文章编号：2095-2163(2020)05-0018-06 中图分类号：TP399 文献标志码：A

基于 DTW 距离的动物声音分类研究

黄 玮[1]，冉启斌[2]
（1 南开大学 汉语言文化学院，天津 300071；2 南开大学 文学院，天津 300071）

摘 要：文章提出从声学距离的角度对动物分类的想法，并通过计算动物声音的 DTW 距离来对动物进行分类。实验发现，DTW 算法在动物分类的运用上具有可行性和科学性，分类结果在一定程度上反映动物的基本特征，每种动物用于实验的声音的数量越多，则分类的效果越理想。

关键词：动物分类；DTW 算法；声学距离

Animal Sound classification based on DTW distance

HUANG Wei[1], RAN Qibin[2]
（1 College of Chinese Language and Culture, Nankai University, Tianjin 300071, China;
2 School of Literature, Nankai University, Tianjin 300071, China）

【Abstract】 This paper puts forward the idea of animal classification from the perspective of acoustic distance, and classifies animals by calculating the DTW distance of animal sound. The experiment found that DTW algorithm is feasible and scientific in the application of animal classification, and the classification results reflect the basic characteristics of animals to a certain extent. The more sounds each animal uses in the experiment, the better the classification effect will be.

【Key words】 animal classification; DTW algorithm; Acoustic distance

0 引 言

现在所用的动物分类系统，大多是以动物形态或解剖的相似性和差异性的总和为基础的。根据古生物学、比较胚胎学、比较解剖学上的许多证据，基本上能反映动物界的自然类缘关系[1]。在分类特征的依据方面，迄今形态学特征尤其是外部形态仍然是最直观和常用的依据。

从动物声音的角度出发，对动物进行分类，有别于传统的动物分类体系，可以丰富动物分类的依据，让动物分类体系更加立体化，帮助人类从听觉角度建立对动物世界的再认识，形成全新的认知系统。从动物声学距离的角度出发对动物进行分类，能够在一定程度上反映出动物声音的声学特征，进而揭示动物声音的发展规律、演变顺序等信息，有助于对动物演化过程的研究。

在动物声音的研究方面，早在 1995 年 Kurt 等就设计了一款基于特征提取算法的程序，对海洋哺乳动物的声音进行识别和归类[2]；Yuanfeng Ma 等(2008)使用短时傅里叶变换(STFT)、摩尔模型等方法，从时频感知的角度对海洋哺乳动物的声音进行分类[3]；Che Yong Yeo 等(2011)提出了基于动物声音模式识别的动物识别和检测系统，该系统使用零交叉率(ZCR)、梅尔频率倒谱系数(MFCC)和动态时间规整(DTW)联合算法，并用狗的声音做出了检验[4]；Fernando 等(2017)使用平行识别模型和倍率分析对海洋哺乳动物的声音进行了探测和分类[5]，该研究考虑到了每个物种发出的多种声音，但研究只涉及墨西哥湾的 11 种海洋哺乳动物；Tuomas Oikarinen 等(2018)引入了端到端前馈卷积神经网络对圈养狨猴的呼叫声的来源和类型进行了分类[6]；Na Lin 等(2018)提出了一种对动物声音信号进行分类的新方法，即基于稀疏表示法的时频域方法，可以对重叠的动物声音进行分类[7]。综合来看，前人的研究着重于两个方面，一是从声音角度对动物进行识别和归类，主要运用于海洋哺乳动物；二是对动物声音的类型进行探测和归类，主要使用狗、狨猴等较为单一的物种进行检验。

DTW 算法已经被广泛地应用到基于识别、距离计算、数据匹配的各个领域，最具代表性的是应用在人类语音识别领域。吕军等(2007)较早使用 DTW 算法对汉语学习者的发音进行识别并进行评价系统设计[8]；邹韬(2012)使用 DTW 算法对汉语扬州方

基金项目：国家社科基金重大项目(19ZDA300)。

作者简介：黄 玮(1997-)，男，本科生，主要研究方向：语音学；冉启斌(1977-)，男，博士，教授，博士生导师，主要研究方向：语音学、词汇学。
通讯作者：冉启斌 Email：ranqibin@126.com

收稿日期：2020-01-20

言的识别进行了研究和设计[9]；王国林(2017)使用 DTW 算法设计评价系统对我国中学生的英语发音进行自动评价[10]；Hossein Hamooni 等(2016)通过基于 DTW 的分类来对音素序列进行识别，进而实现对话语的识别[11]。由此可见，DTW 算法运用于声音的研究皆有先例可循。

本文将基于 DTW 算法，提出对动物物种从声学角度进行分类，而非对某类动物进行识别和归类，或者对动物的某种声音进行识别。

1　实验情况

1.1　研究对象

本研究对 175 种动物的声音进行分类，每种动物拟使用 3 条声音，即研究对象为 525 条动物声音。在补充实验中，又对其中 43 种动物的声音进行了分类，每种动物的声音增加到 10 条，总计 430 条声音。

1.2　声音情况

1.2.1　声音参数

本研究中使用的声音均下载自 www.animal-sounds.org 等 8 个国外声音网站，下载的声音文件格式有.mp3、.aiff、.wav 等，采样率有 11 025 Hz、22 050 Hz 等，存储位数有 8 位、16 位、32 位等，有单声道声音和双声道声音。最终，本文使用 Praat(Praat：doing phonetics by computer，简称 Praat)将声音文件统一为.wav 格式，将声音采样率统一为 22 050 Hz，将存储位数调整为 16 位，将声道设置为单声道，进行保存和实验。

1.2.2　声音处理

本实验中用以剪切声音的软件是 Praat，一款实现跨平台多功能语音学实验的专业软件，主要用于对数字化的声音信号进行分析、标注、处理及合成等，同时输出各类语图和文字报表。在实验中，用 Praat 分别将动物的声音打开，将满足研究需要的声音剪切出来并保存。

在剪切中遵循以下标准：如果有较为明确的周期，则按一个周期剪切出声音，例如布谷鸟的叫声是“布谷布谷”，则剪切出“布谷”；没有明确周期的，或者周期极短、声音急促而连续的，则按 1 秒的声音长度剪切。

1.3　声学距离计算

1.3.1　动态时间规整算法

日本学者 Itakura 提出的动态时间规整(Dynamic Time Warping，DTW)算法是把时间规整和距离测度计算结合起来的一种非线性规整技术，采用动态规划(Dynamic Programming，DP)思想将一个复杂的全局最优化问题转化为许多局部最优化问题，一步一步进行决策，寻找出一条最佳路径。DTW 算法早期广泛应用于语音识别领域，尤其适用于对孤立点的匹配和识别，具有计算速度较快，结论直观等优点。现在，DTW 算法被广泛应用于语音检索[13]、汉语声调识别[14]、汉语方言语音识别[7]、手写签名识别[15]、手势识别[16]、图形识别[17]、空中目标识别[18]、农作物遥感影像识别及归类[19]、电波识别[20]等领域。

在本实验中，DTW 算法作为核心工具，主要用于计算各条声音两两之间的声学距离(即 DTW 距离)，这些声学距离将用于系统聚类分析。

1.3.2　系统聚类分析

本实验使用的系统聚类分析和主成分分析工具是 SPSS，它是一款著名的数据统计与分析软件，全称为 Statistical Product and Service Solutions(统计产品与服务解决方案软件)，最初软件全称为 Statistical Package for the Social Sciences(社会科学统计软件包)，它涵盖了数据管理、统计分析、图表分析、输出管理等功能，其中统计分析又包含系统聚类分析和主成分分析等功能。本实验使用 SPSS 将上述的声学距离进行系统聚类分析，得出以声学距离为基础的谱系图，该谱系图作为以声学距离为基础进行动物分类的直观呈现。

1.3.3　其他实验工具及脚本

由于 DTW 算法属于一种计算思维，没有具体的操作工具，本实验使用了承载 DTW 算法的脚本来实现对声学距离的计算。在两两计算动物声音的声学距离之前，需要将录音名称修改为该动物的名称，由于逐个修改工作量大，容易出错，本实验使用了重命名脚本来实现，该脚本可以在几秒钟之内将文件夹中的成百条录音以上级文件夹的名称批量重命名，并将重命名之后的文件汇集到同一个文件夹之中。

使用距离计算脚本计算动物间声学距离后，距离文件以文本文档格式保存，为了使其适用于 SPSS 的运行方式，本实验使用了作者自己编写的制表工具 sound2xls-full 将文本文档转存为 Excel 表格。该制表工具是基于 Python 设计的应用软件，操作简单、实用高效，极大地简化了数据整理工作，保证了数据另存过程的准确性。

1.4　补充实验

由于客观条件的限制，要保证有 175 种动物，而每种动物的有效声音只有 3 条，声音数量较少，对实

验结果有一定的影响,为了验证本方法的科学性和可行性,做了一项补充实验。补充实验中每种动物采用 10 条有效声音。经过筛选,175 个物种中有 43 种能够提取出 10 条有效声音。

2 实验结果

2.1 主实验结果

通过计算和数据处理,得到 175 种动物的聚类分析谱系图,如图 1 所示。

图 1 是 SPSS 生成的谱系图,依据动物之间距离的远近进行分类。横坐标表示距离,该距离是将计算出的 DTW 距离进行标准化转换以后得到的,区间是 0~25(左开右闭区间,下同)。当从横坐标上取一个点,从该点做一条垂直于横轴的直线,该线与连接着动物的水平线的交点数目,就是动物被划分出的类别数量。例如:当距离取 24 时,有 2 个交点,则在这个距离上,动物被分为两类,pewee 至 chaffinch(指纵轴上由 pewee 向下至 chaffinch 范围内的所有动物)为一类,dog 至 lion 为一类;当距离取 20 时,有 3 个交点,则在这个距离上,动物被划分为三类,pewee 至 macaw 为一类,hamster 至 chaffinch 为一类,dog 至 lion 为一类;当距离取 13 时,有 5 个交点,则在这个距离上,动物被划分为五类,pewwee 至 hyena 为一类,alligator 至 macaw 为一类,hamster 至 blackbird 为一类,chaffinch 单独为一类,dog 至 lion 为一类。另外,单从距离与动物类别的数量来看,当距离为 0~1 时,动物被划分为 175 类;距离为 1~2 时,划分为 102 类;距离为 2~3 时,划分为 47 类;距离为 3~4 时划,分为 25 类;距离为 4~5 时,划分为 19 类;距离为 5~6 时,划分为 13 类;距离为6~8 时,划分为 9 类;距离为 8~9 时,划分为 8 类;距离为 9~10 时,划分为 7 类;距离为 10~11 时,划分为 6 类;距离为 11~12 时划分为 14 类;距离为 14~23 时划分为 3 类;距离为 23~25 时划分为 2 类。

将图 1 所示的分类结果与传统动物分类方法进行比较,发现有许多相吻合之处。例如,blue jay 和 crow 在距离大于 1 时,被划分为一类,它们在传统的动物分类上都是雀形目鸦科动物;hamster 和 mouse 在距离大于 1 时,被划分为一类,它们都是啮齿目鼠形亚目动物;bobcat 和 cheetah 在距离大于 1 时,被划分为一类,它们都是食肉目猫科动物;tiger、cougar 和 lion 在距离大于 2 时,被划分为一类,它们都是食肉目猫科动物;dog 和 coyote 在距离大于 3 时,被划分为一类,它们都是食肉目犬科犬属动物。

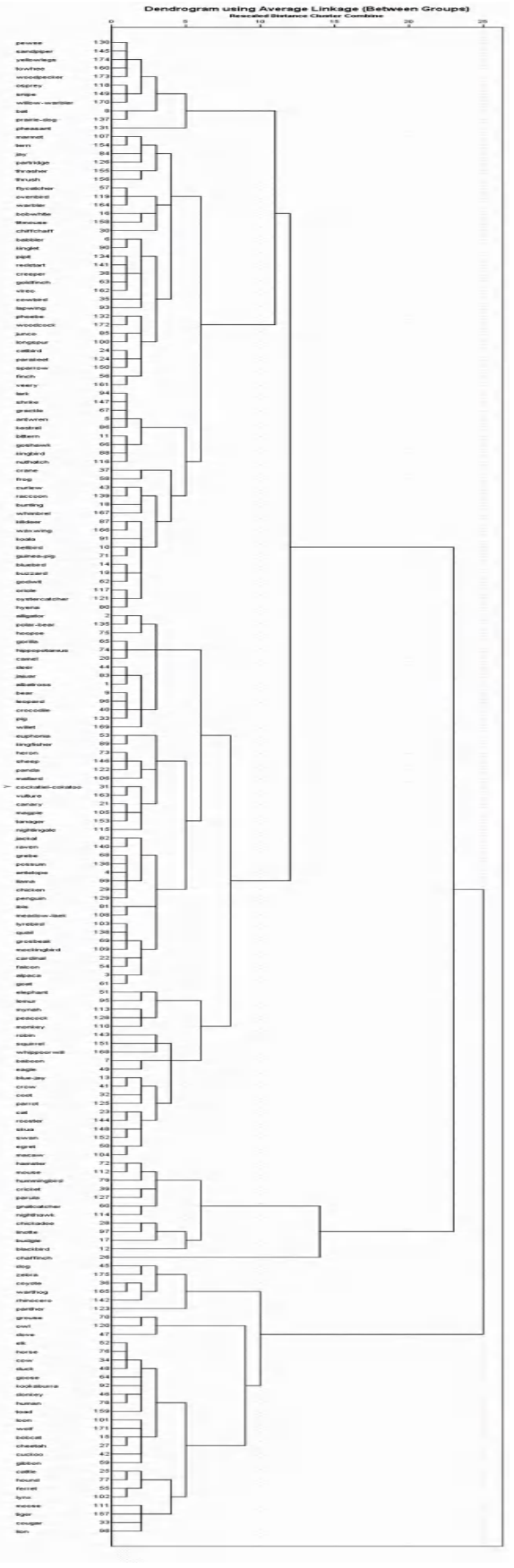

图 1 175 种动物的分类结果

Fig. 1 Classification results of 175 species of animals

同时，更多的是与传统的动物分类方法相异的地方。例如：polar bear（属于哺乳纲）和 alligator（属于爬行纲）在距离大于 1 时，被划分为一类；hippopotamus（属于鲸偶蹄目）和 camel（属于偶蹄目）在距离大于 1 时，被划分在一类；elephant（属于长鼻目）和 lemur（属于灵长目）在距离大于 1 时，被划分在一类。

图 2 显示的是类别数量与距离之间的对应关系，以及相应的变化趋势。可见，距离区间与类别数量成负相关关系，距离在 0～6 区间时，类别数量的变化率较大，在 6～25 区间时，变化平缓。在较小的距离内，类别数量产生了较大的变化，说明在动物声音之间的细微差异还是比较小。

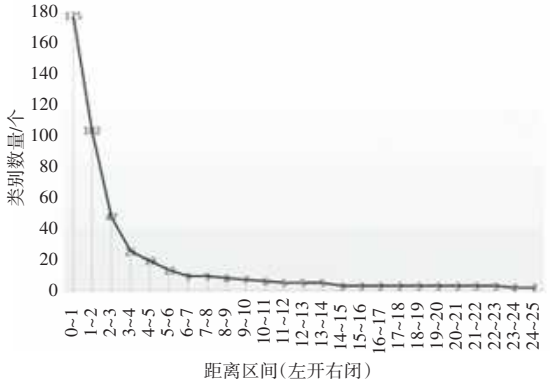

图 2　175 种动物的类别数量与距离区间的关系

Fig. 2　The relationship between the number of categories and the distance interval of 175 species of animals

2.2　补充实验结果

通过重复主实验过程，对动物声音样本进行计算和数据处理之后，得到补充实验中 43 种动物的聚类分析谱系图。

如图 3 所示，当距离为 0～1 时，动物被划分为 43 类；距离为 1～2 时，划分为 39 类；距离为 2～3 时，划分为 34 类；距离为 3～4 时，划分为 25 类；距离为 4～5 时，划分为 20 类；距离为 5～6 时，划分为 17 类；距离为 6～7 时，划分为 14 类；距离为 7～8 时，划分为 10 类；距离为 8～9 时，划分为 8 类；距离为 9～10 时，划分为 7 类；距离为 10～12 时，划分为 6 类，距离为 12～13 时，划分为 5 类；距离为 13～17 时，划分为 3 类；距离为 17～25 时，划分为 2 类。

从传统动物分类的角度对图 3 的结果进行了分析，本次补充实验验证了本研究所用方法的科学性。首先，有证据显示，在传统分类中属于同一科的动物，在研究中随着声音数目的增加，距离更近。例如：jay 和 blue jay 在距离大于 5（主实验为 11）时被划分为同一类，它们都是雀形目鸦科动物；dog 和 wolf 在距离大于 4（主实验为 9）时被划分为同一类，它们都是食肉目犬科犬属动物。其次，虽然有的动物之间的距离有所拉大，但是还在可以接受的范围之内。例如：leopard 和 jaguar 在距离大于 3（主实验为 2）时被划分为同一类，它们都是食肉目猫科豹属动物；tiger 和 lion 在距离大于 3（主实验为 2）时被划分为同一类，它们都是食肉目猫科豹属动物；goat 和 antelope 在距离大于 6（主实验为 3）时被划分为同一类，它们都是偶蹄目牛科动物。当然，这是以传统动物分类体系为参照做出的比较，因为是从声音角度对动物进行分类，与传统分类方法截然不同，但是目前尚无别的办法。

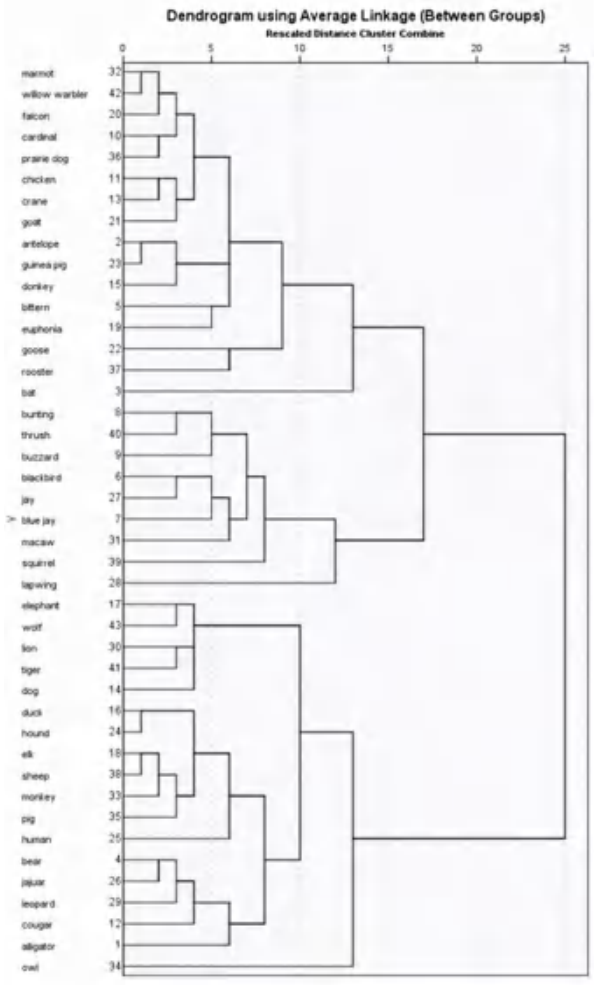

图 3　43 种动物的分类结果

Fig. 3　Classification results of 175 species of animals

图 4 显示的是补充实验中动物类别数量与距离区间之间的关系，与主实验相比，本图显得较为平缓。在距离较小的区间内，类别数量没有出现断崖式的下跌，也反映出声音内部的特征距离比较大，这也可能是由于物种数量与主实验相比较少造成的。

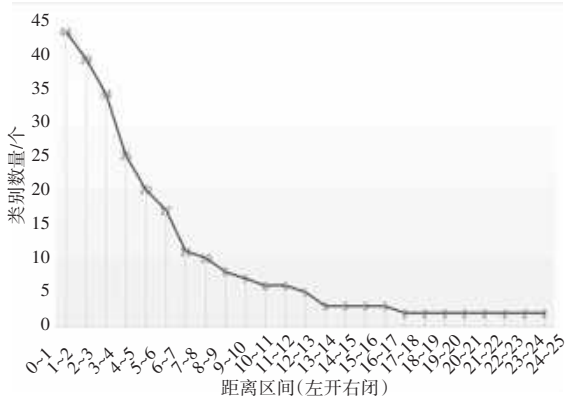

图 4 43 种动物的类别数量与距离区间的关系

Fig. 4 The relationship between the number of categories and the distance interval of 175 species of animals

统计了鸟类在两次实验中的区分率情况,发现补充实验能够更好地将鸟类与其他动物分开。在主实验中,鸟类占总数的 66.28%,当只划分为两个大类时(即距离大于 23),鸟类在第一类中占比为 75.52%,在第二类中占比为 25%;在补充实验中,鸟类占总数的 46.51%,只划分为两个大类时(即距离大于 17),鸟类在第一类中占比为 72%,在第二类中占比 11.11%。综合来看,本研究能够较好地将鸟类和其他动物区别开来,尤其是在补充实验中,第二类里鸟类的占比已经非常少了。

3 讨 论

本研究使用 DTW 算法计算出动物声音之间的声学距离,通过数据分析对动物进行聚类分析,旨在探索一种新的动物分类维度和方法。除了主实验外,还做了补充实验论证本方法的科学性。

3.1 距离与类别数量

在主实验中,当距离区间在较小范围内(0~6)时,类别数量的变化较为剧烈,即当距离尺度稍微放大,类别数量就会大量减少,这说明在动物声音之间的差别较小,对距离尺度的变化做出的反应比较敏感。在补充实验中,距离区间的变化与类别数量的变化较为平缓,在较小距离区间也没有出现类别数量急剧变化的情况,这说明动物声音之间的差别比较大,对距离尺度的变化做出的反应比较迟钝。本文认为,在主实验中,每种动物只有 3 条声音,不能较好地反映该种动物的声音所具有的区别于其他动物的特征,所以算法没有能很好地捕捉到动物声音体现出的特征,进而表现出类别数量与距离之间较为敏感的对应;在补充实验中,每种动物有 10 条声音,声音数量的增加,更好地反映出每种动物的声音特征,使动物之间的区别更加明显,距离更大,因此在类别数量与距离区间的对应上,显得不那么敏感;另外,本文认为与动物种类的数量有关,主实验中有 175 种动物,补充实验只有 43 种动物,因而在主实验中,由于基数较大,当在较小的距离区间内时,类别数量产生了大幅的变化;在补充实验中,动物数量较少,类别数量的变化范围就会较小。但是本质上还是和动物声音特征的区别度有关。

3.2 鸟类声音的区别性

声音的物理特征包括音高、音强、音长和音质。其中音质是声音的基本属性,由发音体、发音方法和共鸣器决定。由于哺乳动物和鸟类在共鸣器上存在较大的差别,所以有理由相信在很大程度上,鸟类声音会与其他动物的声音有较大差别。实验结果验证了这一猜想,在主实验和补充实验中,鸟类声音大体上都能与其他动物的声音区别开来,并且在补充实验中这一现象更加明显。另外,实验结果中对鸟类的划分与传统动物分类体系的划分相差较大,有很多鸟类不是同一科,甚至不是同一目,会在很小的距离内被划分到一起。相比之下,属于同一目或同一科的哺乳动物,尤其是猫科和犬科动物,在音质上统一性更好,所以被划分到一起的几率更大。这也说明,基于形态学方法对动物的划分,存在声音维度上的欠缺。

3.3 人耳感知和声音本质

本研究使用的 DTW 算法是将动物声音的频率(单位:Hz)转化为梅尔刻度(Mel scale,单位:Mel)计算的。Mel 与 Hz 是心理-声学相关的等价单位,它体现的是人耳对声音的感知,这种感知与声音的客观频率 Hz 是非线性对应关系[21]。在研究中使用梅尔刻度,是立足于从人类听觉感知的角度对动物进行分类。

另外,王士元(1998)曾提出人类学、遗传学和语言学是一种综合体,考古、遗传和语言是了解人类过去历史的 3 个窗口[22],语言的演化能反映人类的发展。本文认为,从声学意义上,动物的声音蕴含着动物的特征,动物声音的演化也能反映动物的演化,动物之间声音的关系在一定程度上也能揭示动物之间的关系。在以后的研究中,应该探究距离数据所代表的声音特征,探索其中的联系和规律。

3.4 不足与改进方向

由于传统的动物分类体系几乎不考虑动物的声音,目前没有与本研究类似的从动物声音角度对动物进行分类的研究结果可供对比,所以在接下来的研究中应该弥补不足,进一步验证研究方法的科学

性，深入挖掘研究的意义。

由于客观条件的限制，主实验中涉及 175 种动物，每种动物只有 3 条声音，样本数量较少。通过补充实验发现，通过增加声音样本的数量，会使分类结果更科学。但是，补充实验中只涉及了 43 种动物，动物的物种数量过少。在接下来的研究中，应该在增加物种数量的同时，增加每种动物的声音数量。另外，对于没有声音和声音较小的物种，比如鱼类和小型昆虫，没有办法进行分类。

研究中使用的声音材料下载自不同的网站，声音质量不统一，可能会对研究结果造成影响。目前缺少高质量、广博齐全的动物声音数据库，所以这个问题还没有办法很好的解决，只能在筛选声音的时候更加仔细。

附录一 动物声音网站

http://www.animal-sounds.org.

https://www.seaworld.org/animals/sounds.

http://www.findsounds.com/animals.html.

http://www.animalsoundarchive.org.

https://www.naturebits.org.

https://www.freesound.org.

http://www.grsites.com/archive/sounds.

https://www.freesoundeffects.com.

参考文献

[1] 刘凌云，郑光美. 普通动物学(第三版)[M]. 北京：高等教育出版社：2009.1-5

[2] Kurt M. Fristrup, William A. Watkins. Marine animal sound classification [J]. The Journal of the Acoustical Society of America, 1995, 97(5).

[3] MA Y, CHEN K. A time-frequency perceptual feature for classification of marine mammal sounds [C]//2008 9th International Conference on Signal Processing. IEEE, 2008: 2820-2823.

[4] YEO C Y, AL-HADDAD S A R, NG C K. Animal voice recognition for identification (ID) detection system [C]//2011 IEEE 7th International Colloquium on Signal Processing and its Applications. IEEE, 2011: 198-201.

[5] FERNANDO RUBÉN GONZÁLEZ-HERNÁNDEZ, LUIS PASTOR SÁNCHEZ-FERNÁNDEZ, SERGIO SUÁREZ-GUERRA, et al. Marine mammal sound classification based on a parallel recognition model and octave analysis [J]. Applied Acoustics, 2017, 119.

[6] OIKARINEN T, SRINIVASAN K, MEISNER O, et al. Deep convolutional network for animal sound classification and source attribution using dual audio recordings [J]. The Journal of the Acoustical Society of America, 2019, 145(2): 654-662.

[7] LIN N, SUN H, ZHANG X P. Overlapping animal sound classification using sparse representation [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 2156-2160.

[8] 吕军，曹效英. 基于语音识别的汉语发音自动评分系统的设计与实现[J]. 计算机工程与设计，2007(5)：1232-1235.

[9] 邹韬. 基于孤立字词的扬州方言语音识别研究[D]. 扬州大学，2012.

[10] 王国林. 基于 DTW 的语音评价系统及其中学英语教学中的应用[D]. 华中师范大学，2017.

[11] Hossein Hamooni, Abdullah Mueen, Amy Neel. Phoneme sequence recognition via DTW-based classification [J]. Knowledge and Information Systems, 2016, 48(2).

[12] ITAKURA F. Minimum prediction residual principle applied to speech recognition[J]. IEEE Trans.Acoust.Speech Signal Process. 1975, 23.

[13] 张利平. 汉语连续语音识别系统的研究与实现[D]. 西北大学，2010.

[14] 万春. 基于 DTW 的语音识别应用系统研究与实现[J]. 集美大学学报(自然科学版)，2002(2)：104-108.

[15] 全中华. 基于动态手写签名的身份认证研究[D]. 中国科学技术大学，2007.

[16] 荆雷，马文君，常丹华. 基于动态时间规整的手势加速度信号识别[J]. 传感技术学报，2012，25(1)：72-76.

[17] 张湘莉兰. 若干图像和语音数据分类问题研究[D]. 国防科学技术大学，2013.

[18] 姚佩阳，周旺旺，张杰勇，等. 基于动态时间规整的空中目标机动识别[J]. 火力与指挥控制，2018，43(9)：15-18+24.

[19] 翟涌光，屈忠义. 基于非线性降维时序遥感影像的作物分类[J]. 农业工程学报，2018，34(19)：177-183.

[20] 王振浩，杜凌艳，李国庆，等. 动态时间规整算法诊断高压断路器故障[J]. 高电压技术，2006(10)：36-38.

[21] 李爱军. 语调研究中心理和声学等价单位[J]. 声学技术，2005(24)3.

[22] 王士元. 王士元语言学论文集[M]. 北京：商务印书馆，2002.41-68.

(上接第 17 页)

[13] KIAPOUR M H, HAN X, LAZEBNIK S, et al. Where to Buy It: Matching Street Clothing Photos in Online Shops [J]. IEEE Transactions on Multimedia, 2014, 16(1): 253-265.

[14] 覃蕊，梁惠娥，陈东生，等. 服装风格评价体系探讨[J]. 山东纺织科技，2007，48(5)：33-35.

[15] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[16] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// International Conference on International Conference on Machine Learning. JMLR.org, 2015.

[17] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]//Thirty-first AAAI conference on artificial intelligence. 2017.

[18] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: Generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[19] LIU Z, LUO P, QIU S, et al. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.