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

INTERSPEECH 2024

KEY CONTRIBUTION

Adaptation of the Whisper model, originally designed for human speech recognition, to bird call classification, demonstrating a 15% F1-score improvement over the baseline. Introduction of targeted data augmentations to address class imbalance, showcasing the model's adaptability in non-human acoustic domains.

INTRODUCTION

- Bird call classification is crucial for ecological monitoring and biodiversity conservation.
- Advances in deep learning have enhanced automatic bird species classification from vocalizations.
- Adapting pre-trained models like Whisper, designed for human speech, to bird call classification poses challenges due to data distribution mismatches and inadequate feature extraction.
- Ensuring the reliability of these models in non-human acoustic domains is essential for effective wildlife monitoring.
- Adapting pre-trained acoustic models to new domains, especially non-human sounds, is a promising but underexplored research area.

METHODOLOGY

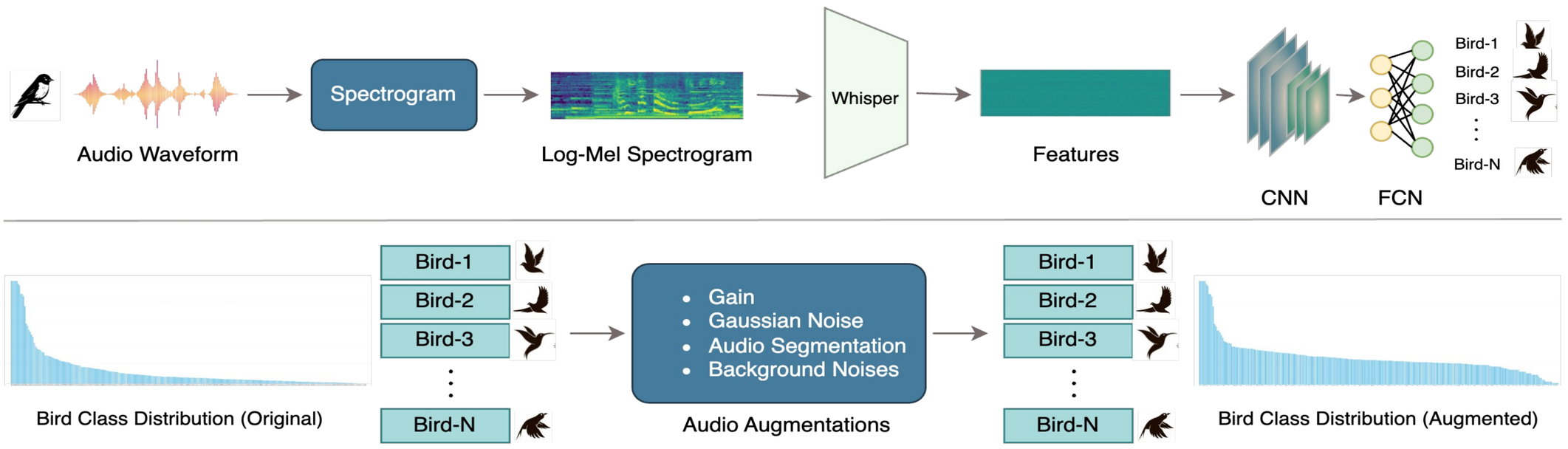

Figure 1: *Methodology Overview: The top row shows the bird call classification pipeline, where the Whisper encoder extracts temporal features from the spectrogram of a bird call, followed by processing through a CNN and FCN to predict bird species. The bottom row highlights data augmentations applied to address class imbalance in the dataset.*

- Utilized the BirdCLEF 2023 dataset, comprising audio recordings from 264 bird species, with significant class imbalance.
- Applied data augmentations, including audio segmentation, Gaussian noise, gain adjustments, and background noise addition, to address class imbalance and enhance model generalization.
- Employed the Whisper model's encoder for feature extraction, followed by processing with a small CNN and fully connected network (FCN) for bird species classification.
- Fine-tuned the Whisper model on bird calls to improve feature extraction, addressing the initial shortcomings in distinguishing avian vocalizations.
- Conducted experiments comparing different training strategies, including full training, linear probing, and fine-tuning, to evaluate model performance.

RESULTS

We assessed the model's performance using macro-averaged F1-score across different training configurations on both original and augmented datasets (see Table 1). The configurations included **Full Training** (training from scratch), **Linear Probing** (freezing the Whisper encoder and training only the final layers), and **Fine-Tuning** (updating the entire model from pre-trained weights). Data augmentation significantly improved performance, with Fine-Tuning yielding the highest F1-scores, underscoring the effectiveness of adapting pre-trained models for bird call classification.

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**

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

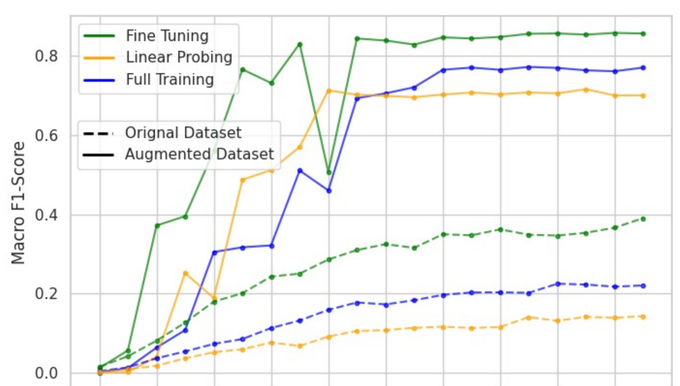

Figure 2: F1-score over different epochs while training Whisper-Base encoder on BirdCLEF-2023 dataset. Model converges faster on augmented dataset and Fine-Tuning strategy as compared to other settings.

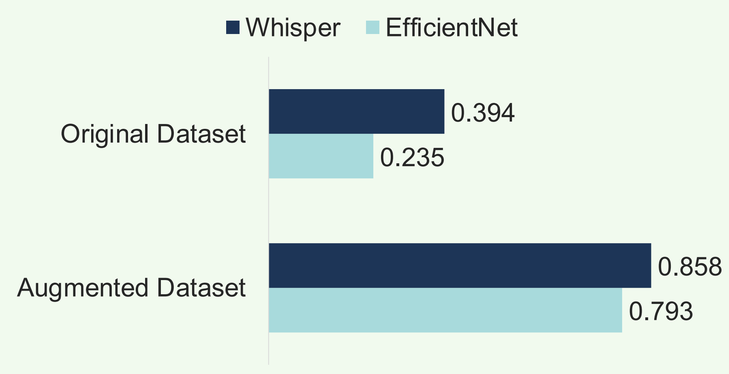

Figure 3: Comparison of EfficientNet-B4 and Whisper-Base encoder (F1-score) under fine-tuning shows Whisper outperforms EfficientNet on both original and augmented datasets.

CONCLUSION

This study demonstrates the successful adaptation of large pre-trained acoustic models, originally designed for human speech, to bird call classification. The results highlight the effectiveness of using pre-trained models for non-human acoustic tasks, offering potential applications in broader audio classification. By incorporating data augmentations and fine-tuning strategies, we addressed challenges like computational constraints and class imbalance, paving the way for future research aimed at enhancing model adaptability across diverse acoustic domains.

GITHUB REPO