第 42 卷 第 4 期
2023 年 7 月

应用声学
Journal of Applied Acoustics

Vol. 42, No. 4
July, 2023

◇ 研究报告 ◇

基于Mel频谱值和深度学习网络的鸟声识别算法*

李大鹏　周晓彦†　王基豪　王丽丽　叶　如

(南京信息工程大学电子与信息工程学院　南京　210044)

摘要： 为了增强网络对鸟鸣声信号的特征学习能力并提高识别精度，提出一种基于深度残差收缩网络和扩张卷积的鸟声识别方法。首先，提取鸟鸣声信号的对数Mel特征及其一阶和二阶差分系数组成log-Mel特征集，作为网络模型的输入；其次，通过深度残差收缩网络自动学习噪声阈值，减少噪声干扰；然后，引入扩张卷积增大卷积核感受野并利用注意力机制使网络聚焦于关键帧特征；最后，通过双向长短时记忆网络从学到的局部特征中学习长期依赖关系。以北京百鸟数据库中的19种中国常见鸟类作为实验对象，识别正确率可以达到96.58%，并对比模型在不同信噪比数据下的识别结果，结果表明该模型在噪声环境下的识别效果优于现有模型。

关键词： 鸟声识别；log-Mel特征；深度残差收缩网；扩张卷积神经；注意力机制

中图法分类号： TN912.34　　**文献标识码：** A　　**文章编号：** 1000-310X(2023)04-0825-08

DOI: 10.11684/j.issn.1000-310X.2023.04.018

Bird voice recognition algorithm based on Mel spectrum value and deep learning network

LI Dapeng　ZHOU Xiaoyan　WANG Jihao　WANG Lili　YE Ru

(*College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing* 210044, *China*)

Abstract: In order to enhance the feature learning ability of the network to the birdsong signal and improve the recognition accuracy, a birdsong recognition method based on depth residual shrinkage network (DRSN) and expanded convolution is proposed. Firstly, the logarithmic Mel feature and its first-order and second-order difference coefficients of birdsong signal are extracted to form a log-Mel feature set as the input of the network model; Secondly, the noise threshold is automatically learned through the DRSN to reduce the noise interference; Then, the expanded convolution is introduced to increase the receptive field of convolution kernel, and the attention mechanism is used to make the network pay more attention to the characteristics of key frames; Finally, the long-term dependence is learned from the learned local features through the two-way long-term and short-term memory network. Taking 19 kinds of common Chinese birds in birdsdata as the experimental object, the recognition accuracy can reach 96.58%. Compared with the recognition results of the model under different signal-to-noise ratio data, the results show that the recognition effect of the model in noise environment is better than that of the existing model.

Keywords: Bird song recognition; log-Mel feature; Depth residual shrinkage network; Dilated convolutions; Attention mechanism

2022-04-25 收稿; 2022-08-02 定稿

*国家自然科学基金项目 (62076064)

作者简介: 李大鹏 (1997–), 男, 江苏徐州人, 硕士研究生, 研究方向: 模式识别, 声音信号处理。

†通信作者 E-mail: xiaoyan_zhou@nuist.edu.cn

0 引言

鸟类是生态系统的重要组成部分。对鸟类活动及其分布的监测，为了解一个地区的生物多样性变化和气候变化提供了重要的依据[1−2]。鸟鸣声是区分鸟类的重要特征。鸟鸣声识别也是目前鸟类物种识别普遍采用的方式之一。通过鸟声识别实现鸟类监测具有高效、稳定、范围广的优点，具有巨大的应用价值。

鸟鸣声识别的关键在于减少自然环境下噪声的影响，提取合适的鸣声特征，匹配分类器进行识别。目前，鸟声识别的分类方法大致有3种：(1) 基于模板匹配的分类方法。最常见的是动态时间规整(Dynamic time warping, DTW)算法。例如，徐淑正等[3]使用基于音长、Mel频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)、线性预测系数(Linear prediction coefficient, LPCC)和时频域纹理特征的DTW算法并结合多种分类器进行鸟声识别。此类算法时间复杂度较高，容易受到噪声干扰。(2) 基于传统机器学习的分类方法。此类方法多采用手工提取特征，利用支持向量机(Support vector machine, SVM)[4]、随机森林(Random forest, RF)[3]等分类器进行识别。例如，张赛花[4]提取了一种Mel子带参数化特征，使用SVM对野外11种鸟鸣声进行分类识别，结果表明该方法对11类鸟声查全率、查准率和F1-score均高于89%。目前该类算法正确率的提高多依赖于对特征的优化与选择，其主要适应于小样本数据集，在样本充足的情况下识别效果低于深度学习的方法。(3) 基于深度学习的方法。深度学习网络具有很好的自动学习特征的能力，近年来在鸟类物种识别中得到了广泛的应用并取得了良好的效果。例如，Cakir等[5]提出了基于卷积递归神经网络(Convolutional recurrent neural networks, CRNN)的方法实现鸣声的高维特征及短时帧间的相关性特征提取，对Freesound数据中的鸟鸣声进行分类实验，正确率达到88.5%。冯郁茜[6]提出了基于双模态特征融合的鸟类物种分类算法，融合卷积网络提取的语图特征和长短时记忆结构提取的鸣声时序序列特征，自适应完成鸟鸣声的物种识别。Naranchimeg等[7]利用卷积神经网络(Convolutional neural networks,CNN)提取语图特征并且提出跨模态结合特征，提高了分类识别的性能。谢将剑等[8]采用3种不同语谱图作为输入特征并进行特征融合，利用VGG16网络进行鸟类物种识别，实验表明特征融合模型具有更好的识别效果。Puget[9]将通过短时傅里叶变换(Short time Fourier transform, STFT)生成的STFT语谱图经过网格化处理后作为Transformer神经网络的输入，并通过Xeno-Canto鸟声数据库中397类鸟声识别，测试后准确率达到77.55%。邱志斌等[10]将Mel语谱图输入自搭建的24层CNN 模型中，并通过反复执行卷积、池化操作及微调内部参数，在40类鸟类鸣声中识别准确率能达到96.1%。Liu等[11]提出了一种将双向长短期记忆网络(Bidirectional long-short term memory, BiLSTM)和DenseNet卷积神经网络级联组合的鸟声分类模型，将Mel语谱图作为输入，在北京百鸟数据库中20种鸟类声频中平均准确率能达到92.2%。上述文献[5–11]基于深度学习的方法主要以语谱图作为模型的输入，通过CNN、RNN等网络进一步提取高等级特征进行分类识别，取得了良好的识别效果。但上述文章均未考虑噪声对于网络性能的影响。鸟鸣信号在自然环境中获取，往往包含大量噪声，为了增强对含噪鸟鸣声特征的学习能力，本文受深度残差收缩网络(Deep residual shrinkage networks, DRSN)[12]、卷积块注意力模块(Convolutional block attention module, CBAM)[13]、通道注意力(Efficient channel attention, ECANet)[14]的启发，结合扩张卷积[15]和残差思想[16]，设计了基于DRSN和扩张卷积的鸟声识别网络，以提高模型在自然场景下鸟声识别的分类精度。本文的主要工作如下：

(1) 提取鸟鸣声信号的对数Mel特征及其一阶和二阶差分系数组成log-Mel特征向量作为网络模型的输入。

(2) 设计了更加高效的深度残差收缩模块。结合ECANet网络的思想对DRSN进行改进，即通过一维卷积替代DRSN模型注意力机制中的两层全链接，降低参数量的同时增强对含噪鸟鸣声的特征学习能力。

(3) 基于扩张卷积、残差连接和结合空间注意力机制构建局部特征提取模块，将提取到的局部特征输入BiLSTM，考虑时间依赖性关系进一步提取全局特征。

(4) 在北京百鸟数据birdsdata鸟声库上进行

实验，分析本文不同网络的作用并与其他基于深度学习网络的鸟声识别算法进行对比，最后研究本文方法在不同信噪比数据下的识别效果。

1　基于 Mel 频谱值和深度学习网络的鸟声识别算法

本文所提出的鸟声识别算法总体框架如图 1 所示。首先，对于输入的鸟鸣声信号进行预加重、分帧、加窗，通过 STFT 和 Mel 滤波操作得到 MFCC 并计算得到其一阶差分、二阶差分系数组成 3 维 log-Mel 特征向量；其次，将特征向量输入一个卷积单元进行特征提取，通过池化层缩小特征图大小，并输入深度残差收缩模块减弱噪声干扰；然后，通过残差连接和 3 个扩张卷积单元结合空间注意力机制 (Spatial attention module, SAM) 组成扩张卷积注意力模块 (DilatedSAM) 进一步提取高等级空间局部特征；最后，输入 BiLSTM 层来捕获时间序列特征，再经过全连接、softmax 层实现鸟鸣声的分类识别。

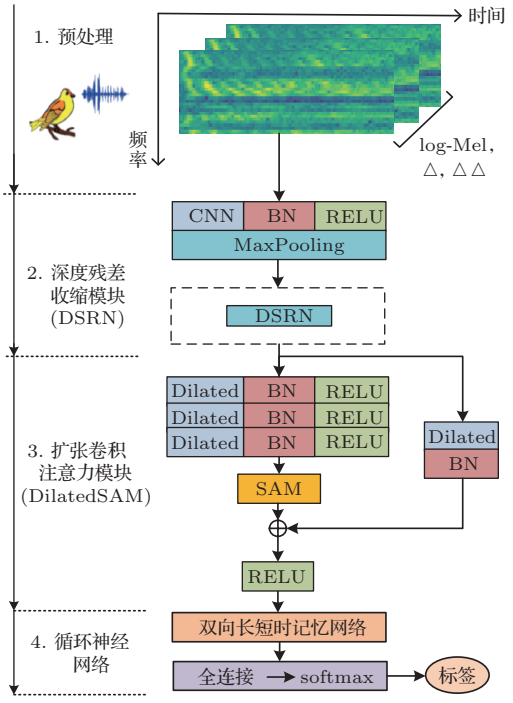


图 1　鸟声识别网络总体结构

Fig. 1　General structure of the bird sound recognition network

1.1　对数 Mel 特征 (log-Mel)

静态特征仅描述了帧级声频的能谱包络，而声频具有一定的动态信息。在语声情感识别领域的相关文献 [17–18] 将静态特征和动态信息相结合取得了较好的识别效果，因此本文提取鸟鸣声信号的 log-Mel 特征并计算其一阶差分和二阶差分系数，将静态和动态信息相结合组成 3 维 log-Mel 特征向量。处理过程如图 2 所示。

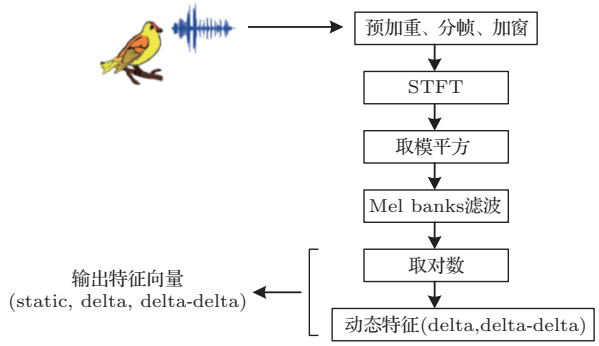


图 2　log-Mel 特征提取过程

Fig. 2　log-Mel feature extraction process

(1) 将鸟鸣声通过高通滤波器进行预加重处理，高通滤波器表示为

$$H(z) = 1 - \mu z^{-1}, \tag{1}$$

其中，μ 的取值范围为 0.9~1，本文取 0.94；

(2) 对预加重后的鸟鸣声信号进行分帧、汉明窗加窗，其中帧长为 25 ms、帧移为 10 ms；

(3) 对每一帧进行离散傅里叶变换 (Discrete Fourier transform, DFT) 后得到各帧的频谱，并对频谱取模平方得到对应的功率谱，将时域信号转换为频域上的能量分布；

(4) 将功率谱输入到 Mel 滤波器组中得到能量值，对于第 i 个滤波器 ($0 < i \leqslant 40$)，能量为 p_i，对 p_i 进行对数变换后计算出倒谱 Mel 频率 $y_i = \log(p_i)$；

(5) 为了更好地体现时域连续性，可在静态特征增加前后帧动态信息，可由 y_i 计算一阶差分 z_i^d 和二阶差分 z_i^{dd}：

$$z_i^d = \frac{\sum_{n=1}^{N} n(y_{i+n} - y_{i-n})}{2\sum_{n=1}^{N} n^2}, \tag{2}$$

$$z_i^{dd} = \frac{\sum_{n=1}^{N} n(z_{i+n}^d - z_{i-n}^d)}{2\sum_{n=1}^{N} n^2}, \tag{3}$$

其中，$N=2$，计算得到信号的动态信息 z_i^d、z_i^{dd}，与静态特征 y_i 组成 3 维 log-Mel 特征向量 $\boldsymbol{X}\in\boldsymbol{R}^{t\times f\times k}$，其中，$t$ 表示时间帧个数，f 表示 Mel 滤波器的个数，k 表示特征的通道数，这里 $t=200$、$f=40$、$f=3$。

1.2 DRSN

在实际环境中采集到的鸟鸣声数据，往往存在大量的背景噪声，影响模型识别的准确率。为解决此问题，本文提出一种改进的 DRSN，从而减弱环境噪声对识别结果的影响。文献 [11] 为解决滚动轴承故障诊断中的高噪声问题，将信号去噪中经常使用的软阈值函数引入深度残差神经网络中，并利用通道注意力机制 [19] 自动确定噪声阈值，提出了一种能够自适应软阈值的 DRSN。本文为了进一步降低 DRSN 网络的参数量，利用一维卷积替代 DRSN 模型注意力机制中的两层全链接，其具体结构如图 3 所示。

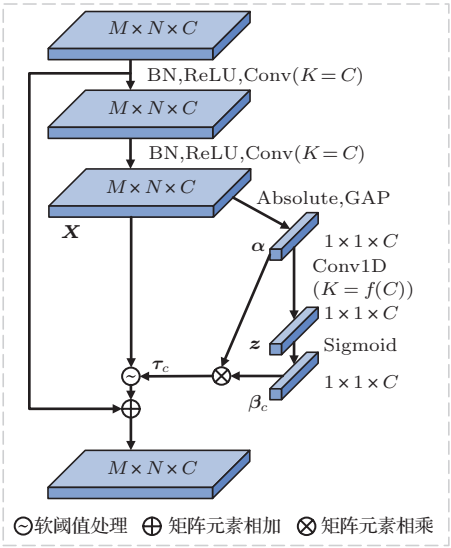


图 3 改进的深度残差收缩单元

Fig. 3 Improved depth residual shrinkage unit

对于输入的三维特征图 $\boldsymbol{X}(M\times N\times C)$ 首先通过取绝对值和全局平均池化操作将特征信息进行压缩得到维度为 $1\times1\times C$ 的向量 $\boldsymbol{\alpha}$，计算公式如下：

$$\boldsymbol{\alpha}=\left|\text{GolbalAverage}(\boldsymbol{X}_{M,N,C})\right|. \tag{4}$$

其次通过一维卷积得到每个通道的注意力参数，同时在两层全连接网络后应用 sigmoid 函数，使注意力参数缩放到 (0,1)，其计算公式如下：

$$\boldsymbol{\beta}_C=\frac{1}{1+\mathrm{e}^{-\boldsymbol{z}}}, \tag{5}$$

其中，$\boldsymbol{z}$ 为一维卷积的输出，$\boldsymbol{\beta}_C$ 为注意力参数。

最后注意力参数 $\boldsymbol{\beta}_C$ 乘以向量 $\boldsymbol{\alpha}$，得到最终阈值 $\boldsymbol{\tau}_c$，从而确保阈值为正同时不会太大。

综上所述，软阈值可以表示为

$$\boldsymbol{\tau}_C=\boldsymbol{\beta}_C\odot\text{Average}\left|\boldsymbol{X}_{M,N,C}\right|, \tag{6}$$

其中，$\boldsymbol{\tau}_C$ 为特征矩阵对应通道的阈值；M、N、C 分别为特征图 $\boldsymbol{X}$ 的宽度、高度和通道，$\odot$ 为矩阵的哈达玛积。

图 3 中 ⊖ 为软阈值操作，即将每个通道特征图参数在 $-\boldsymbol{\tau}_C\leqslant\boldsymbol{X}\leqslant\boldsymbol{\tau}_C$ 的特征设为 0，其他特征参数向 0 收缩。具体计算公式为

$$\boldsymbol{Y}=\begin{cases}\boldsymbol{X}-\boldsymbol{\tau}_C, & \boldsymbol{X}>\boldsymbol{\tau}_C,\\ 0, & -\boldsymbol{\tau}_C\leqslant\boldsymbol{X}\leqslant\boldsymbol{\tau}_C,\\ \boldsymbol{X}+\boldsymbol{\tau}_C, & \boldsymbol{X}<-\boldsymbol{\tau}_C,\end{cases} \tag{7}$$

其中，$\boldsymbol{X}$ 为输入特征参数，$\boldsymbol{Y}$ 为输出特征参数，$\boldsymbol{\tau}_C$ 为阈值。

在经典的信号去噪算法中，设置合适的阈值往往需要大量经验，残差收缩单元通过注意力机制实现了不同通道阈值的自动确定，避免了人工设置的麻烦。为了进一步减少确定阈值所需的计算量、降低模型复杂度，本文借鉴 ECANet 网络的方法，用一维卷积替代残差收缩单元中两层全连接网络，实现跨通道信息的交互，并通过选择一维卷积核大小确定局部跨通道交互的覆盖范围。

对于给定的通道维度 C，一维卷积核大小 K 计算公式如下：

$$K=f(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|. \tag{8}$$

对于参数 γ 和 b 采用 ECA-Net 网络中的设定，将 γ 和 b 分别设置为 2 和 1。

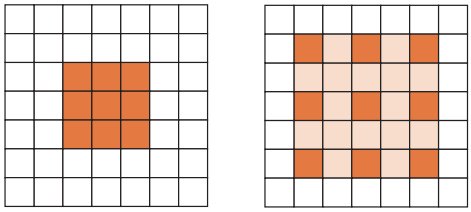
1.3 扩张卷积残差注意力结构

为了进一步有效提取鸟鸣声特征，减少池化带来的信息丢失，同时希望网络能够聚焦于关键帧信息，本文结合扩张卷积和 CBAM 网络中的空间注意力机制及残差的思想，提出了扩张卷积残差注意力结构。传统的 CNN 主要由卷积层和池化层组成，其中，卷积层用来提取局部特征；池化层用来对特征图进行下采样减小特征图尺寸，间接提高下层卷积感知的范围。然而池化层在减小特征图尺寸的过程中，可能会造成一些信息的丢失，对于此问题，在本

模块中采用扩张卷积来代替传统的CNN，在特征提取过程中不丢失信息和增加计算量的情况下获得更大的感受野。扩张卷积的结构比较简单，通过在标准卷积中增加空洞的方式，实现感受野的扩大。如图4所示，在标准卷积行列权值中插入 $r-1$ 个值为0的权值，γ 为扩张率，其感受野的计算公式如下：

$$l_j = l_{j-1} + \left((f_j - 1) * \prod_{i=1}^{j-1} s_i \right), \tag{9}$$

其中，j 表示卷积层序号，l_j 为第 j 个卷积层的感受野大小，f_j 表示该层卷积核尺寸，s_i 表示卷积步长大小。



(a) 二维标准卷积　　(b) 二维扩张卷积($r=2$)

图4　标准卷积与扩张卷积示意图

Fig. 4 Schematic diagram of standard convolution and dilation convolution

扩张卷积残差注意力网络主要的特征提取部分由扩张卷积层(DiltedCNN)、批量归一化层(Batch normalization, BN)和RELU层组成扩张卷积单元。由于扩张卷积层的存在，可以在不使用池化层的情况下获得更大的感受野，提取局部特征。BN层对特征进行归一化处理，提高结构的性能和稳定性。

1.4　BiLSTM

LSTM模型是一种改进的时间递归神经网络，解决了循环神经网络梯度爆炸和梯度消失的问题[20]。LSTM在时间序列信息处理中得到了广泛的应用，尤其在声频领域[5,21]。LSTM 可以选择性地学习长期信息序列信息，拥有3个"门"对信息进行控制，即输入门、输出门和遗忘门，遗忘门根据输入和前次输出来帮助模型遗忘一些无用的信息。

鸟鸣声信号是一种时序信号，具有动态特性，而LSTM内部的循环机制使其具有对时序序列的记忆能力，能综合考虑时序序列前后帧特征之间的联系。本文使用BiLSTM，结合前向信息和后向信息，其中，前向层捕获序列的历史信息；后向层捕获序列的未来信息。然后将前向层和后向层的隐藏状态连接起来，得到单个序列的隐藏状态，作为BiLSTM隐藏层的输出。

2　实验设置与分析

2.1　鸟声数据库

为了验证模型的有效性，本文选用的鸟类鸣声声频文件均来自Birdsdata手工标注自然声音标准大数据集[22]，该数据集由百鸟数据科技有限责任公司发布，其公开部分共收集了中国常见鸟种20种，该数据集共有进行过2 s标准化切割的44.1 kHz、wav声频文件14311个，各类鸟鸣声文件数量如表1所示。

表1　北京百鸟数据库

Table 1　Birdsdata

物种名称	数量	物种名称	数量
麻雀	1195	普通鸬鹚	852
林鹛	825	苍鹭	850
红脚鹛	790	红喉潜鸟	835
白腰草鹛	710	雉鸡	797
凤头麦鸡	814	西鹌鹑	738
黑翅长脚鹛	786	灰山鹑	29
骨顶鸡	460	绿翅鸭	602
西方秧鸡	680	绿头鸭	766
欧亚鸲	290	大天鹅	800
苍鹰	733	灰雁	759

由于数据库中灰山鹑数量过少，实验中删除该鸟类，采用19种鸟类，共计14282个声频文件。

2.2　实验设置

本文网络模型的搭建采用谷歌公司发布的基于TensorFlow 2.4.0的Keras2.4.3深度学习框架，硬件环境租用MistGPU平台的NVIDIA RTX 2080Ti显卡。模型训练的参数如表2所示。

表2　训练参数

Table 2　Train parameters

参数类型	值或方法
优化器	Adam
batch_size	32
最大训练次数	100
学习率	0.0001
损失函数	交叉熵函数

网络中所有卷积层卷积核个数K均设为128，padding设为same卷积模式，BiLSTM层的单元大小设置为128。

为了避免网络发生过拟合问题，文章采用了3种方法：(1) 每个卷积层后均添加BN层，提高网络的泛化能力。(2) 在全连接层之前采用dropout技巧，并设为0.5。(3) 对于每个卷积层采用L2正则化技巧，正则化参数设为0.0001。

为评估模型性能，本文将准确率(Accuracy)和F1-score作为自身模型和其他对比模型的评价指标。F1-score得分由精确率(Precision)和召回率(Recall)两项指标加权得到，具体计算公式如下：

查准率(精准率)：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{10}$$

查全率(召回率)：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

正确率(准确率)：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{12}$$

F1-score：

$$\text{F1} = \frac{2\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{13}$$

其中，TP为正确地预测为正例，TN为正确地预测为反例，FP为错误地预测为正例，FN为错误地预测为反例。

实验协议采用五折交叉验证的方式，即将数据集分成5份，轮流将其中4份作为训练数据，1份作为测试数据进行实验。

2.3 实验与分析

本文实验采用北京百鸟数据库，为验证本文算法的有效性，实验共分为3个部分。首先对比一维静态log-Mel特征和3维log-Mel特征在本文模型上的识别效果，同时对比近年来相关论文所提算法；其次在不同强度的高斯白噪声背景下进行实验，验证本文算法在噪声环境下的识别效果；最后对本文模型进行消融实验，分析各个模块对模型识别结果的影响。

2.3.1 消融实验

为了验证深度残差收缩模块、扩张卷积和空间注意力模块的有效性，进行了消融实验，输入特征均为三维log-Mel频谱值。实验中将普通卷积加BiLSTM模型(CNN+BiLSTM)作为基准模型，分别对比基于扩张卷积的残差块加BiLSTM模型(dilatedCNN+BiLSTM)、基于扩张卷积和空间注意力的残差块加BiLSTM模型(DilatedSAM+BiLSTM)和DRSN加于扩张卷积和空间注意力的残差块加BiLSTM模型(DSRN+DilatedSAM+BiLSTM)。

如表3所示，将基线模型(CNN+BiLSTM)中普通卷积换成扩张卷积并增加残差技巧，识别正确率提高0.63%，在此基础上增加空间注意力机制，识别精度有少幅提升；原始数据均在自然环境中采集，包含大量背景噪声，增加DRSN后，识别正确率提高了0.87%。实验结果：(1) 说明残差结构可以在学习当前层鸟鸣声特征的同时避免丢失之前的信息，提高信息的复用率，引入了残差技巧和扩张卷积可以提高网络的识别效果；(2) 由于数据集本身在自然场景中获取包含一定噪声，因此在添加DRSN后识别正确率得到较大提高。

表3 消融实验结果

Table 3 Results of ablation experiments

(单位：%)

	Accuracy	F1-score
CNN+BiLSTM(baseline)	95.14	94.55
DilatedCNN+BiLSTM	95.64	95.22
DilatedSAM+BiLSTM	95.74	95.48
DSRN+DilatedSAM+BiLSTM	96.65	96.54

2.3.2 噪声实验

鸟鸣信号往往包含大量环境噪声，为了验证模型在噪声环境下的识别效果，本文设置了噪声实验，通过在原始数据库中添加高斯白噪声进行实验，来判断模型在噪声环境下的有效性。在实验中向数据库中分别加入不同强度的高斯白噪声，使原始信号与高斯白噪声的信噪比为−5 dB、−2 dB、0 dB、2 dB、5 dB和10 dB，并对比了log-Mel+CRNN[5]模型和3维log-Mel+DSRN+DilatedSAM+ BiLSTM的识别效果，同时为了验证本文引入的DSRN模块的有效性，实验也对比了在本文模型基础上去除DSRN模块的识别效果。

表4为不同信噪比下各个模型的识别正确率。从中可以看出：(1) 随着噪声强度的提高，3种方法识别精度都在降低。(2) 相比文献[5]采用的CRNN

方式，本文设计的基于扩张卷积和注意机制的残差连接模块 (DilatedSAM) 可以有效地在噪声环境下提取关键特征，在不同强度的背景噪声下均优于 CRNN。(3) 由于 DSRN 中软阈值操作的存在，模型可以有效将噪声特征值降低或置 0，因此该网络对于噪声有着良好的抑制作用，增加 DSRN 模块可以有效提高模型在噪声环境下的识别效果。

表 4　噪声实验结果

Table 4　Results of noise experiments

(单位：%)

	−5 dB	−2 dB	0 dB	2 dB	5 dB	10 dB
CRNN[5]	78.85	82.24	85.91	87.68	89.67	91.67
DilatedSAM +BiLSTM	83.44	86.76	88.48	90.19	91.82	92.71
DSRN+ DilatedSAM +BiLSTM	85.51	88.34	89.9.3	91.80	93.28	93.87

2.3.3　特征和其他模型对比实验

为了验证所提方法的有效性，本文进行了不同特征的对比实验，具体特征为：一维静态 MFCC 特征、包含动态信息的三维 MFCC 特征、一维静态 log-Mel 特征、包含动态信息的三维 log-Mel 特征。同时与其他学者的方法进行对比，log-Mel+CNN[7] 和 log-Mel+CRNN[5] 采用一维静态 log-Mel 频谱值作为输入特征，分别通过 CNN 和 CNN+GRU 模型进行识别；Mel 语谱图 +VGG16 提取鸟声信号的 log-Mel 特征并将其转化成尺寸为 256 × 256 语谱图图片，采用经典 VGG16 网络进行识别。

表 5 和表 6 为不同特征和算法识别正确率，从中可以看出：(1) 上述 4 种特征在不同网络上的识别结果差距较小，由 log-Mel 特征经过离散余弦变换得到的 MFCC 特征，在深度学习网络上的识别结果稍低于 log-Mel 特征，这可能是离散余弦变换操作造成了部分信息的丢失；结合动态信息的 3 维特征相较于仅包含静态信息的特征在不同模型上的提升并不明显，主要是由于深度学习网络可以有效地从静态特征中获取有效信息。(2) 本文所提算法识别效果相较于其他算法有着明显优势，识别正确率和 F1-score 得分分别可以达到 96.65% 和 96.54%。(3) 由于 DSRN 对于噪声信息的印制、残差技巧对于信息的复用和通过扩张卷积减少池化操作带来的信息丢失问题，本文所提的方法相较于其他网络获得了更好的识别效果。

表 5　特征对比实验结果

Table 5　Results of feature comparison experiments

(单位：%)

	MFCC	3 维 MFCC	log-Mel	3 维 log-Mel
CNN[7]	91.46	91.74	92.37	92.56
CRNN[5]	94.29	94.35	94.67	94.86
DSRN+ DilatedSAM +BiLSTM(our)	96.23	96.36	96.59	96.65

表 6　其他模型对比实验结果

Table 6　Results of other model comparison experiments

(单位：%)

	Accuracy	F1-score
log-Mel+CNN[7]	92.52	92.34
log-Mel+CRNN[5]	94.86	94.53
语谱图 +VGG16	93.24	92.87
Bi-LSTM-DenseNet[11]	92.48	—
3log-Mel+DSRN+ DilatedSAM+BiLSTM	96.65	96.54

3　结论

本文结合一些深度学习方法，提出了一种新的网络结构实现对噪声环境下鸟鸣声的识别，研究了如何从 log-Mel 频谱值中有效学习局部信息和全局信息。首先结合注意力机制的方法实现对噪声软阈值的自动确定，提出了一种改进的 DRSN；然后为了进一步提取有效特征，设计了一个基于扩张卷积和空间注意力机制的残差连接模块以获取更有效的局部特征；最后通过 BiLSTM，从局部特征中学习前后的依赖关系，获取全局特征。以北京百鸟数据库 20 类鸟声为实验对象结果表明：DRSN 中软阈值操作可以有效降低噪声干扰，相较于其他模型具备明显优势。因此本文模型在自然场景下具有良好的应用价值，可以有效降低环境中噪声干扰，提高识别正确率。在未来的研究中还会进一步探讨 DRSN 模块堆叠数量与对于不同强度噪声的抑制效果，从而将本文模型更好地应用于自然环境下的鸟声识别中。

参 考 文 献

[1] 赵洪峰, 雷富民. 鸟类用于环境监测的意义及研究进展 [J]. 动物学杂志, 2002(6): 74–78.
Zhao Hongfeng, Lei Fumin. Birds as monitors of environmental change[J]. Chinese Journal of Zoology, 2002(6): 74–78.

[2] 吴伟伟, 徐海根, 吴军. 气候变化对鸟类影响的研究进展 [J]. 生物多样性, 2012, 20(1): 108–115.
Wu Weiwei, Xu Haigen, Wu Jun. The impact of climate change on birds: a review[J]. Biodiversity Science, 2012, 20(1): 108–115.

[3] 徐淑正, 孙忆南, 皇甫丽英, 等. 基于 MFCC 和时频图等多种特征的综合鸟声识别分类器设计 [J]. 实验室研究与探索, 2018, 37(9): 81–86, 91.
Xu Shuzheng, Sui Yinan, Haungfu Liying, et al. Design of synthesized bird sounds classifier based on multi feature extraction classifiers and time-frequency chat[J]. Research and Exploration in Laboratory, 2018, 37(9): 81–86, 91.

[4] 张赛花. 面向鸟声传感网的鸟鸣自动分类方法研究 [D]. 南京: 南京理工大学, 2018.

[5] Cakir E, Adavanne S, Parascandolo G, et al. Convolutional recurrent neural networks for bird audio detection[C]. 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017: 1744–1748.

[6] 冯郁茜. 基于深度学习的双模态特征融合鸟类物种识别算法 [D]. 北京: 北京林业大学, 2019.

[7] Naranchimeg B, Zhang C, Akashi T. Cross-domain deep feature combination for bird species classification with audio-visual data[J]. IEICE Transactions on Information and Systems, 2019, E102. D(10): 2033–2042.

[8] 谢将剑, 杨俊, 邢照亮, 等. 多特征融合的鸟类物种识别方法 [J]. 应用声学, 2020, 39(2): 199–206.
Xie Jiangjian,Yang Jun, Xing Zhaoliang, et al. Bird species recognition method based on multi-feature fusion[J]. Journal of Applied Acoustics, 2020, 39(2): 199–206.

[9] Puget J. STFT transformers for bird song recognition[C]//2021 Conference and Labs of the Evaluation Forum(CLEF). Bucharest, Romania, 2021: 21–29.

[10] 邱志斌, 卢祖文, 王海祥, 等. 基于 Mel 频谱图和 CNN 的电网涉鸟故障鸟声识别 [J]. 华南理工大学学报 (自然科学版), 2022, 50(2): 129–136.
Qiu Zhibin, Lu Zuwen, Wang Haiyang, et al. Recognition of bird sounds related to power grid faults based on mel spectrogram and convolutional neural network[J]. Journal of South China University of Technology(Natural Science Edition), 2022, 50(2): 129–136.

[11] Liu H, Liu C, Zhao T, et al. Bird song classification based on improved Bi-LSTM-DenseNet network[C]//2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE). 2021: 152–155.

[12] Zhao M, Zhong S, Fu X, et al. Deep residual shrinkage networks for fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, 16(7): 4681–4690.

[13] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]. Berlin: Springer, 2018: 3–19.

[14] Wang Q, Wu B, Zhu P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11531–11539.

[15] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]// International Conference on Learning Representations (ICLR), 2016.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770–778.

[17] Meng H, Yan T, Yuan F, et al. Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network[J]. IEEE Access, 2019. 7: 125868–125881.

[18] Chen M, He X, Jing Y, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10): 1440–1444.

[19] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 44(8): 2011–2023.

[20] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735–1780.

[21] Jiang P, Fu H, Tao H, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition[J]. IEEE Access, 2019, 7: 90368–90377.

[22] 北京智源人工智能研究院. Birdsdata 数据集 [DB/OL]. [2021-03-10]. http://open.baai.ac.cn/data-set-detail/MTI2NDg=/NjQ=/true.