Bird Whisperer: Leveraging Large Pre-trained Acoustic Model for Bird Call Classification

Muhammad Umer Sheikh, Hassan Abid*, Bhuiyan Sanjid Shafique*, Asif Hanif, Muhammad Haris*

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

{muhammad.sheikh,hassan.abid,bhuiyan.shafique,asif.hanif,muhammad.haris}@mbzuai.ac.ae

Abstract

Adapting large pre-trained acoustic models across diverse domains poses a significant challenge in speech processing, particularly when shifting from human to non-human contexts. This study aims to bridge this gap by utilizing the pre-trained Whisper model, initially intended for human speech recognition, for classifying bird calls. Our study reveals that when employed solely as a feature extractor, the Whisper encoder fails to yield meaningful features from bird calls, possibly due to categorizing them as background noise. We propose a simple but effective technique to enhance Whisper's ability to extract distinctive features from avian vocalizations, resulting in a remarkable 15% increase in F1-score over the baseline. Furthermore, we mitigate the issue of class imbalance within the dataset by introducing a series of data augmentations. Our findings underscore the potential of adapting large pre-trained acoustic models to tackle broader bioacoustic classification tasks. The code is available at https://github.com/umer-sheikh/bird-whisperer

Index Terms: bird call classification, birdclef-2023, fine-tuning, whisper

1. Introduction

Birds, as integral components of our ecosystem, play multifaceted roles beyond pollination, seed dispersal, and habitat maintenance [1]. Their activities influence nutrient cycling, pest control, and even disease regulation, thus exerting far-reaching impacts on ecosystem dynamics. Moreover, their sensitivity to environmental changes makes them invaluable bioindicators, reflecting alterations in habitat quality, climate patterns, and overall ecosystem health [2]. Effective monitoring of bird species and populations not only aids in gauging the success of ecological restoration efforts but also provides early warnings of broader ecological imbalances [3]. Furthermore, the study of avian behavior and distribution patterns can inform land management strategies, conservation planning, and even urban development practices, fostering more sustainable coexistence between humans and wildlife. Therefore, recognizing the intricate connections between avian biodiversity and ecosystem resilience underscores the importance of bird monitoring programs in preserving global biodiversity and ecological stability [4].

However, traditional bird observation survey methods are labor-intensive, time-consuming, and logistically challenging [5]. In response to the challenges posed by traditional bird observation methods, technological innovations have emerged to revolutionize avian monitoring efforts. Acoustic sensors deployed in natural environments have emerged as a promising solution, capturing rich auditory data that offer insights into avian presence, behavior, and biodiversity [6]. However, the sheer volume of data generated by these sensors presents a formidable obstacle to manual analysis, underscoring the need for automated analysis of bird sounds [7]. To unlock the full potential of acoustic monitoring, there is a pressing need for automated analysis tools capable of efficiently processing and interpreting large-scale bird sound datasets. Machine learning algorithms, trained on annotated audio recordings, hold promise in automating species identification and quantifying avian activity patterns.

Advancements in deep learning have paved the way for utilizing these technologies in automatically extracting features from complex datasets, including for the acoustic classification of bird species [8]. Despite these advancements, challenges persist, particularly the lack of large, labeled datasets and the presence of class imbalance across bird species [9].

One widely employed strategy to address the scarcity of annotated training data is leveraging pre-trained models through linear probing and fine-tuning for the downstream task. In the domain of acoustic classification for different species, numerous studies have explored the application of linear probing, predominantly relying on various adaptations of pre-trained Convolutional Neural Networks (CNNs) for feature extraction [10, 11]. These pre-trained models have primarily been trained on natural images. However, we advocate for an alternative approach: utilizing a model pre-trained on human speech data. This choice seems natural considering that bird calls also fall within the category of audio data.

In recent years, the domain of human speech processing has witnessed remarkable progress with the development of unsupervised pre-training techniques, exemplified by Wav2Vec 2.0 [13]. These methods are able to learn directly from raw audio in the absence of labels. Given the extensive availability of unlabeled human speech data, audio encoders are able to learn high-quality representations of speech. A question then arises: ***How effectively can encoders pre-trained on human speech be adapted for extracting informative features from non-human audio data, such as bird calls?***

In this study, we explore the adaptation of a pre-trained audio encoder, originally trained on human speech, for the task of extracting distinctive features for fine-grained classification of bird calls. Specifically, we utilize the encoder from OpenAI's Whisper model, which has undergone training on an expansive multilingual dataset, comprising 680,000 hours of labeled audio [12]. The inherent diversity present in such a large dataset allows Whisper's encoder to be incredibly robust and versatile. Additionally, we conduct data augmentations on the original dataset to mitigate the issue of class imbalance. The combi-

*Equal contribution

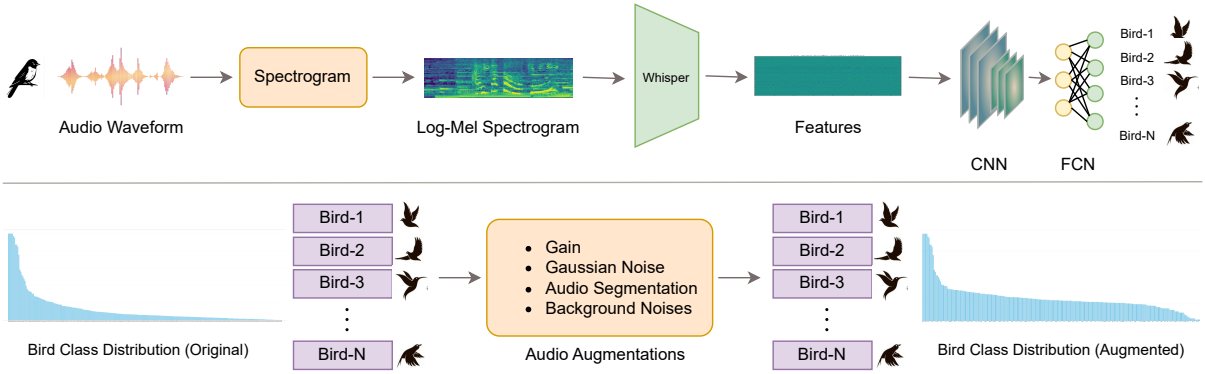

Figure 1: ***Methodology Overview****: The top row illustrates the pipeline for classifying bird calls. Initially, the spectrogram of an audio waveform (a recording of a bird call) is processed by the Whisper [12] encoder to extract temporal features. Subsequently, these features are further processed by a small convolutional neural network (CNN) and then by a fully connected network (FCN), which collectively generate predictions for N bird species. The bottom row illustrates the data augmentations applied to audio recordings of bird calls. The heavy-tailed distribution (class imbalance issue) in the original dataset has been mitigated in the augmented dataset.*

nation of data augmentations and fine-tuning of the Whisper encoder results in a significant performance improvement, elevating the F1-score from 39.4% to 85.8%.

2. Problem Description

This research examines the feasibility of adapting Whisper's encoder, trained on extensive human speech datasets, for non-human sound classification, particularly avian vocalizations. Despite its proficiency in human speech feature extraction, challenges arise in data distribution mismatch and feature relevance between human and non-human speech. Moreover, non-human sounds often contain higher noise levels, complicating classification. Yet, the encoder's adaptability suggests potential in non-human speech classification, aligning with domain adaptation principles to extend pre-trained models to new tasks. This exploration contributes to transfer-learning/fine-tuning and model adaptability in diverse domains. In essence, this work focuses on domain adaptation, assessing how a deep learning model trained on human speech performs in classifying non-human speech, in particular avian vocalizations.

3. Method

The method section is structured as follows: Firstly, the dataset containing recordings of bird calls is described, followed by an overview of the data augmentation techniques employed to mitigate class imbalance. Subsequently, we provide a brief description of the backbone used for feature extraction, specifically the encoder of the Whisper model [12], and outline various training strategies utilized to adapt the Whisper encoder for bird call classification.

3.1. Dataset

This study utilizes the BirdCLEF 2023 dataset, provided by the Cornell Lab of Ornithology [14]. The dataset comprises audio recordings from 264 distinct bird species, with the number of audio samples per species ranging from 1 to 500. This significant variation in sample size across classes introduces a notable imbalance within the dataset. The distribution of classes in original dataset is shown in Figure 2(a), highlighting a heavy-tailed distribution and the challenges posed by this imbalance.

3.2. Data Pre-processing and Augmentation

Given the dataset's inherent class imbalance, we implemented a comprehensive pre-processing and augmentation strategy to enhance the representation of underrepresented classes. Our approach involves applying four distinct audio augmentations directly to the raw audio data and two additional augmentations to the Mel-spectrogram representations. The Mel-spectrogram augmentations are applied randomly with a 50% probability during the training phase. The audio augmentations have been strategically employed for classes with fewer than 100 samples to ensure generalization and mitigate the dataset's imbalance. We have tested different combinations of the augmentations to examine their effects. These audio augmentations included:

1. **Audio Segmentation:** To increase the dataset size and ensure compatibility with the Whisper encoder's 30-second window, audio files exceeding 30 seconds are segmented into 30-second chunks. This effectively expands the dataset while preserving all the information.
2. **Gaussian Noise:** Random Gaussian noise was added to the audio signal to introduce variations and enhance the model's robustness to noise. This augmentation aims to account for noise in the audio recordings.
3. **Gain:** Random variations were applied to the amplitude (volume) of the audio signal to further augment the dataset, allowing the amplification of bird calls with very low sound levels.
4. **Background Noise:** Environmental noise is added to the background of the birds' call as the dataset is real world and has background noise such as rain, forest and others.

The impact of audio augmentations on the original dataset is demonstrated in Figure 2(b), showcasing an increased number of samples for previously underrepresented classes. On the Mel-spectrogram, frequency and time masking techniques have been applied.

For training and evaluation, we divide the dataset into training and testing subsets using an 80/20 split ratio. To maintain representation from all classes, we employ a stratified sampling technique and fix the seed to ensure consistency across multiple

experiments.

3.3. Feature Extraction and Model Training

OpenAI's Whisper [12] model is a state-of-the-art speech recognition model that leverages a transformer architecture to achieve remarkable performance in transcribing human speech. The model is composed of two main components: an encoder and a decoder. The encoder is a crucial component of the Whisper speech-to-text model, responsible for extracting meaningful representations from raw audio input. It employs a combination of convolutional layers, self-attention mechanisms, and residual connections to effectively transform audio signals into high-level contextual features.

We follow the same pre-processing steps as used by Whisper [12] model training scheme. The steps are as follows: Initially, the input audio signal undergoes resampling to a standardized sample rate of 16 kHz. Subsequently, audio files are either padded or trimmed to a consistent length of 30 seconds, ensuring uniform duration for all input sequences. To capture both temporal and spectral features of the audio, an 80-channel log-magnitude Mel-spectrogram is computed using 25-millisecond windows with a 10-millisecond stride, resulting in overlapping windows. Finally, the spectrogram is normalized to achieve a mean of zero and a standard deviation of one, thus ensuring numerical stability during subsequent processing stages.

Upon extracting feature vectors with Whisper's encoder, a small sized CNN following a fully connected network (FCN) is employed to convert Whisper encoder's temporal features to classification scores. An overview of the model pipeline can be found in Figure 1. We observe that when employed solely as a *fixed* feature extractor (linear probing), the Whisper encoder fails to yield meaningful features from bird calls, possibly due to categorizing them as background noise as demonstrated by Figure 3(left). The pre-trained Whisper encoder features appear uniform and lack discernible patterns specific to bird calls, indicating a poor representation of the audio content. To address this, we fine-tune the Whisper model to better adapt to our bird calls dataset. This improvement is evident in the comparative analysis of feature vectors before and after fine-tuning, as shown in Figure 3.

We explore different configurations during training, encompassing `Full Training` with random initialization of weights, `Linear Probing` where the Whisper encoder layers are frozen, and `Fine-Tuning` which involves starting from pre-trained weights and subsequently updating them using the BirdCLEF dataset. Additionally, we assess the impact of augmenting the dataset on performance by conducting experiments using both the original and augmented datasets.

4. Experiments

4.1. Experimental Setup

In our experimental setup, we utilize single Nvidia Quadro RTX 6000 GPU with 24 GB memory. The software environment included PyTorch (version 1.11.0+cu113), along with libraries such as openai-whisper, numpy, pandas, matplotlib, and scikit-learn. We use `base` Whisper model [12] with its encoder containing around $20M$ parameters. We use a CNN (consisting of two convolutional layers, each containing 5×5 kernels) to further process temporal features provided by Whisper encoder. Features from CNN's output are flattened and fed to a three-layer FCN to generate predictions of N bird species.

We implement four types of augmentations on raw audio

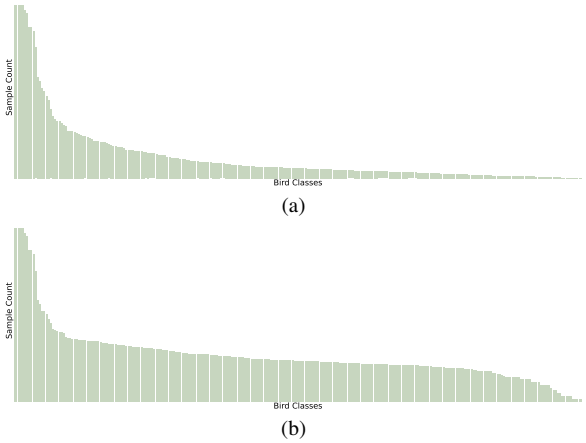

Figure 2: *Class Distribution of BirdCLEF-2023 Dataset: (a) original dataset with heavy-tailed distribution, (b) augmented dataset with more samples of previously underrepresented classes*

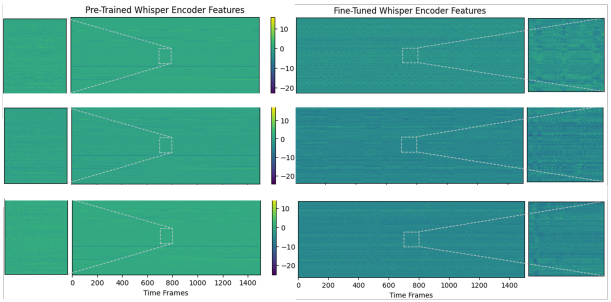

Figure 3: *Whisper features of three bird call recordings, with extreme left and right showing zoomed-in parts, (left) pre-trained, (right) after fine-tuning, revealing richer information content.*

data and two types of augmentations for Mel-spectrograms on original BirdCLEF-2023 dataset [14]. The augmentations for Mel-spectrograms are applied randomly only during training with a 50% chance. For raw audio augmentations, we segment files with fewer than 100 samples and then apply other audio augmentations to properly balance the classes. Figure 2 illustrates the class distribution of the dataset before and after augmentations. The augmentation variations are as follows:

- **Audio Waveform**
  1. **Gain**: Randomly amplify the audio amplitude from 13dB to 17dB.
  2. **Gaussian Noise**: Add Gaussian noise with zero mean and unit variance to the raw audio waveform.
  3. **Audio Segmentation**: Segment the audio into 30-second intervals to introduce new data, discarding the last segment if it is less than 10 seconds.
  4. **Background Noise**: Add different environmental noises such as breeze, forest sounds, and rain to the raw audio waveform.
- **Mel-spectrogram**
  1. **Time Masking**: Randomly chooses a starting point and width ranging from 300 to 500.
  2. **Frequency Masking**: Randomly chooses a starting point

and bands ranging from 12 to 15 Hz, however the zero frequency (DC component) is unmodified.

In our experimental setup, the model is trained using a cross-entropy loss function and optimized with stochastic gradient descent (SGD) with a learning rate of 0.01. Training is performed with batch size of 16 samples over 20 epochs, and the dataset is split into training and testing sets with an 80/20 ratio.

As Whisper is primarily a Transformer-based architecture [15], we opt to compare it with a CNN-based baseline model having roughly equal number of parameters. Following the work of [11], we select EfficientNet-B4 [16] as our baseline model due to its CNN-based architecture and having around $19.3M$ parameters. It must be noted that Ansar *et. al.* [11] uses an ensemble of two EfficientNet models (B3 and B4 variants). Given the computational expense of the ensemble approach, we opt to compare solely with the B4 variant. This choice is made because the number of parameters in the B4 variant is similar to that of the encoder in the Whisper-Base model.

4.2. Results & Discussion

We evaluate the model performance using macro-averaged F1-score under different training configurations. Experiments are conducted with two variants of the dataset: the original dataset and the augmented dataset. For each variant, experiments are performed in three training configurations: `Full Training` with random initialization of weights, `Linear Probing` where the Whisper encoder layers are frozen, and `Fine-Tuning` which involves starting from pre-trained weights and subsequently updating them. The results are presented in Table 1.

Comparing the performance on the original dataset (first three rows) to that on the augmented dataset, a significant boost in performance is observed. This indicates that incorporating data augmentation significantly enhances performance, demonstrating its effectiveness in presenting a diverse range of audio patterns for the model to learn from.

Among the three training configurations, the `Fine-Tuning` strategy demonstrates a favorable edge over the others. The high F1-scores achieved with the `Fine-Tuning` strategy underscores the advantages of adapting a pre-trained model, originally trained on human speech, to non-human speech data, such as bird vocalizations.

Table 1: *F1-score on original and augmented datasets under three training strategies for Whisper encoder.*

Config #	Dataset		Whisper Encoder			F1-Score
	Original	Augmented	Full Training	Linear Probing	Fine Tuning	
1	✓	✗	✓	✗	✗	0.2247
2	✓	✗	✗	✓	✗	0.1429
3	✓	✗	✗	✗	✓	0.3944
5	✗	✓	✓	✗	✗	0.7705
4	✗	✓	✗	✓	✗	0.7159
6	✗	✓	✗	✗	✓	**0.8582**

Findings deduced from Table 1, underline the importance of data augmentation and fine-tuning in developing efficient audio classification models. They contribute valuable insights into audio-based species classification, essential for ecological studies and biodiversity monitoring. We have also shown evaluation on test dataset and Whisper-Base model after each epoch in Figure 4. Model converges faster on augmented dataset and `Fine-Tuning` strategy as compared to other settings.

We compare results of EfficientNet-B4 and Whisper-Base encoder on original and augmented dataset with `Fine-Tuning` strategy in Figure 5. As compared to EfficientNet, Whisper provides higher F1-score.

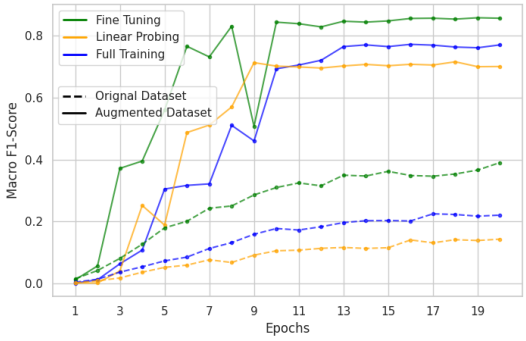

Figure 4: *F1-score over different epochs while training Whisper-Base encoder on BirdCLEF-2023 dataset.*

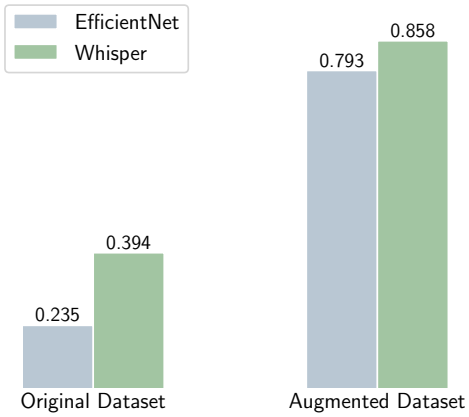

Figure 5: *Comparison of EfficientNet-B4 and Whisper-Base encoder performance (F1-score) on original and augmented dataset under fine-tuning strategy.*

5. Conclusions

This study underscores the adaptability of large pre-trained acoustic models, originally trained on human speech, for bird call classification—a significant advancement in the realms of domain adaptation and non-human speech recognition. Our findings affirm the effectiveness of harnessing pre-trained models focused on human speech to classify avian vocalizations, thus illuminating potential applications in broader audio classification tasks. Through the utilization of data augmentations and the exploitation of large pre-trained acoustic models, this research tackles challenges such as computational constraints and class imbalance in datasets. Moreover, it provides promising insights into the realm of linear probing and fine-tuning, laying the groundwork for future endeavors aimed at enhancing model adaptability across diverse acoustic domains.

6. References

[1] C. J. Whelan, D. G. Wenny, and R. J. Marquis, "Ecosystem services provided by birds," *Annals of the New York academy of sciences*, vol. 1134, no. 1, pp. 25–60, 2008.

[2] U.S. Geological Survey, "Birds as indicators of ecosystem health," https://www.usgs.gov/centers/forest-and-rangeland-ecosystem-science-center/science/birds-indicators-ecosystem-health, 2024, accessed: 2024-02-12.

[3] S. Chowfin and A. Leslie, "Using birds as biodindicators of forest restoration progress: A preliminary study," *Trees, forests and people*, vol. 3, p. 100048, 2021.

[4] M. A. Tabur and Y. Ayvaz, "Ecological importance of birds," in *Second International Symposium on Sustainable Development Conference*, 2010, pp. 560–565.

[5] R. D. Gregory, D. W. Gibbons, and P. F. Donald, "Bird census and survey techniques," *Bird ecology and conservation*, pp. 17–56, 2004.

[6] J. Zhang, K. Huang, M. Cottman-Fields, A. Truskinger, P. Roe, S. Duan, X. Dong, M. Towsey, and J. Wimmer, "Managing and analysing big audio data for environmental monitoring," in *2013 IEEE 16th International Conference on Computational Science and Engineering*. IEEE, 2013, pp. 997–1004.

[7] J. Xie and M. Zhu, "Acoustic classification of bird species using an early fusion of deep features," *Birds*, vol. 4, no. 1, pp. 138–147, 2023.

[8] J. Xiie and M. Zhu, "Handcrafted features and late fusion with deep learning for bird sound classification," *Ecological Informatics*, vol. 52, pp. 74–81, 2019.

[9] J. Xie, K. Hu, Y. Guo, Q. Zhu, and J. Yu, "On loss functions and cnns for improved bioacoustic signal classification," *Ecological Informatics*, vol. 64, p. 101331, 2021.

[10] M. Zhong, R. Taylor, N. Bates, D. Christey, H. Basnet, J. Flippin, S. Palkovitz, R. Dodhia, and J. L. Ferres, "Acoustic detection of regionally rare bird species through deep convolutional neural networks," *Ecological Informatics*, vol. 64, p. 101333, 2021.

[11] W. Ansar, A. Chatterjee, S. Goswami, and A. Chakrabarti, "An efficientnet-based ensemble for bird-call recognition with enhanced noise reduction," *SN Computer Science*, vol. 5, no. 2, p. 265, 2024.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[13] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 27 826–27 839.

[14] "Birdclef 2023 competition," https://www.kaggle.com/competitions/birdclef-2023, accessed on 2023-11-21.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.