，在实际应用中满足要求，但随着网络层数，尤其是深层的网络层数增加，模型大小有明显增大，因此对比性能和实时性，在Resnet19和Resnet27具有同等分类性能的情况下，模型复杂度更低的Resnet19更适合应用于婴幼儿哭声识别任务中。

4 结束语

本文提出了使用语谱图结合残差网络的方法进行婴幼儿哭声识别，在较小语料库中没有出现过拟合或欠拟合，$F1$-score达到0.9965，比基线系统SVM和CNN模型分别高出0.0507和0.0476，比基于Gammatone听觉滤波器的听觉谱残差网络高出0.0162，证明了语谱图在婴幼儿哭声检测问题上能直观全面地反应声学特征，结合残差网络强大的学习和泛化能力，是解决婴幼儿哭声识别任务的优秀方法。在3种模型中残差网络的复杂度较高，但本文实验说明调用已训练完成的模型所需时间远小于特征提取的时间，且约0.15 s的识别时间满足实时性的要求，在实际应用中，通过DSP等硬件实现还可进一步降低特征提取所需的时间。未来的工作需要收集更丰富，标注更详细的语料库，进行婴幼儿哭声病理性分析，辅助医生或看护人判断婴幼儿状态，这也是当前哭声识别相关工作的难点。

参 考 文 献

[1] 于洪志, 刘思思. 三个月婴儿啼哭声的声学分析[C]. 全国人机语音通讯学术会议, 西安, 2011: 1–4.
YU Hongzhi and LIU Sisi. Crying sound learning analysis of three months baby[C]. National Conference on Man-Machine Speech Communication, Xi'an, China, 2011: 1–4.

[2] 王之禹, 雷云珊. 婴儿啼哭声的声学特征[C]. 中国声学学会2006年全国声学学术会议, 厦门, 2006: 389–390.
WANG Zhiyu and LEI Yunshan. Acoustic characteristic of infant cries[C]. National Conference on Acoustics. Acoustical Society of China, Xiamen, China, 2006: 389–390.

[3] ABDULAZIZ Y and AHMAD S M S. Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients[C]. International Conference on Information Retrieval & Knowledge Management, Shah Alam, Malaysia, 2010: 260–263. doi: 10.1109/INFRKM.2010.5466907.

[4] COHEN R and LAVNER Y. Infant cry analysis and detection[C]. Electrical & Electronics Engineers in Israel, Eilat, Israel, 2012: 1–5.

[5] LAVNER Y, COHEN R, RUINSKIY D, *et al*. Baby cry detection in domestic environment using deep learning[C]. 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, Israel, 2016: 1–5. doi: 10.1109/EEEI.2012.6376996.

[6] TORRES R, BATTAGLINO D, and LEPAULOUX L. Baby cry sound detection: A comparison of hand crafted features and deep learning approach[C]. International Conference on Engineering Applications of Neural Networks. Springer, Cham, 2017: 168–179. doi: 10.1007/978-3-319-65172-9_15.

[7] CHANG Chuanyu and LI Jiajing. Application of deep learning for recognizing infant cries[C]. IEEE International

Conference on Consumer Electronics, Nantou, China, 2016: 1–2. doi: 10.1109/ICCE-TW.2016.7520947.

[8] SHARAN R V and MOIR T J. Cochleagram image feature for improved robustness in sound recognition[C]. IEEE International Conference on Digital Signal Processing, Singapore, 2015: 441–444. doi: 10.1109/ICDSP.2015.7251910.

[9] PATTERSON R D, NIMMO-SMITH I, HOLDSWORTH J, *et al*. An efficient auditory filterbank based on the gammatone function[C]. Proceedings of the 1987 Speech-Group Meeting of the Institute of Acoustics on Auditory Modelling, RSRE, Malvern, 1987: 2–18.

[10] 刘文举, 聂帅, 梁山, 等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报, 2016, 42(6): 819–833. doi: 10.16383/j.aas.2016.c150734.
LIU Wenju, NIE Shuai, LIANG Shan, *et al*. Deep learning based speech separation technology and its developments[J]. *Acta Automatica Sinica*, 2016, 42(6): 819–833. doi: 10.16383/j.aas.2016.c150734.

[11] MITTAL V K. Discriminating features of infant cry acoustic signal for automated detection of cause of crying[C]. International Symposium on Chinese Spoken Language Processing, Tianjin, China, 2017: 1–5. doi: 10.1109/ISCSLP.2016.7918391.

[12] RPSITA Y D and JUNAEDI H. Infant's cry sound classification using Mel-Frequency Cepstrum Coefficients feature extraction and Backpropagation Neural Network[C]. International Conference on Science and Technology-Computer, Yogyakarta, Indonesia, 2017: 160–166. doi: 10.1109/ICSTC.2016.7877367.

[13] 雷云珊. 婴儿啼哭声分析与模式分类[D]. [硕士论文], 山东科技大学, 2006.
LEI Yunshan. Analysis and pattern classification of infants' cry[D]. [Master dissertation], Shandong University of Science and Technology, 2006.

[14] KRIZHEVAKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems, Nevada, USA, 2012: 1097–1105.

[15] HE Kaiming, ZHANG Xianyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Computer Vision and Pattern Recognition, Nevada, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[16] GVERES. donateacry-corpus[OL]. https://github.com/gveres/donateacry-corpus, 2017.3.

[17] 彭天强, 栗芳. 基于深度卷积神经网络和二进制哈希学习的图像检索方法[J]. 电子与信息学报, 2016, 38(8): 2068–2075. doi: 10.11999/JEIT151346.
PENG Tianqiang and LI Fang. Image retrieval based on deep convolutional neural networks and binary hashing learning[J]. *Journal of Electronics & Information Technology*, 2016, 38(8): 2068–2075. doi: 10.11999/JEIT151346.

[18] CHANG Chihchung and LIN Chihjen. LIBSVM: A library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 1–27. doi: 10.1145/1961189.1961199.

[19] 徐利强, 谢湘, 黄石磊, 等. 连续语音中的笑声检测研究与实现[C]. 全国声学学术会议, 武汉, 2016: 581–584.
XU Liqiang, XIE Xiang, HUANG Shilei, *et al*. Research and implementation of laughter detection in continuous speech[C]. National Conference on Acoustics. Acoustical Society of China, Wuhan, China, 2016: 581–584.

谢　湘：男，1976年生，副教授，研究方向为语音识别.
张立强：男，1995年生，硕士生，研究方向为语音人格感知.
王　晶：女，1980年生，副教授，研究方向为音频信号处理.