摘　　要

本文基于美国国家海洋和大气管理局(NOAA)、康奈尔大学、纽约大学在 kaggle 平台上共同发起的 ICML 2013 露脊鲸识别挑战赛数据集，在 STFT 和图像识别结合的上曲叫声检测方法[2]基础上，提出一种结合 STFT 提取时频特征，维纳滤波、小波阈值进行时频特征增强，输入 CNN 分类器学习的方法，处理露脊鲸海洋声学数据集。结果表明，低信噪比的情况下，能够有效地检测低信噪比的上曲叫声，在自己划分的验证数据集（约 10000 条）上达到 98.3%的 acc。

关键词：露脊鲸　STFT　Wiener滤波 小波阈值降噪　CNN分类

目　　录

1 引言

露脊鲸是一种由于滥捕濒临绝灭的海洋哺乳动物，现经国际协议进行保护。主要分布在太平洋、大西洋等海域，北太平洋现存约 1000 只，北大西洋仅约 100 只 [1]。为此，美国国家海洋和大气管理局(NOAA)、康奈尔大学、纽约大学在 kaggle 平台上共同发起的 ICML 2013 露脊鲸识别挑战赛[1]，旨在提出有效和高效的算法自动化地处理大型海洋声学数据集，从中检测出海洋哺乳动物的叫声，以更好地了解海洋哺乳动物的生物声学行为。

其中海洋声学数据集由康奈尔大学提供，包含了船只搭载的 Auto-buoy 水中听音器监测平台或 MARU 平台于北大西洋超过十年的大量采集数据，这些数据跨越几个海洋盆地，涵盖了各种海洋哺乳动物物种。而本课题使用 kaggle 平台上公开的包含露脊鲸标注的海洋声学数据，旨在设计准确的露脊鲸叫声分类器。

为了与同类远距离交流，露脊鲸的能发出 50~250HZ 频带频率上曲的叫声，记作上曲叫声。由于采集海洋声音含有很大的底噪，且混有其他海洋物种的叫声，上曲叫声的信噪比很低，给检测带来了困难。为此，有学者提出了基于 STFT 和图像识别结合的上曲叫声检测方法[2]。但该文章提出的方法需要使用改进的 Moore 邻域法进行连续区域处理，提取感兴趣的上曲叫声频谱轮廓，并将处理后的时频图进行均值池化，同时利用设计的 3 种二值特征掩模进行平均池化，提取出 20 种对角特征频谱输入后续分类器并进行学习。

上述特征提取过程现今看来过于繁琐，因而本文在其基础上提出一种结合 STFT 提取时频特征，维纳滤波、小波阈值进行时频特征增强，输入 CNN 分类器学习的方法，处理露脊鲸海洋声学数据集。结果表明，低信噪比的情况下，能够有效地检测低信噪比的上曲叫声，在自己划分的验证数据集（约 10000 条）上达到 98.3%的 acc。

2 方法和原理

2.1 STFT

由于时间有限，在此不再赘述。

2.2 维纳滤波

[1] The ICML 2013 Whale Challenge - Right Whale Redux
数据集 https://www.kaggle.com/c/the-icml-2013-whale-challenge-right-whale-redux

由于时间有限，在此不再赘述。

2.3 小波阈值降噪

小波变换是一种信号的时间-尺度（时间-频率）分析方法，通过尺度函数可以进行多分辨率分析，在时频两域都具有表征信号局部特征的能力，是一种窗口大小固定不变但其形状可改变，时间窗和频率窗都可以改变的时频局部化分析方法。即在低频部分具有较低的时间分辨率和较高的频率分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率，适合于分析非平稳的信号和提取信号的局部特征[3]。

在图像分析领域，小波变换把图像分解成逼近图像和细节图像之和，它们分别代表了图像的不同结构，对图像进行一次小波分解后，可分为 LL，LH，HL 和 HH 子频带。其中 LL 反映的是水平和竖直方向的低频信息；LH 反映的是水平方向的低频信息和垂直方向的高频信息；HL 反映的是水平方向的高频信息和垂直方向的低频信息；HH 反映的是水平和竖直方向的高频信息。将小波变换用于图像多尺度分解，优点是重构后没有信息损失，缺点对小波选取要求高。

将小波变换用于软阈值降噪，是为了弥补硬阈值降噪去噪后产生局部的抖动的缺点。由于信号在信号空间上是连续性的，因此在小波域，有效信号所产生的小波系数其模值往往较大。而高斯白噪声则是非连续性的，因此噪声经过小波变换分布近似为高斯。因而在小波域，有效信号对应的系数很大，而噪声对应的系数很小。因此，只要将区间三倍方差内的系数置零，就能有效抑制噪声的，同时将经过阈值处理后的小波系数重构，就能得到降噪后的信号[3]。

2.4 CNN 分类器

由于这方面不属于书本内容，不在此不再详述

3 实验过程及现象分析

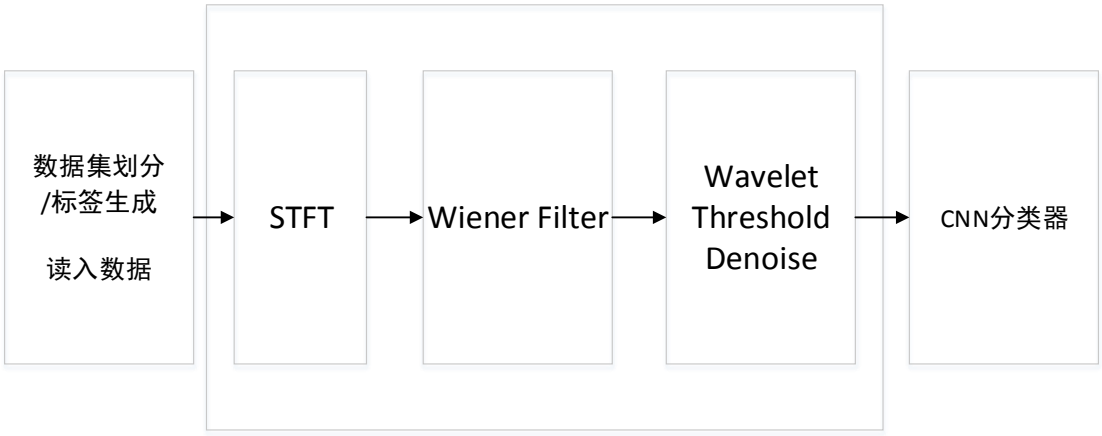

图 3.1　实验流程框图

本节主要介绍提出地识别露脊鲸上曲叫声方法的实现流程细节，实验流程框图如图 3.1 所示。

3.1 数据集介绍/音频帧基础概念

图 3.2　实验数据集示意

本文采用的海洋音频数据的训练集包含 42565 负样本，5276 正样本，且含有噪声标注。

内为采样率为 2kHZ、持续时间不一（1~2s，2000~4000 个采样点）的.aif 无损音频文件，编码格式为 PCM（未经编码的音频数据）。由于该格式未经编码压缩，因而一音频帧包括 *nchannels*samplesize* bytes[4]，这里 *nchannels* 为 1，*samplesize* 为 8bit(1Bytes)，因而 *aifc.readframes* 读取最大帧数 *n_frame* 即采样点数 4000。

输入音频的二维波形图 3. 3 如下所示：

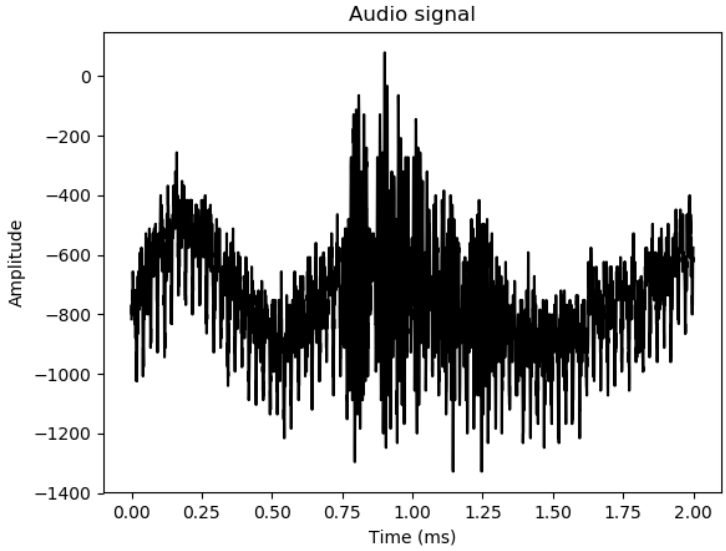

图 3.3　海洋音频数据波形图

3.2 使用 STFT 提取二维时间-幅值频谱图

参考[2]，本文采用如下参数：*128 ms*的*Hann*窗函数（即256采样点，*50%*重叠)，*800HZ*采样率以提取*0~400HZ*的有效信息（露脊鲸叫声频带范围），对输入数据进行STFT变换，生成二维时间-幅值频谱图，后续简称二维时频谱图。通过STFT输入维数由*(batch_size, 4000)*变为*(batch_size, 129, 33)*。

时间-幅值频谱可视化成三维如图 3.4 所示，可见露脊鲸上曲叫声频谱被直流分量和部分低频分量等窄带噪声所掩盖，其中窄带噪声可能包括船航行时与水面碰撞、船内发动机发出的声音以及风声，平面上还存在波浪状宽带高斯噪声。因而要增强上曲叫声，首先要考虑去除直流分量，否则分类器无法学到有效信息。

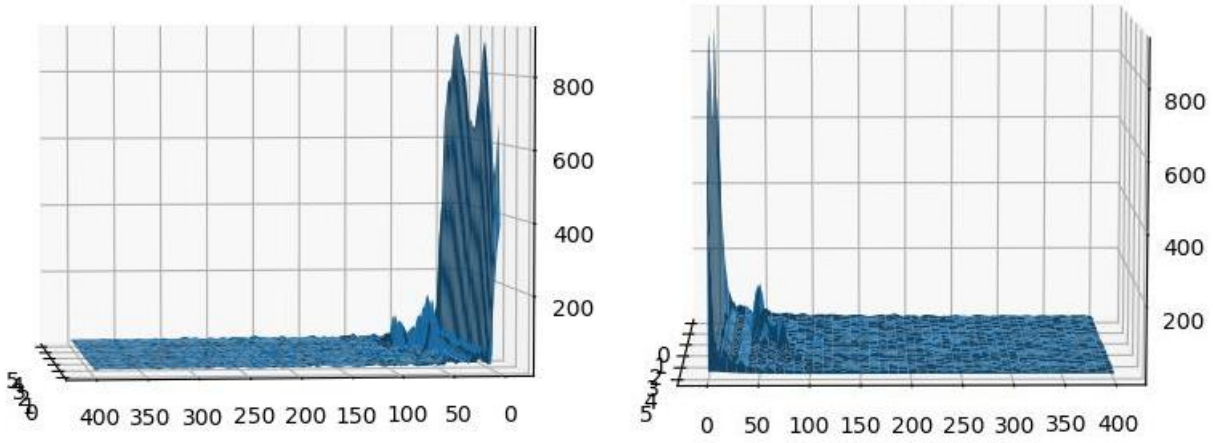

图3.4　三维时间-幅值频谱示意

3.3 频谱标准化和去噪

为去除频谱的低频窄带分量，首先对频谱进行标准化操作。这方面尝试了很多不同的标准化方法。首先是全局标准化，标准化公式如式（3-1）所示，标准化后的三维谱图如图3.5所示。

$$x_N = (x - x_{global\_mean}) / (x_{global\_max} - x_{global\_min}) \qquad (3\text{-}1)$$

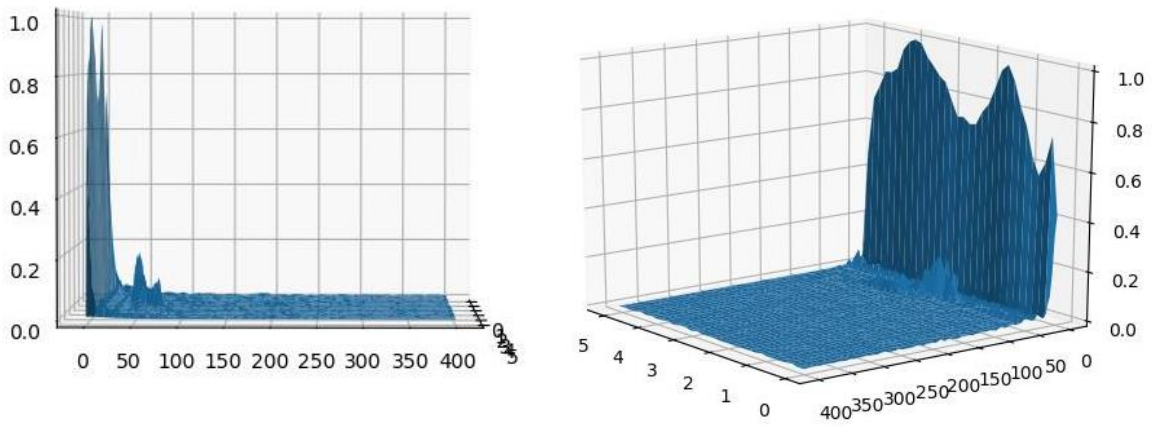

图3.5　全局标准化后的三维谱图

如上图所示，标准化前后三维谱图几乎没有变化。可见全局标准化方法没能有效去除直流分量。究其原因，STFT 幅值主要少部分分布在 50~250HZ 的上曲时频分量和窄带低频分量。若直接进行全局标准化，由于除上曲时频分量外大部分区域 STFT 幅值很小，因而加权下来全局均值很小。上曲时频分量和窄带低频分量反而相当于离群点，不能很好的对直流分量进行标准化。

因而可提出设想，使用直流分量的均值进行标准化。当然这样做可以有效去除直流分量，但由于直流分量相对上曲叫声分量差距较大，会淹没其有效的 STFT 特征信息。

最后采用的方法是对每个频率点进行标准化，标准化公式如式（3-2）所示。这样既不会淹没上曲叫声分量的 STFT 特征，又能有效去除直流分量，但缺点是放大了非上曲时频分量区域的噪声，因而需要进一步进行降噪。

$$x_N = (x - x_{per\_fre\_mean}) / (x_{per\_fre\_\max} - x_{per\_fre\_\min}) \qquad (3\text{-}2)$$

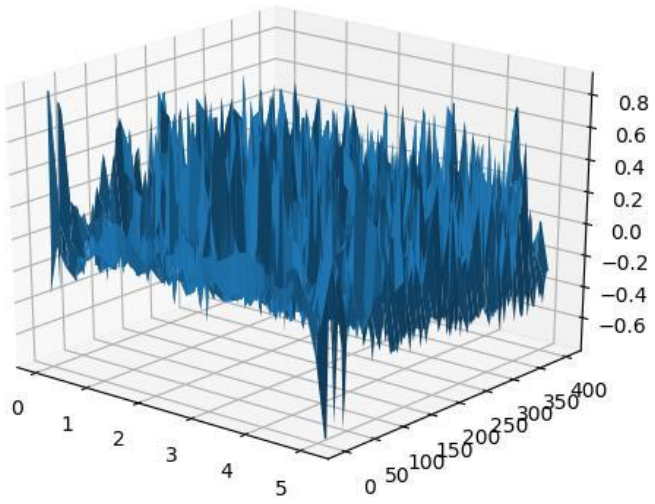

图 3.6　局部标准化后的三维谱图

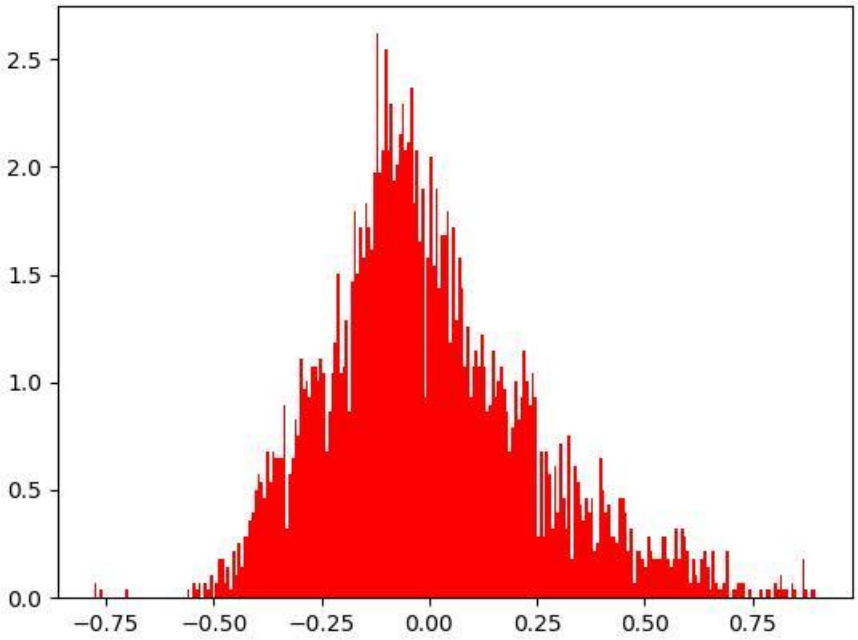

图 3.7　局部标准化后的直方图分布

局部标准化后的三维谱图如图 3.6 所示，直方图分布如图 3.7 所示。

如图 3.6 所示，可假设频谱空间上的噪声为高斯分布。在假定噪声为高斯分布的情况下，可利用 Wiener 滤波器可根据最小方差原则，用于滤除高斯噪声得到最优估计。使用维纳滤波是为了平滑频谱，提取轮廓特征。这里采用窗大小为（*5*，*5*）的二维维纳滤波器得到的结果如图 3.8、图 3.9 所示，可见确实平滑了噪声（限制了噪声的幅度），增强了上曲叫声。

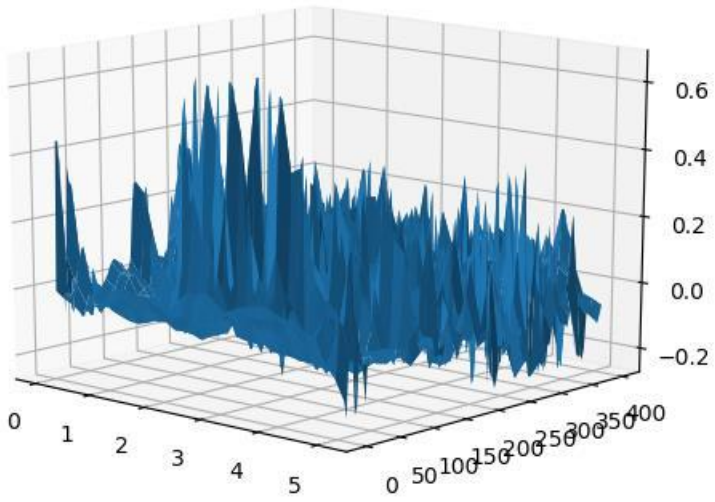

图 3.8　维纳滤波后的三维谱图

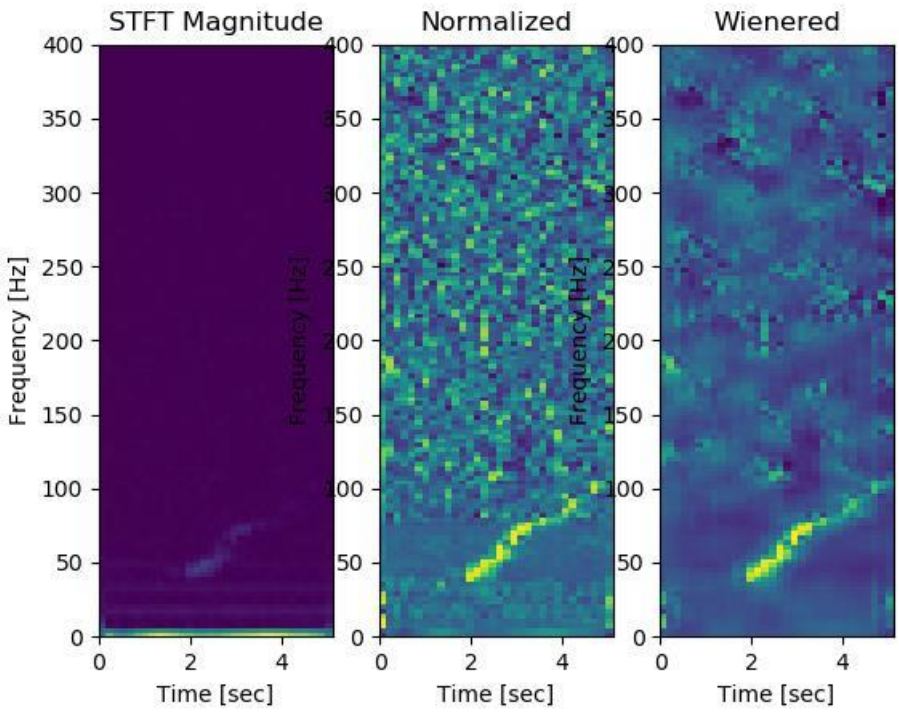

图 3.9　维纳滤波后的二维谱图

接下来需要进一步对噪声进行抑制，这里探讨两种方法：（1）直方图均衡；（2）小波阈值均衡。

直方图均衡化是指将一幅图像的灰度直方图变平，使变换后的图像中每个灰度值的分布概率都相同，可以增强图像的对比度[5]。直方图均衡后的结果如图 3.10 所示，可见既增强了上曲叫声，也增强了我们不想要的时频成分（可能是其他海洋物种的声音）。

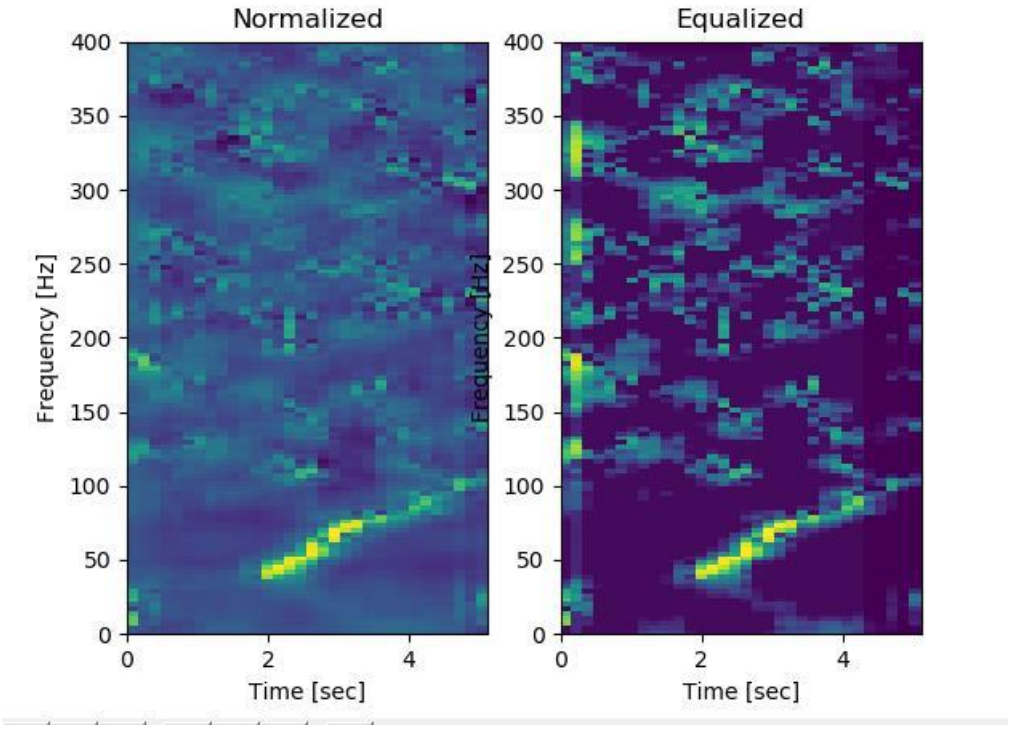

图 3.10　直方图均衡后的二维谱图

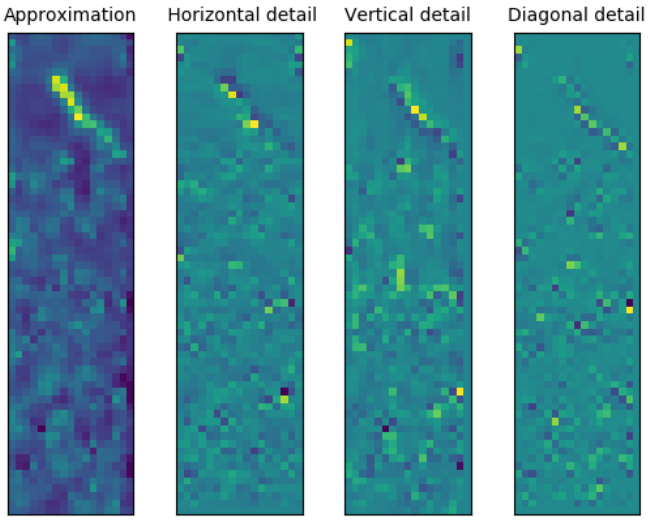

图3.11　小波分解

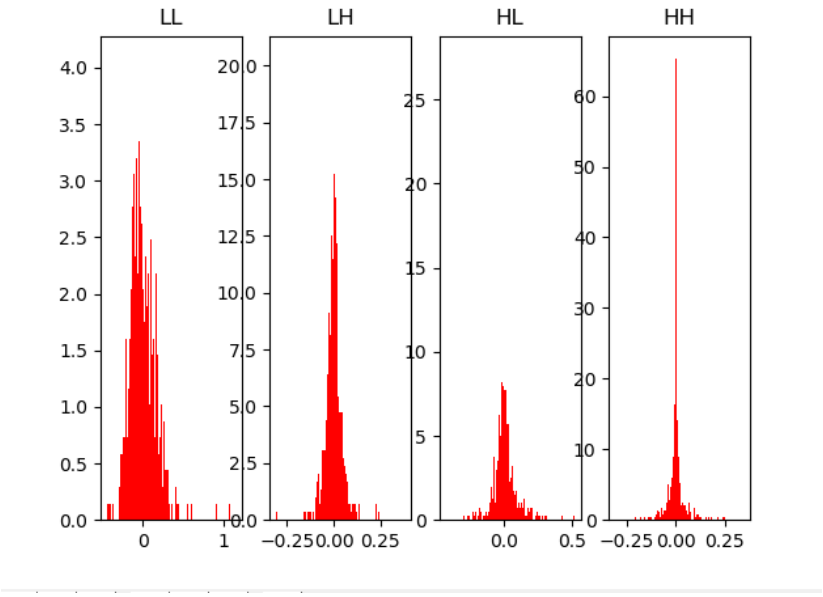

图3.12　小波分解

小波阈值降噪本质上是对信号进行了subsampling并进行滤波并重构，小波分解后的频谱如图3.11和图3.12所示。这里使用*pywt*小波分析包进行小波软阈值软置换，阈值为*0.15*，降噪结果如图3.13所示。

综上所述，本文采用小波阈值降噪的方法。

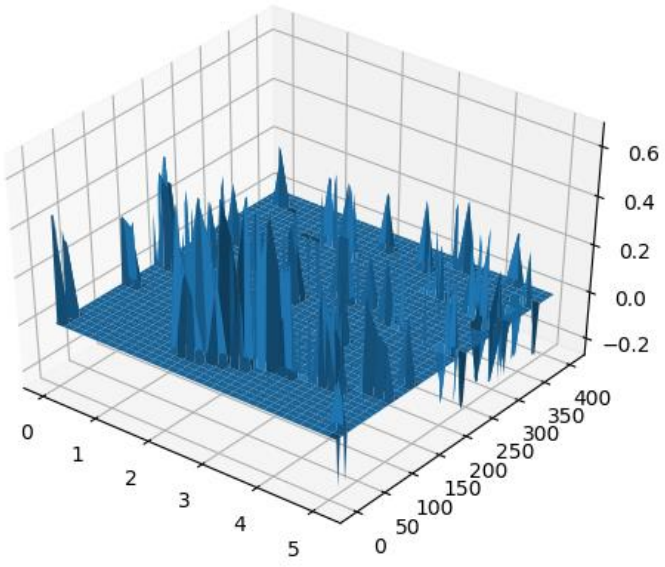

图3.13　小波阈值降噪

至此完成了最重要的一步——对输入音频序列提取时频特征谱图并进行预处理，后续只需将其输入分类器即可。

3.4 分类器

本文采用CNN分类器，卷积神经网络应用于时频谱图检测上曲叫声的优势在于：（1）具有尺度不变性，在第二节里提到输入音频的长度是不唯一的，根据CNN的特性可以resize同样大小的时频谱图再作为输入。同时由于上曲的角度是不固定，因而对于CNN池化层的旋转不变性。（2）由于数据集只给了标签信息而没有具体到频谱，所以最好通过二分类的方法解决。当然也可以手动构建特征，再进行二分类，端到端的方式可以节省很大人工开销。

首先进行数据集的划分，根据0.2比例随机划分验证集。根据文件名生成训练集和验证集标签并写入txt文件，如图3.14所示。数据读入和预处理代码根据*tf.data API*编写。

```
./data/train2/20090328_000000_002s3ms_TRAIN0_0.aif 0
./data/train2/20090328_000000_010s8ms_TRAIN1_0.aif 0
./data/train2/20090328_000000_021s6ms_TRAIN2_0.aif 0
./data/train2/20090328_000000_059s0ms_TRAIN3_0.aif 0
./data/train2/20090328_000000_068s4ms_TRAIN4_0.aif 0
./data/train2/20090328_000000_076s8ms_TRAIN5_0.aif 0
./data/train2/20090328_000000_090s4ms_TRAIN6_0.aif 0
./data/train2/20090328_000000_095s6ms_TRAIN7_0.aif 0
./data/train2/20090328_000000_103s2ms_TRAIN8_0.aif 0
./data/train2/20090328_000000_107s4ms_TRAIN9_0.aif 0
./data/train2/20090328_000000_118s3ms_TRAIN10_0.aif 0
./data/train2/20090328_000000_120s9ms_TRAIN11_0.aif 0
./data/train2/20090328_000000_154s6ms_TRAIN12_0.aif 0
./data/train2/20090328_000000_167s1ms_TRAIN13_0.aif 0
./data/train2/20090328_000000_169s5ms_TRAIN14_0.aif 0
./data/train2/20090328_000000_174s0ms_TRAIN15_0.aif 0
./data/train2/20090328_000000_189s6ms_TRAIN16_1.aif 1
./data/train2/20090328_000000_210s2ms_TRAIN17_0.aif 0
./data/train2/20090328_000000_213s2ms_TRAIN18_0.aif 0
./data/train2/20090328_000000_215s2ms_TRAIN19_0.aif 0
./data/train2/20090328_000000_220s7ms_TRAIN20_0.aif 0
./data/train2/20090328_000000_222s8ms_TRAIN21_0.aif 0
./data/train2/20090328_000000_228s2ms_TRAIN22_0.aif 0
./data/train2/20090328_000000_230s0ms_TRAIN23_0.aif 0
./data/train2/20090328_000000_234s0ms_TRAIN24_0.aif 0
./data/train2/20090328_000000_236s4ms_TRAIN25_1.aif 1
./data/train2/20090328_000000_246s0ms_TRAIN26_0.aif 0
./data/train2/20090328_000000_256s8ms_TRAIN27_0.aif 0
./data/train2/20090328_000000_260s7ms_TRAIN28_0.aif 0
./data/train2/20090328_000000_264s9ms_TRAIN29_0.aif 0
./data/train2/20090328_000000_273s9ms_TRAIN30_0.aif 0
./data/train2/20090328_000000_288s1ms_TRAIN31_0.aif 0
./data/train2/20090328_000000_301s9ms_TRAIN32_0.aif 0
./data/train2/20090328_000000_316s7ms_TRAIN33_0.aif 0
./data/train2/20090328_000000_324s7ms_TRAIN34_0.aif 0
```

图3.14　划分的数据集标签示意

本文采用的CNN而分类器结构如图3.15所示，由二层*[5, 5]*卷积层，第一层*featuremap=32*，第二层*featuremap=64*，插夹二层*stride=2*的*[2, 2]*池化层，dropout层和logit层和softmax分类层构成，其中*dropout rate=0.4*。

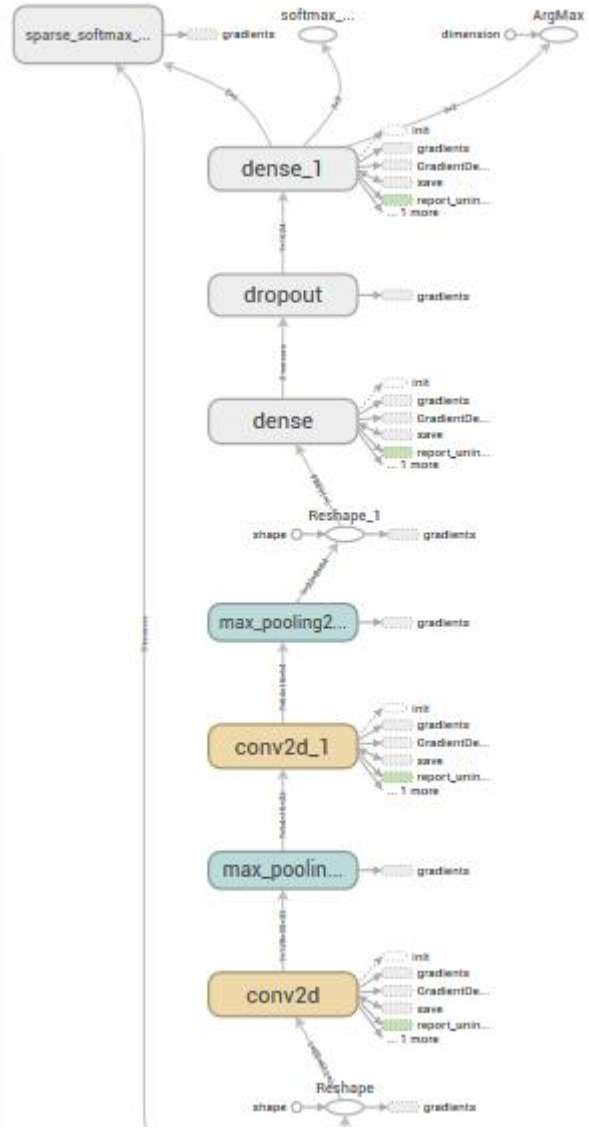

图3.16　采用的CNN分类器结构

训练和评估根据*tf.estimator* API编写，使用*SGD*优化器，训练参数为*learning rate=0.01*，*batch size=4* 训练*steps=108000*时loss曲线和验证集acc曲线如图3.17、图3.18所示。可以看到在验证集的检测准确率虽步长最高达到98.3%，说明输入的特征频谱是稳定有效的。

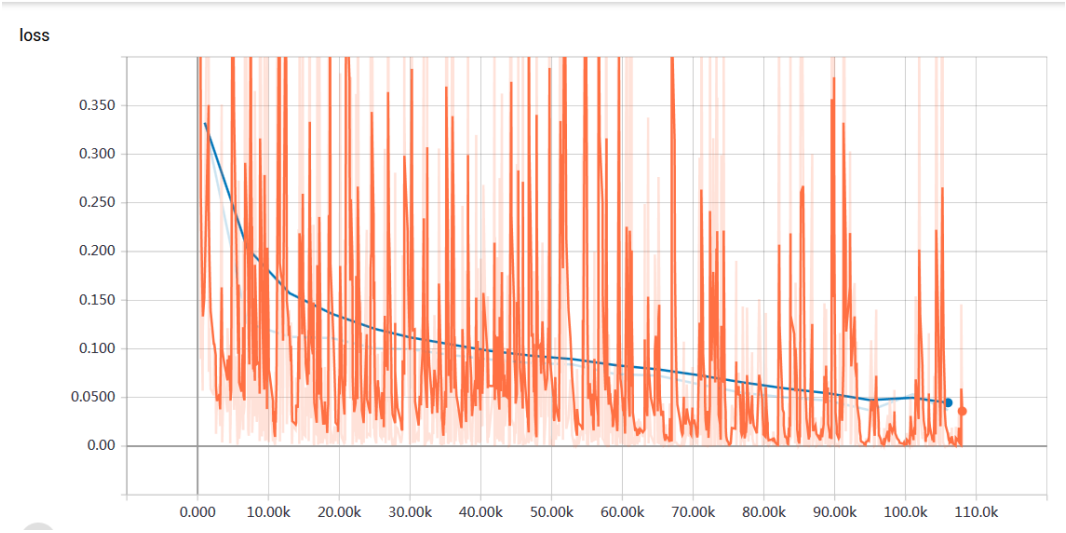

图3.17 训练loss曲线

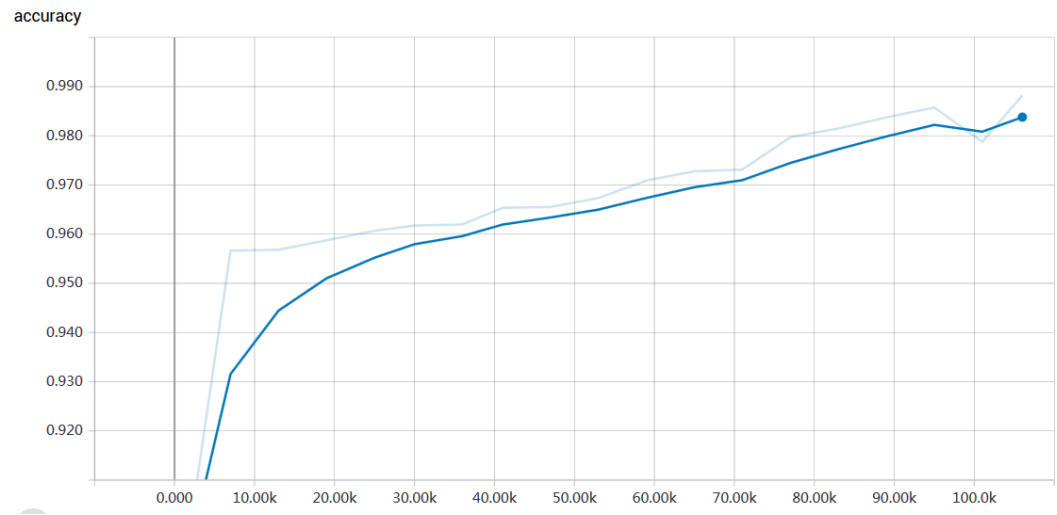

图3.18 验证集acc曲线

4 结论

本文利用了现代数字信号处理中的 STFT/Wiener 滤波/小波变换等方法，构建了从读取音频序列，到 STFT 生成时频二维谱图，到后续的利用维纳滤波和小波阈值变换，最后到 CNN 分类器的从海洋声音数据检测露脊鲸上曲叫声的流水线。文中分析了方案选取方法，尽可能地简化处理流程。

最后通过实验验证在低信噪比的情况下，上述过程提取的特征是稳健有效的。因而该方法也能够有效地检测低信噪比的上曲叫声，在划分的约 10000 条验证数据集上达到 98.3%的 acc。

参考文献

[1] 露脊鲸[EB/OL].https://baike.baidu.com/item/%E9%9C%B2%E8%84%8A%E9%B2%B8/1396279?fr=aladdin.

[2] Pourhomayoun M, Dugan P J, Popescu M,et al. Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network[J]. 2013.

[3] 小波阈值分析[EB/OL].https://blog.csdn.net/zhang0558/article/details/76019832.

[4] python aifc reference[EB/OL].https://docs.python.org/2/library/aifc.html.

[5] 直方图均衡 [EB/OL].http://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_equalize.html#sphx-glr-auto-examples-color-exposure-plot-equalize-py.