On the Similarity Between Bird Calls Using BirdNET Embeddings

Niels van Harten s1012159

July 2022

1 Introduction

Can we measure the similarity between (collections of) bird calls to do interesting observations? We could use this measure to find better ways to distinguish bird calls for two species or give the incorrect answer options in a multiple choice bird call question. Both are applications of measuring how similar a bird call is to bird calls of another species, *interspecies*. Another aspect of measuring similarity is among multiple bird calls of a single species, *intraspecies*. Could one distinguish different types of calls a bird makes with an automatically generated clustering?

The field of speaker recognition is concerned with identifying persons based on the characteristics of their voice. Current approaches use deep learning methods converting speech into abstract feature vectors, also called embeddings, trying to capture the relevant information for identification [1]. Deep learning methods can also be used for identification of bird species. Birds often have their own distinguishable calls which makes that other features can be used than those for identifying people. BirdNET [2] is an advanced deep classifier identifying bird species in three second long audio segments. It is a decent standard for identifying bird calls even outperforming specific solutions for the BirdCLEF 2022 competition [3]. However, we are not interested in bird species recognition but similarity between (collections of) bird calls. O'Reilly et al. (2015, [9]) propose a measure of bird call similarity inspired by human dialect separation measurement. The measure considers a representation of the pitch contour micro-structure of the bird call. This method is more fully developed and analysed in later research [8]. Other work by O'Reilly et al. (2017, [10]) uses a feature extraction method for bird species recognition. We also extract features from audio fragments but use the deep neural network of BirdNET to do so and use it as similarity measure for (collections of) bird calls. Using cosine similarity and UMAP, we will analyze the similarity of different bird call audio fragments using experiments to see whether the BirdNET embeddings can be used for answering questions about the similarity of bird calls.

2 Method

To see whether we can use BirdNET embeddings as a similarity measure, we use a quantitative experimental approach. We use an experimental methodology while we want to evaluate the applicability of our approach by answering specific questions which can be done by running experiments on our data. In sections 3 and 4 we explain the specific questions we want to answer for the evaluation of our approach and with that our experiments. The data we use for the experiments comes from the Xeno-Canto [11] free-to-use bird sound library. The audio fragments there are community uploaded can be used non-commercially.

We will analyze the data by feeding it into the embed functionality of the BirdNET Analyzer. It transforms audio recordings in feature vectors consisting of 420 floating point numbers for every three second long segment and tries to capture the identification characteristics of the audio fragment. We can subsequently compare these embeddings and do so by making use of cosine similarity. A relatively simplistic metric used regularly for the comparison of feature vectors [1], [7]. To compare collections of bird calls with each other, we make use of *mean embeddings*. A mean embedding contains for each dimension the average of that dimension for all embeddings of the collection. Given this average representation for a group, we can compare groups of bird calls as being a single call, losing some details in the process.

For distinguishing different types of calls we want to be able to plot embeddings to discover groups of embeddings that are close together which should be similar. However, to do so, we need to reduce the dimensionality of the embeddings from 420 to either two (2D) or three (3D). We do so using UMAP [6], an often-used manifold learning

technique for dimensionality reduction, applicable on real world data. After having reduced the dimensionality, we can plot the reduced embeddings. By reducing the dimensionality, we lose some of the information of the original feature vector. However, otherwise we would not be able to visualize the data and UMAP works in such a way that information about similarity is kept as much as possible.

3 Set-up

To aid reproducibility of the results, our source code is publicly available at `https://github.com/nielsvharten/birdnet_embedding_analyzer/`.

3.1 Data

From the Xeno-Canto library we limit ourselves to Dutch recordings with the highest quality rating A and at least twenty recordings per species. We are interested in the BirdNET embeddings which we analyze. These are feature vectors with 420 dimensions (floats) for every three second long segment. However, the data is weakly labeled. Each recording has a label containing the species for which the fragment is recorded. So, not every three second long segment contains the labeled bird and only that bird. Therefore, we first want to strengthen our labels. We do this by using BirdNET as classifier, ignoring all three second long segments predicted to contain other bird species or not to contain the bird species of the main label. For this we use a minimal confidence level of 0.5. This cleans to some extend background bird calls and incorrect labeling for a segment. It is not perfect but better than leaving it as noisy as is. Besides, we do not have access to the same amount of strongly labeled data which would have been ideal. The amount of data we use for our experiments can be found in Table 1. The recordings are the audio recordings retrieved from Xeno-Canto. The embeddings are generated for three second long segments predicted by the BirdNET classifier to contain calls of the bird of the main label without any other species calls. Note that for some recordings we do not use a single embedding while for others we use multiple.

3.2 Experimental

First, we calculate the accuracy classifying the feature vectors of all three second long audio fragments to the mean embedding most similar using cosine similarity. This way we quantitatively measure how well the mean embedding for a species represents all species bird calls and is able to distinguish calls of different species. We considered using a confusion matrix. However, given the many classes and large number of embeddings, and to make results more clear we restrict ourselves to classification accuracy. We also reduce the dimensionality from 420 to two and three using UMAP and subsequently calculate the accuracy. This is done as we use dimensionality reduction later for visualization purposes and want to know to what extend it is able to cluster species. One should note that dimensionality reduction using all 86 species is significantly harder than dimensionality reduction for a single or a couple of species. As in the later case, it can focus on the values that differ for those birds specifically.

Next, we execute experiments using our similarity score to answer multiple questions about the similarity among calls of a bird species and between multiple bird species. We do these experiments both to illustrate that our approach yields fair results as well as a contribution in itself.

For intraspecies call diversity we execute the following experiments:

a) We give a ranking of the most and least diverse species based on the average cosine distance to its mean embedding for all the embeddings of that species. One would think that species with more diverse calls would be harder to recognize. Can we make a ranking that makes sense?

	Recordings	Embeddings
#fragments	4068	55750
#classes	87	86
min(fragments per class)	20	67
max(fragments per class)	148	3788
avg(fragments per class)	46.76	648.26

Table 1: Amount of data used for the experiments

b) For one of the more diverse species based on call, we reduce the dimensionality to two. But any species could be an interesting choice for closer inspection. Next, we plot all the reduced embeddings to see whether clusters can be found. Finally, we check six interesting embeddings from the plot to see whether different types of calls can be distinguished from the plot.

For interspecies call similarity we try obtaining knowledge with the following experiments:

a) We check how similar each pair of two species is. We do this by calculating the distance between each two mean embeddings. We are mainly interested in the most similar bird pairs as we suspect those to be hard to distinguish.

b) We plot the mean embedding for all birds which should give a good overview of which birds are more and less similar to each other. We compare this to the results of a) checking whether they are similar.

c) We plot two cases of similar pairs using UMAP to show how well the two classes are separated. It could also show interesting edge cases or outliers which might be interesting to analyze manually in later research. It is not done here due to time constraints.

Not all parameters for UMAP are self-evident and differ per situation. Therefore, we consider which basic parameters [5] to use.

- As n_neighbors, we use different numbers for different use cases. When reducing dimensionality for all embeddings and the mean embeddings for the species we use the value of 200 for 2D and 150 for 3D as it seems to be a good trade-off and resulted in best classification accuracies. When comparing the embeddings for two birds or comparing the mean embedddings we use the default value of 15 as in those cases much less data points are considered and with a value of 15, results look good.
- For min_dist, we used 0.0 in cases where we compared multiple embeddings for the the same species. As multiple embeddings for one species could well be very similar. Especially because multiple embeddings can be generated from the same recording, containing the exact same bird and having the same background noise. When using only mean embeddings we kept min_dist at the default 0.1.
- For n_components, we considered only two for generating plots as 2D-plots are the most intuitive to use though extensions to 3D could be made. Both two and three components are used for calculating accuracy. We included three to see how much better classification accuracy is having an extra dimension.
- As metric, we used "cosine" as we use cosine similarity as our similarity measure. Besides, we found classification accuracy for 2D to be significantly higher using cosine compared to euclidean.

4 Experiments

4.1 Classification with embeddings

We classify all 55750 embeddings to the bird species of the mean embedding with the minimal cosine distance to the embedding. As measure, we use accuracy, i.e. the percentage of correctly classified birds.

We do this for the embeddings generated by BirdNET with 420 dimensions as well as embeddings with dimensionality reduction applied using UMAP with the following parameters:
For 3D: random_state=42, n_components=3, min_dist=0.0, n_neighbors=150, metric="cosine"
For 2D: random_state=42, n_components=2, min_dist=0.0, n_neighbors=200, metric="cosine"

4.2 Intraspecies call diversity

For each bird species we calculate its call diversity. We do so in the following way. First, we calculate its mean embedding as explained in Sec 2. Next, we calculate the cosine distance for each embedding to the mean embedding and take the average of these distances. This metric we will use as diversity measure. We can than rank all species by diversity and show the most and least diverse species according to our metric.

Of the more diverse birds, we look at the Little Grebe more closely. We reduce the dimensionality for all its

Dimensionality	Accuracy
420 dimensions	98.2%
3 dimensions	90.4%
2 dimensions	70.9%

Table 2: Accuracy classifying all embeddings to the mean embedding most similar.

embeddings to two using UMAP with the following parameters:
random_state=42, min_dist=0.0, n_neighbors=15, metric="cosine"

Next, we plot all these reduced embeddings and pick six embeddings which we deem interesting. Of these six embeddings we show its spectrogram and we annotate these in the plot.

4.3 Interspecies call similarity

We calculate the similarity between each pair of species using the cosine distance between the means of the two birds. We rank all pairs by similarity and show the five most and five least similar pairs of species.

Next, we reduce the dimensionality for all mean embeddings to two using UMAP and plot these. As important parameters for UMAP we use: random_state=42, min_dist=0.1, n_neighbors=15, metric="cosine".

Furthermore, we again reduce the dimensionality to two using UMAP to being able to plot embeddings. However, we now do so for all the embeddings of a pair of bird species. For UMAP we use the following parameters: random_state=42, min_dist=0.0, n_neighbors=15, metric="cosine". We do this for the pair *Eurasian Blackbird* & *Song Thrush* as well as for the pair *Great Spotted Woodpecker* & *Lesser Spotted Woodpecker*.

5 Analysis & Results

5.1 Classification with embeddings

The classification accuracy using our approach of Section 4.1 can be found in Table 2. Given the full feature vector, the mean embedding of each bird is a very good representation for classifying a call, correctly classifying 98.2%. When dimensionality reduction is applied using UMAP the classification accuracy decreases as suspected, as the amount of information stored in the embedding is greatly reduced. However, the accuracy is still 90.4% having three and 70.9% having only two dimensions. So, with three or even two dimensions the method is able to classify bird calls quite well. This supports the use of UMAP for visualization and analyzation purposes.

5.2 Intraspecies call diversity

By looking at the average cosine distance to its mean embedding for all segments of each bird species, we rank how diverse the embeddings are which should be a good indication of the diversity of the calls. As can be seen in table 3, the difference in average distance to the mean between all birds is not that large. With the most diverse having a somewhat less than three times as large average distance than the least. It seems that there exists a significant distance of a few percent even if calls do not differ much. The order of the results seem very reasonable. When listening to the different bird calls for the five least diverse bird calls resulting from the algorithm, it seems that those birds have only a few different calls which are all highly repetitive, do not contain much of a melody and sound quite alike. The most diverse bird calls on the other hand have multiple different calls that differ much from each other.

Least diverse bird calls	avg distance mean	Most diverse bird calls	avg distance mean
Spotted Crake	3.92%	European Starling	9.33%
Eurasian Jackdaw	4.81%	Little Grebe	8.13%
Eurasian Nightjar	4.87%	Icterine Warbler	7.96%
Common Quail	5.03%	Song Trush	7.80%
Common Grasshopper-Warbler	5.07%	Blyth's Reed Warbler	7.78%

Table 3: The most and least diverse species calls based on average cosine distance to its mean.

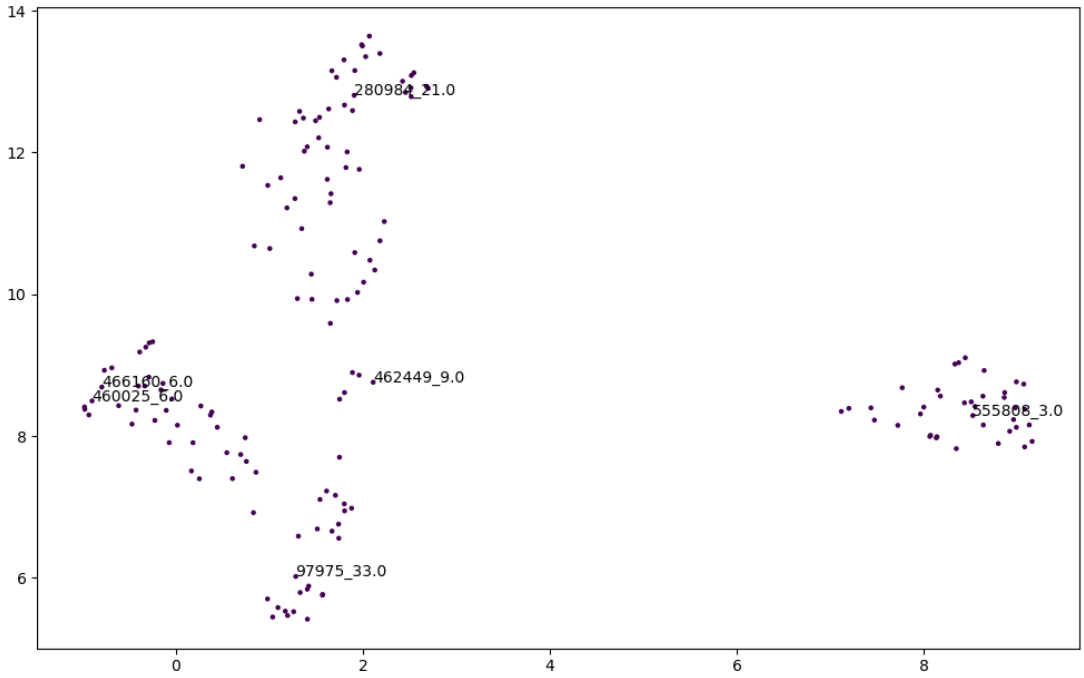
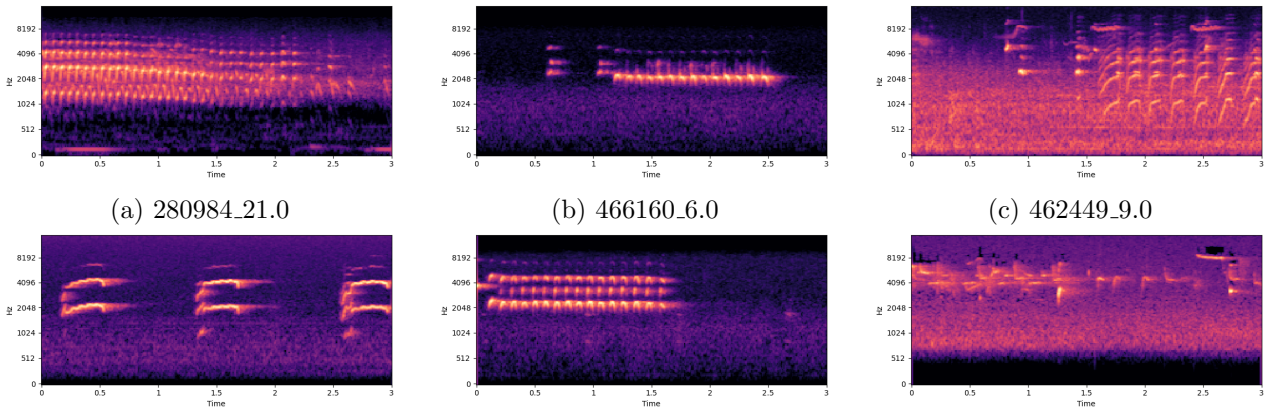


Figure 1: All embeddings for the Little Grebe reduced using UMAP.

We can also look at the embeddings for a single bird. In Figure 1, all embeddings for the Little Grebe, a bird with relatively diverse calls, are reduced to two dimensions using UMAP and subsequently plotted. As can be seen, some clusters appear. Six interesting segments are annotated and displayed in Figure 2. It seems to do a quite decent job of discovering types of calls given that 466160_6.0 is both close to and quite similar to 460024_6.0. Besides, the other embeddings being further apart seem to contain significant differences. But a more thorough analysis for multiple bird species would be useful and needed for drawing any strong conclusions.



(a) 280984_21.0 (b) 466160_6.0 (c) 462449_9.0

(d) 555808_3.0 (e) 460024_6.0 (f) 97975_33.0

Figure 2: Six interesting 3s long audio fragments of the Little Grebe, annotated in Figure 1.

Most similar bird pairs	Distance	Least similar bird pairs	Distance
Blyth's Reed Warbler vs Marsh Warbler	2.90%	Gadwall vs House Sparrow	11.92%
Eurasian Blackcap vs Garden Warbler	3.16%	Common Grasshopper-Warbler vs Eurasian Jackdaw	11.74%
Eurasian Blackbird vs Song Thrush	3.23%	House Sparrow vs Mute Swan	11.63%
Icterine Warbler vs Marsh Warbler	3.31%	Common Grasshopper-Warbler vs Mute Swan	11.63%
Bluethroat vs Marsh Warbler	3.37%	Common Grasshopper-Warbler vs Common Raven	11.55%

Table 4: The most and least similar bird call pairs based on cosine distance between means.

5.3 Interspecies call similarity

We first calculate the distances between all pair of birds, rank them and show the five most and least similar birds in Table 4. Results for the most similar pairs seem quite accurate. For example, the Eurasian Blackcap and Garden Warbler are hard to distinguish [4] and the other bird species are quite similar. The least similar birds seem reasonable, though we cannot think of practical use cases for these results.

When plotting all mean embeddings, as shown in Figure 3, a quite similar picture appears, where the most similar pairs of Table 4 are relatively close together and the least similar pairs are relatively far apart. However, other most similar and least similar pairs appear. For example, European Goldfinch and Greenfinch are the closest together but not mentioned in the list of most similar bird pairs. It could be that for the two results other types of birds are found to be most closely together. When comparing pairs, only certain types of song birds are in the most similar five while for the plot other types of bird species are also in the top five of closest birds.

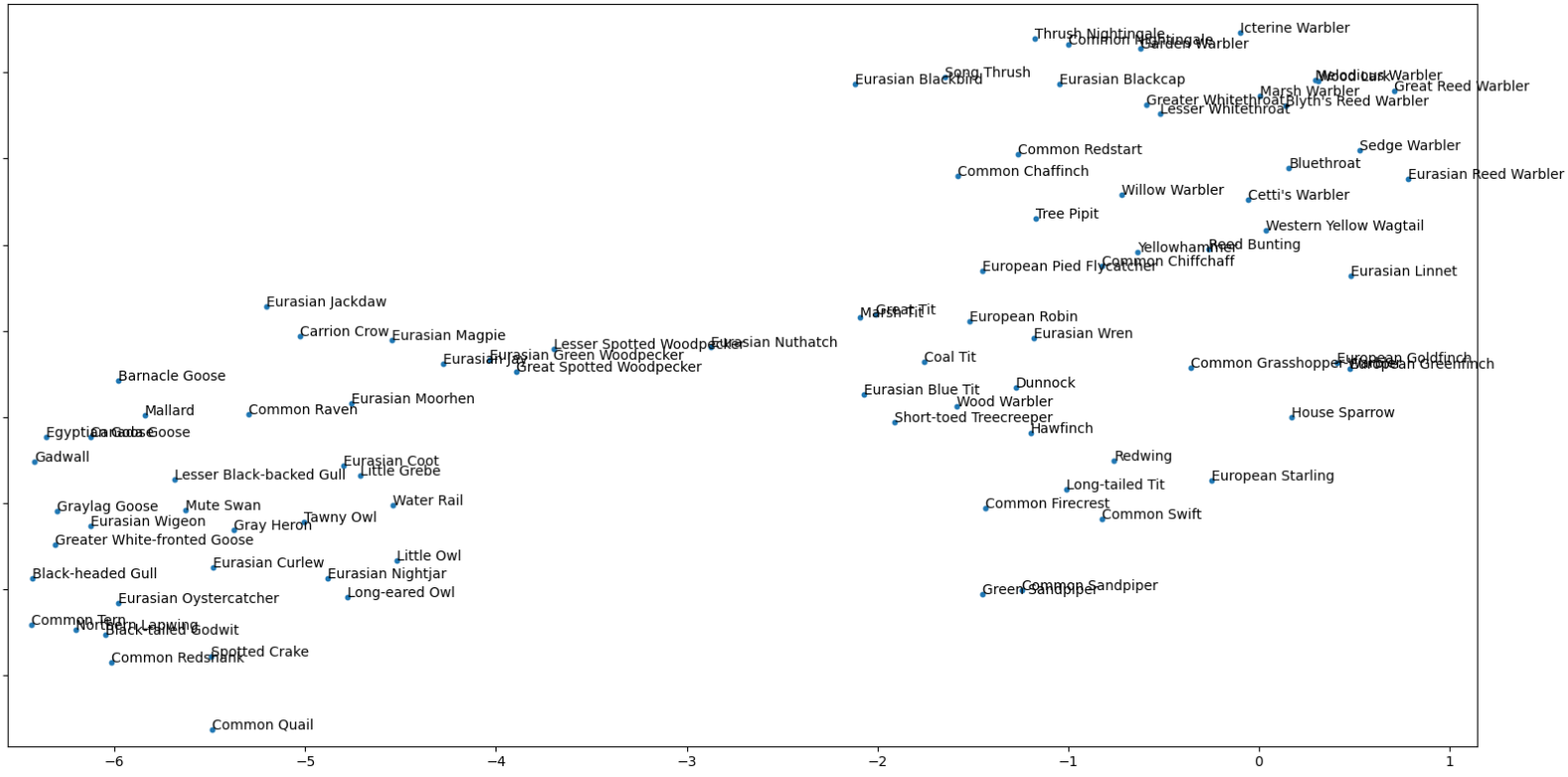
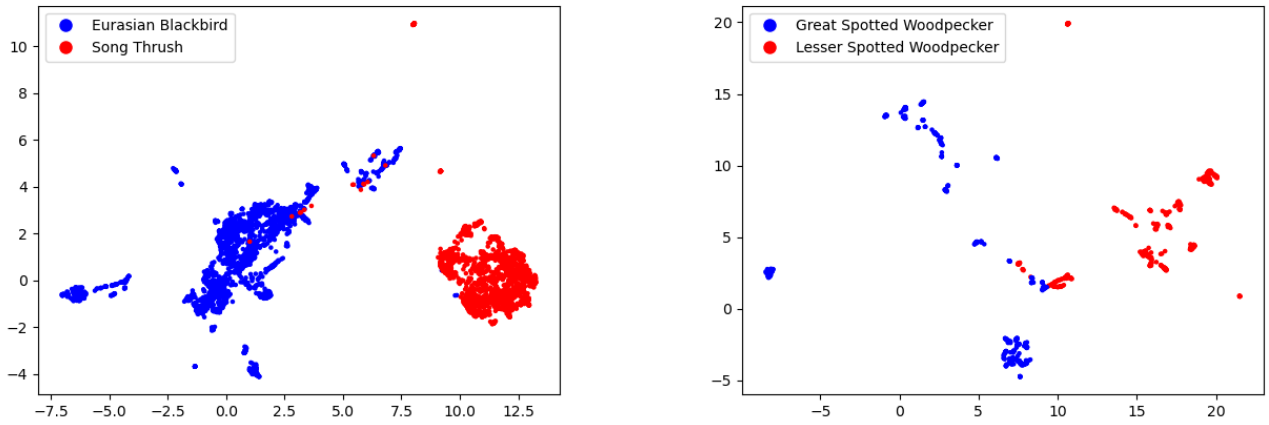


Figure 3: An overview of all the mean embeddings plotted together.

In figure 4 we only look at two pairs of birds for which all embeddings are plotted. As can be seen, a) and b) are clearly different plots. a), containing the Blackbird and Song Thrush contains dense clusters with a few outliers. Most interesting could be the few Song Thrush outliers that lie precisely in a Blackbird cluster. It would be interesting to compare those two. For the Great & Lesser Spotted Woodpecker the picture is different showing a more spread out picture. There is a region in which both Lesser and Greater Spotted Woodpecker segments exist. Besides that,



(a) Blackbird & Song Thrush

(b) Great & Lesser Spotted Woodpecker

Figure 4: Plots showing all embeddings for a pair of bird species.

most points are separated well. All in all, it seems that even for birds with similar calls can still be separated quite well in two dimensions.

6 Discussion

It seems that BirdNET's feature vectors in combination with cosine similarity and UMAP can be really useful as you are able to compare calls and groups of calls in ways that are not possible using only the BirdNET classifier. One could extend this research in many ways. First, one could extend to using other/more bird species. Also, it would be interesting to look in more detail at specific species. By for example looking at edge cases, maybe even to try improving the BirdNET classifier, or by analyzing clusters of calls generated by plotting all embeddings for a bird. Besides, a visualization extension could be made using 3D plots instead of 2D having information richer datapoints. Furthermore, one could try using another similarity measure than cosine similarity or another dimensionality reduction technique than UMAP. Another research direction is that as stated, one of the limitations of our technique is that the dataset we use is filtered using BirdNET classification. Manipulating the data in a way that could influence results negatively. Therefore, using a dataset with strongly labeled audio fragments per three seconds segment would be better and a comparison with our data interesting. Last, one could generalize our results to other of sound recognition, for example human speech. Using an embedder optimized for speaker recognition it would be possible to conduct similar experiments on the similarity of human speech. For example, how much do family members sound alike or how (much) does your speech change in different settings or when sick?

7 Conclusion

We looked at the similarity of (groups of) bird calls using features vectors of BirdNET to see whether we can retrieve useful information from them. Classification using mean embeddings worked very well and even decently after dimensionality reduction to two or three components using UMAP indicating the strength of the embeddings. Looking at the different calls for a single species we used the average distance to its mean to determine which species are most and least diverse. Plotting all reduced embeddings for a species might be used to discover the types of calls for a bird as one can look at multiple individuals that are either close together or far apart more closely. On the similarity between different bird species, we looked at which birds are most and least similar both using distance between their mean embeddings and by plotting all reduced mean embeddings. Also, for two pairs of species we plotted all embeddings, indicating possibly interesting outliers as well as which groups of calls for a bird are most similar to which group of calls of the other species. All in all, we find analysis of bird calls using BirdNET embeddings to be a promising research direction.

References

[1] Zhongxin Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99.

[2] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* 61 (2021), 101236.

[3] LeonShangguan. 2022. BirdCLEF 2022 [Public 1 Private 2] + [Private 7/8 (potential)] solutions. The host wins. `https://www.kaggle.com/competitions/birdclef-2022/discussion/326950`. Accessed: 2022-07-04.

[4] Amy Lewis. 2022. Bird song identification: UK warblers. `https://www.woodlandtrust.org.uk/blog/2022/04/warbler-song-identification/`. Accessed: 2022-07-11.

[5] Leland McInnes. 2022. Basic UMAP Parameters. `https://umap-learn.readthedocs.io/en/latest/parameters.html`. Accessed: 2022-07-11.

[6] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[7] Hieu V Nguyen and Li Bai. 2010. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*. Springer, 709–720.

[8] Colm O'Reilly, Kangkuso Analuddin, David J Kelly, and Naomi Harte. 2018. Measuring vocal difference in bird population pairs. *The Journal of the Acoustical Society of America* 143, 3 (2018), 1658–1671.

[9] Colm O'Reilly, Nicola M Marples, David J Kelly, and Naomi Harte. 2015. Quantifying difference in vocalizations of bird populations. In *Sixteenth Annual Conference of the International Speech Communication Association.*

[10] Colm Or'Reilly, Münevver Kcöküer, Peter Jančović, Regan Drennan, and Naomi Harte. 2017. Automatic frequency feature extraction for bird species delimitation. In *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 1959–1763.

[11] Willem-Pier Vellinga and Robert Planqué. 2015. The Xeno-canto Collection and its Relation to Sound Recognition and Classification.. In *CLEF (Working Notes)*.