Computer Science and Application 计算机科学与应用, 2025, 15(9), 115-122
Published Online September 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.159229

Hans汉斯

基于深度学习模型的鸟类保护研究

魏倩楠，张志华，纪雨欣，李佳硕，刘洪源，全　尧

辽宁科技大学电子与信息工程院，辽宁 鞍山

收稿日期：2025年7月29日；录用日期：2025年8月28日；发布日期：2025年9月11日

摘　要

为了保护鸟类，维持森林生态系统的稳定，针对声音识别技术在鸟类保护系统中的重要性，本文旨在运用深度学习模型实现鸟类声音，伐木，枪声，电锯声的识别，首先通过对采集到的声音运用麦克风阵列法进行声源定位等预处理；其次对声音进行识别，运用梅尔倒谱系数(MFCC)进行特征提取是识别声音的关键步骤，模型测试结果表明，鸟声的识别率为90.2%，电锯和枪声的识别率为96.6%。

关键词

深度学习，模型，鸟类保护系统，声音识别，声源定位

Research on Bird Conservation Based on Deep Learning Models

Qiannan Wei, Zhihua Zhang, Yuxin Ji, Jiashuo Li, Hongyuan Liu, Yao Quan

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: Jul. 29th, 2025; accepted: Aug. 28th, 2025; published: Sep. 11th, 2025

Abstract

To protect birds and maintain the stability of forest ecosystems, and considering the significance of sound recognition technology in bird protection systems, this paper aims to apply a deep learning model to recognize bird sounds, logging sounds, gunshots, and chainsaw sounds. First, the collected sounds are preprocessed for sound source localization using the microphone array method. Second, for sound recognition, extracting features with Mel-Frequency Cepstral Coefficients (MFCC) is a crucial step. The model test results show that the recognition rate of bird sounds is 90.2%, and the recognition rate of chainsaw sounds and gunshots is 96.6%.

Keywords

Deep Learning, Model, Bird Protection System, Voice Recognition, Sound Source Localization

文章引用: 魏倩楠, 张志华, 纪雨欣, 李佳硕, 刘洪源, 全尧. 基于深度学习模型的鸟类保护研究[J]. 计算机科学与应用, 2025, 15(9): 115-122. DOI: 10.12677/csa.2025.159229

1. 引言

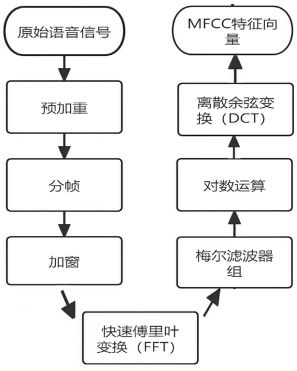
当今时代，人类对鸟类动物肆意捕杀将会加剧鸟类动物的灭绝速度。在进行生态系统保护的过程中融入科学发展观念，能够有效地遏制非法采伐和捕杀野生动物的行为发生。声音识别技术在鸟类保护系统中变得至关重要，扮演着重要的角色。声音识别技术可以实现对森林中声音的自动识别和定位，为鸟类保护和维持森林生态系统的稳定等领域提供了有效的解决方案。因此，声音识别技术尤为重要，本文将利用深度学习模型(Deep learning models)设计一种准确率较高的鸟类保护系统。声音定位识别的流程主要包括以下 4 个部分：数据收集与处理、特征提取、声源识别、声源定位。

2. 数据收集

在进行鸟类声音相关研究时，选定 xeno-canto 世界野生鸟类声音公开数据(https://xeno-canto.org/)作为鸟类音频文件数据源[1]。此数据集依据录音清晰程度，将数据分为了 a、b、c、d、e 共 5 个等级。利用 xeno-canto API，批量获取数据中质量达到 c 级及以上的声音数据。xeno-canto 数据集的所有音频均为没有经过任何加工的 MP3 格式，且包含环境噪声，时长在 10 s 到 5 min 不等，由于网络 数据来源存在差异，其采样率以及频率各不相同。为保持数据一致性，本研究把数据切割为 10 s 时长的音频，并转换为单声道的 wav 格式，重采样至 44.1 kHz [2]。同时从 Freesound.org 收集枪声和电锯声的数据。

3. 特征提取

本研究运用梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)融合特征参数，来提取信号的时频域特征，其突出特点是把信号从线性频域转换到梅尔域[3]。实现对鸟声信号的分类，相关提取流程如图 1 所示。



图注：展示梅尔频率倒谱系数(MFCC)提取流程，从原始语音信号出发，经预加重(补偿高频、提升信噪比)、分帧(切分连续信号为短帧)、加窗(减少帧边缘突变)、快速傅里叶变换(FFT，转时域为频域)、梅尔滤波器组(模拟人耳听觉，提取临界带特征)、对数运算(压缩动态范围)、离散余弦变换(DCT，去相关性)，最终得到 MFCC 特征向量，用于语音识别、情感分析等任务。

Figure 1. Flowchart of MFCC feature extraction
图 1. 梅尔倒谱系数(MFCC)提取流程图

1) 预加重

为实现增强高频信号、抑制环境噪声并提升信噪比的目标，采用一阶高通滤波器增强高频成分，其工作原理可通过差分方程计算公式与传递函数描述，对应形式如式(1)、式(2)所示。

$$y[n]=x[n]-\alpha\cdot x[n-1]\left(\text{通常}\alpha=0.97\right) \tag{1}$$

$$H(z)=1-a\cdot z^{-1} \tag{2}$$

2) 分帧

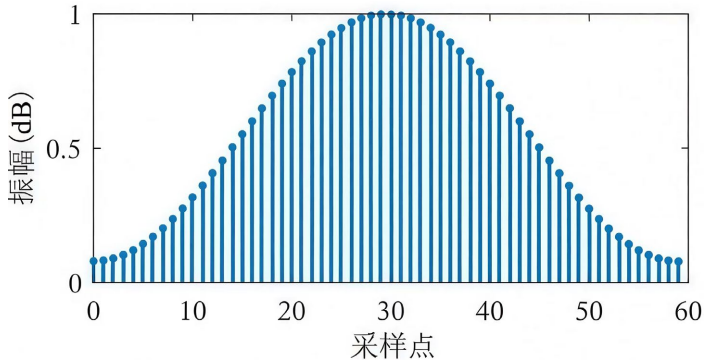
鸟声帧长为 20~40 ms (如 160 点@8 kHz 采样率)，电锯和枪声帧长为 46 ms。帧移通常为 50%。

3) 加窗

汉明窗公式：

$$W(n)=\begin{cases}0.54-0.46\cos\left[\dfrac{2\pi n}{N-1}\right],0\le n\le N-1\\ 0,otherwise\end{cases} \tag{3}$$

汉明窗幅度示意图如图 2。



图注：(N 为帧长，采样点对应窗序列索引)的幅度分布。其“钟形”曲线用于语音信号处理加窗环节，可平滑帧边缘、抑制信号突变，为后续快速傅里叶变换(FFT)等操作提供预处理支撑。

Figure 2. Hamming window amplitude response diagram
图 2. 汉明窗幅度相应示意图

减少帧边缘突变。

4) FFT

计算每帧的功率谱：$|\text{FFT}(x)|^2$。

5) 梅尔滤波器组

将线性频率转换为梅尔刻度：

$$mel(f)=2595\cdot\log10(1+f/700) \tag{4}$$

三角滤波器在梅尔刻度上均匀分布，叠加于功率谱。

6) DCT

对对数滤波器能量做 DCT，解相关并压缩信息：

$$c[n]=\sum\log(E\_k)\cdot\cos\left(\pi n(k-0.5)/N\right) \tag{5}$$

7) 动态特征

一阶差分(Δ)和二阶差分(ΔΔ)可增强时序信息。

4. 声音识别

采用 ResNetSE (带注意力机制的残差网络)作为核心模型架构，在完成语音信号预处理后，基于注意力机制深度学习完成语音特征信息的提取。注意力机制深度学习由神经元网络层级堆叠而成，每一层均负责捕捉并增强语音分析细节。这些神经元网络不仅广泛连接，还融入了注意力机制，能够动态聚焦于语音信号中的关键片段，忽略无关噪声[4]。通过梅尔频谱图特征提取和深度学习技术实现高精度音频识别分类。系统支持 WAV、MP3、OGG 等多种音频格式的自动处理，提供从模型训练到实际预测的全流程解决方案，包含命令行和图形界面两种操作方式，可广泛用于语音识别。

5. 声源定位

聚焦森林鸟类保护及生态监测的实际需求，搭建基于到达时间差(Time Difference of Arrival, TDOA)的麦克风阵列声源定位体系，用于对森林鸟类和环境声音开展精准化定位工作。相较于传统的声源定位方法，TDOA 技术具备更高的定位精度与更强的抗干扰性。能够达成语音控制设备的精准识别与定位[5]。TDOA 算法的核心逻辑在于，利用声音信号传播至多个麦克风时产生的时间差异来确定声源的方位。当声源发出声音信号后，该信号会在不同的麦克风间产生细微的传播时间差。通过核算麦克风阵列间的时间差，并结合其几何结构关系，即可实现对声源位置的估计。

5.1. 基于广义互相关(GCC)计算时延

依据测量环境条件以及信号自身特性，可对时延估计方法进行分类，涵盖相位法、双谱法、相关法、自适应滤波器参数模型法等。在这些方法中，相关法属于最为经典且应用最广泛时延估计方法[6]。

当存在噪声时，对于由远处声源发出的，并被处于不同空间位置两个麦克风监听的信号，可以数学建模为：

$$x_1 = s_1(t) + n_1(t) \tag{6}$$

$$x_2 = \alpha s_1(t - D) + n_2(t) \tag{7}$$

其中，$s(t)$代表声音信号，$n_1(t)$、$n_2(t)$为两个声音传感器检测到的噪声，三者均属于稳定的随机过程，且相互之间不存在相关性，计算 x_1 与 x_2 的互相关函数：

$$R_{x_1x_2}(\tau) = E\left[x_1(t) * x_2(t - \tau)\right] \tag{8}$$

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{T - \tau}\int_{\tau}^{T} x_1(t)x_2(t - \tau)\mathrm{d}t \tag{9}$$

其中，其中估计的时延 D 为互相关函数 τ 值达到最大值时取得的值，即：

$$\hat{D} = \arg\max R_{x_1x_2}(\tau) \tag{10}$$

5.2. 麦克风阵列选型与布局

采用三维阵列心型指向性动圈式麦克风，心型指向性可增强目标方向声音采集、抑制环境噪声。阵列布局为等间距线性排列，设麦克风数量为 N，间距为 d，依据森林监测范围(如单只鸟类活动半径、群落分布跨度)，确定 d 为 0.2~0.5 m，平衡空间分辨率与信号串扰。

在着手对麦克风阵列开展建模工作前，我们需要明晰进场与远场的概念。顾名思义，离麦克风近则

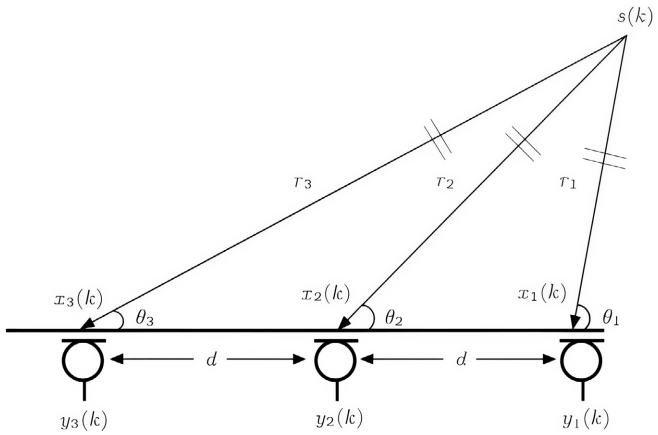
符合进场远，离得远则符合远场模型。

假设 L 代表阵列间距，λ 是声波波长，M 是声源与麦克风的距离，我们把 $\frac{2L^2}{\lambda}$ 设定为远场和近场临界判定值。

当 $M < \frac{2L^2}{\lambda}$ 时，属于近场模型，这时声源传播到麦克风阵列的波形可看作球面波。

当 $M > \frac{2L^2}{\lambda}$ 时，属于远场模型，这时声源传播到麦克风阵列的波形可看作平面波。

5.2.1. 近场模型



图注：三个等间距(间距 d)麦克风线性排列，$s(k)$为远场声源，r_i 是传播路径，θ_i 为入射夹角，用于 DOA 估计及声源定位算法推导。

Figure 3. Structure schematic of near-field model (spherical wave) based on three-microphone array
图 3. 基于三麦克风阵列的近场模型(球面波)结构示意

当 $M < \frac{2L^2}{\lambda}$ 时，符合近场模型，此时声源到达麦克风阵列的波形视为球面波。

近场模型的构建至少需要三个麦克风参与，这里以基础的三麦克风模型(如图 y1，y2，y3 所示)展开分析，如图 3 所示。设定 τ_{12}，τ_{13}、三分别对应第一个麦克风和第二、第三个麦克风间的时延，此时存在如下关系：

$$\tau_{12} = \frac{\gamma_1 - \gamma_2}{c} \tag{11}$$

$$\tau_{13} = \frac{\gamma_3 - \gamma_1}{c} \tag{12}$$

式中，C 代表声速，在标准大气压、15˚环境条件下，声速取值为 340 m/s。

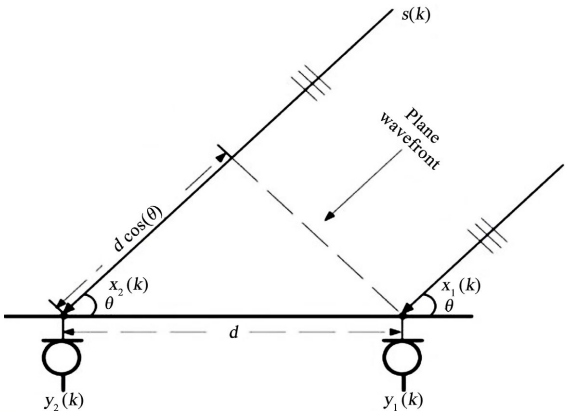
基于麦克风阵列的几何布局，利用余弦定理能够推导出：

$$r_2^2 = r_1^2 + d^2 + 2r_1 d\cos\theta_1 \tag{13}$$

$$r_3^2 = r_1^2 + 4d^2 + 4r_1 d\cos\theta_1 \tag{14}$$

其中 τ_1，τ_2，τ_3 可以借助互相关 GCC 算法获取，且 c (声速)，与 d (阵列相关参数)为已知量，将这些已知条件代入上述公式，便能求解出未知参数 r_1，r_2，r_3，θ，再结合坐标系，就可以确定 $s(k)$的坐标信息。

5.2.2. 远场模型



图注：三个等间距(间距 d)麦克风线性排列，$s(k)$为远场声源，r_i是传播路径，θ_i为入射夹角，用于 DOA 估计及声源定位算法推导。

Figure 4. Schematic of the far-field model (plane wave) structure based on a three-microphone array
图 4. 基于三麦克风阵列的远场模型(平面波)结构示意

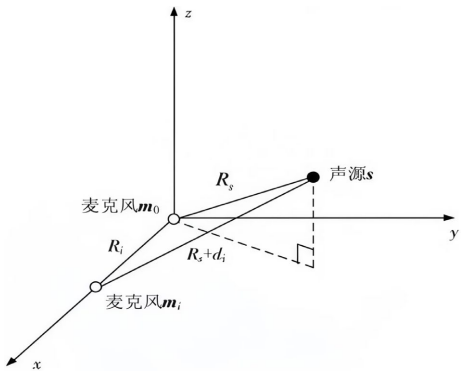
当 $M > \frac{2L^2}{\lambda}$时，满足远场模型条件，此时可将声源传播至麦克风阵列的波形近似看作平面波。

针对仅配置两个麦克风的情形，由于其自身特性限制，仅能够求解方位角，而无法确定方位距离，如图 4 所示。若设 τ 代表声波抵达两个麦克风的时延差，那么存在如下关系：

$$\tau = \frac{d\cos\theta}{c} \tag{15}$$

$$\theta = ar\cos\frac{c\tau}{d} \tag{16}$$

5.3. 三维空间阵列的声源定位系统实现



图注：图中以麦克风 m_0 为坐标原点建立空间直角坐标系，x 轴和 y 轴为水平平面内相互垂直的维度，z 轴为垂直于水平平面的维度。$m_0, m_i(i=1,2...)$表示三维麦克风阵列中的传感器节点，S 为近场声源。模型中，R_0 为声源 S 到 m_0 的距离，R_i 为声源 S 到 m_i 的距离，二者差值与声速 c、传播时延 τ(如公式(15))及方位角(如公式(16))相关联，为推导三维空间内声源坐标下，y，z 的定位算法提供几何依据。

Figure 5. Near-field sound source localization model based on 3D array (taking m_0 as origin)
图 5. 基于三维阵列的近场声源定位模型(以 m_0 为原点)

一个三维麦克风阵列，麦克风分别为 $m_i(i=0,1\cdots n)$，声源 S 符合近场模型。现以麦克风 m_0 为原点，如图 5 所示，建立直角坐标系。推导公式之前，需要先确定以下概念，见表 1：

Table 1. Relationship table of concepts and characters in near-field sound source localization model with 3D array
表 1. 基于三维阵列的近场声源定位模型概念与字符对应关系表

序号	概念	字符
1	麦克风的坐标	Mi i⊂[0, n]
2	声源的估计坐标	S
3	声源 s 到 mi 与 $m1$ 的估计距离差	$\hat{d}i$
4	麦克风 m_i 到原定的距离	R_i
5	声源 s 到原定的距离	R_s

如上图根据三角形 M0，MIS，由余弦定理有：

$$\left(R_S+d_i\right)^2=R_i^2-2m_i^Ts+R_S^2 \tag{17}$$

展开公式(17)得：

$$R_i^2-2m_i^Ts-2Rsd_i-d_i^2=0 \tag{18}$$

由于 di 是通过估计的时延得到的，与实际值之间会有一个偏差，因此公式(15)不为 0，可写为：

$$\varepsilon=\left(R_i^2-d_i^2\right)-2m_i^Ts-2R_sd_i \tag{19}$$

此时已经得到目标值的误差函数，使用最小二乘法求解[7]，问题可以转化为：估计声源坐标 s (x, y, z)，使得最终的误差平方和最小，即：

$$\arg\min\sum_{i=1}^{n}\left[\left(R_i^2-d_i^2\right)-2m_i^Ts-2R_sd_i\right]^2 \tag{20}$$

将公式(19)写成矩阵形式：

$$\varepsilon=2R_s\hat{D}+2Ms-\delta \tag{21}$$

$$M=\begin{bmatrix}m_1\\m_2\\m_3\\....\\m_n\end{bmatrix}=\begin{bmatrix}x_1&y_1&z_1\\x_2&y_2&z_2\\x_3&y_3&z_3\\...&...&...\\x_n&y_n&z_n\end{bmatrix}\hat{D}=\begin{bmatrix}\hat{d}_1\\\hat{d}_2\\\hat{d}_3\\...\\\hat{d}_n\end{bmatrix}\delta=\begin{bmatrix}R_1^2-\hat{d}_1^2\\R_2^2-\hat{d}_2^2\\R_3^2-\hat{d}_3^2\\.....\\R_n^2-\hat{d}_n^2\end{bmatrix} \tag{22}$$

公式(22)可以化简为：

$$\varepsilon=A\mu-b \tag{23}$$

其中，$A=\begin{bmatrix}M&\hat{D}\end{bmatrix}\mu=\begin{bmatrix}s\\R_s\end{bmatrix}b=\frac{1}{2}\delta$。

公式(23)最小二乘解可以表示为：

$$\hat{\mu}=\left(A^TA\right)^{-1}A^Tb \tag{24}$$

公式(24)即为计算结果，下面对结果进行进一步化简：

1) 定义沿到 D 的投影矩阵为：

$$P_A = \frac{\hat{D}\hat{D}^T}{\hat{D}^T\hat{D}} \tag{25}$$

2) 沿 D 到 A 的投影矩阵即为：

$$P_D = I - P_A = I - \frac{\hat{D}\hat{D}^T}{\hat{D}^T\hat{D}}\left(I\text{为单位矩阵}\right) \tag{26}$$

3) 根据投影矩阵的性质，可以得到：

$$\mathrm{A} = \mathrm{P_D}\left[\mathrm{M}\ 0\right] \tag{27}$$

4) 将公式(27)带入(23)中得到：

$$\varepsilon = P_D Ms - b \tag{28}$$

5) 公式(28)的最小二乘解即为最终简化结果。

最终化简结果为：

$$s = \left(M^T P_D M\right)^{-1} M^T P_D b \tag{29}$$

6. 结束语

本文充分梅尔倒谱系数模块实现对音频进行处理，实现对提取信号的时频域特征，实现对鸟声信号，电锯声信号，枪声信号的识别搭建卷积神经网络模型，模型识别准确率较高。在整个声音识别流程中发现，对数运算和离散余弦运算非常重要，会严重影响到声音识别的效果。此外，本文对声音信号搭建了神经网络模型，测试结果表明鸟声的识别率为 90.2%，电锯和枪声的识别率为 96.6%，显示鸟声的识别率比较低。在后期的研究中会对神经网络结构进行调整，并进一步优化。

基金项目

辽宁科技大学大学生创新创业训练计划项目(X202510146100)。

参考文献

[1] 吴天佑. 基于多特征融合的鸟类声音识别方法研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2022.

[2] 刘昱坤, 邓广, 于新文, 等. 基于注意力机制的鸟类声音识别模型——以神农架国家公园为例[J]. 陆地生态系统与保护学报, 2024, 4(3): 39-48.

[3] 陈东, 黄智鹏. 基于梅尔频率倒谱系数和支持向量机的汽车鸣喇叭声识别[J]. 科学技术与工程, 2021, 21(11): 4486-4491.

[4] 朱玉, 李枫. 基于注意力机制深度学习的电力通信网络电话语音识别方法[J]. 科技创新与生产力, 2025, 46(1): 104-106, 110.

[5] 杨曼, 任志国, 薛盼盼. 基于改进鸟群优化算法的 TDOA 定位算法[J]. 兰州文理学院学报(自然科学版), 2025, 39(4): 48-55.

[6] 陈晓辉, 孙昊, 张恒, 等. 基于声源阵列的空间麦克风定位方法研究[J]. 计算机应用研究, 2020, 37(5): 1437-1439, 1444.

[7] 彭烁钟, 张俊杰, 强君宝. 基于 TDOA 的声源定位技术研究与应用[J]. 智能计算机与应用, 2024, 14(12): 163-169.