DEDUCTIVE REFINEMENT OF SPECIES LABELLING IN WEAKLY LABELLED BIRDSONG RECORDINGS

Veronica Morfi, Dan Stowell

Machine Listening Lab, Centre for Digital Music (C4DM), Queen Mary University of London, UK

g.v.morfi@qmul.ac.uk, dan.stowell@qmul.ac.uk

ABSTRACT

Many approaches have been used in bird species classification from their sound in order to provide labels for the whole of a recording. However, a more precise classification of each bird vocalization would be of great importance to the use and management of sound archives and bird monitoring. In this work, we introduce a technique that using a two step process can first automatically detect all bird vocalizations and then, with the use of 'weakly' labelled recordings, classify them. Evaluations of our proposed method show that it achieves a correct classification of 75.4% when used in a synthetic dataset.

Index Terms— bird species classification, event detection, cross-correlation, weak labelling, computational auditory scene analysis

1. INTRODUCTION

The potential applications of automatic species detection and classification of birds from their sounds are many (e.g. ecology, archival). However, automated species identification is a challenging task due to the complexity of bird song, the noise present in most habitats, and the simultaneous song that occurs in many bird communities [1] [2]. Many authors have proposed methods for bird species classification (See [3] for a survey). However, more work is needed to address the problem of identifying all species and the exact times of their vocalizations in noisy recordings with multiple birds. These tasks need to be achieved with minimal manual intervention, in particular without manual segmentation of recordings into birdsong syllables. Some early studies used small datasets, often noise-free and/or manually segmented and with a small number of species. More recent studies have fewer limitations, and introduce useful methods customised to the task [4] [5] [6]. However, these methods are only used for labelling the recordings (identifying the species present) and are not sufficient for detecting the exact times of the vocalizations. Furthermore, techniques for automatic detection of audio events have been of interest to many authors [7] [8] [9] [10] [11] [12].

In this work, our aim is to implement a two step process that using 'weakly' labelled birdsong recordings can automatically detect each bird vocalization, and then classify it to one of these weak labels. By 'weakly' labelled we refer to recordings that are annotated with which bird species are active, but have no information about which individual vocalizations are produced by which species. While manual annotation of the dataset is required to acquire the 'weakly' labelled dataset, precise vocalization annotation is a much more time consuming process which requires expert knowledge. Additionally, there are already quite a few public datasets labelled with the species present in each recording and a lot of methods have been implemented in order to achieve a semi-automatic recording labelling [13] [4] [5] [6]. In order to implement our method, we propose a segmentation-detection process inspired by previously proposed ones [14] [15] [16], followed by a classification process based on finding the best visually similar match of a segment throughout the whole dataset and deductively refining its possible labels.

In the rest of the paper, Section 2 presents our proposed two step process. The evaluation follows in Section 3 with the necessary discussions and conclusions in Section 4.

2. PROPOSED METHOD

To achieve our goal, we implement a two step process, first a segmentation-detection algorithm that detects all vocalizations, followed by a classification method that labels the segments in question.

2.1. Segmentation-Detection

The unsupervised extraction of vocalization segments is of great importance to our classification task. For this process, we employ the event detection paradigm used by Fodor [14], Lasseck [15] and Potamitis [16]. All three methods are very closely related but have some differences. In general, the method proposed by Lasseck produces less segments than the others and is very effective in handling noise (See [16] for an in depth comparison of the three methods). For our purposes, a method that is robust to noise and does not generate noise segments is of great importance. Hence, we implement a close variation and refinement of the Lasseck segmentation, that produces even fewer noise segments and fewer, yet larger, vocalization segments, which is what works better with our proposed classification process.

First we obtain the spectrogram (time-frequency representation) of a recording via the Short-Time Fourier Transform of the librosa library (i.e. *librosa.core.stft*), with window size of 512, Hann window and overlap of 75%. Then, the following steps are performed for the spectrogram derived from each recording (cf. [15]):

1. normalize the spectrogram values to 1.0 using its absolute max value;
2. remove frequencies above 20 kHz and below 340 Hz. Since no bird vocalizations occur in those frequencies, the only audio present there can be considered noise;
3. get binary image via median clipping per frequency and time frame in order to eliminate any noise: we set pixel to 1 if its value is above 3 times the median of its corresponding row and column, otherwise it is set to 0;
4. apply closing [17, pp.657–661] in order to fill any small holes in a present feature (i.e. vocalizations). Closing is applied in a rectangle neighbourhood of size (3,3);
5. remove connected components of less than 5 pixels;
6. apply dilation [17, pp.655–657] in a rectangle neighbourhood of size (7,7). Dilation sets a pixel at (i,j) to the maximum over all pixels in the neighbourhood centred at (i,j). Dilation is applied in order to enlarge the regions that contain features (i.e. vocalizations) and remove small objects that can be considered noise;
7. apply a median filter of size 5;
8. remove connected components of less than 150 pixels;
9. re-apply dilation in a circular region of radius 3;
10. define all connected pixels as a segment (seg_i);
11. find each segment's size and position.

In our implementation, an extra step, compared to the Lasseck method, of removing small segments (step 5) is added before applying the first dilation (step 6). Since dilation enlarges regions where features are present, using dilation without first removing small objects results in expanding these regions. However, such small segments are in majority caused by noise and are not actual vocalizations. Eliminating them in this early step can further reduce the noisy segments produced at the end of the segmentation-detection process. Additionally, an extra dilation (step 9) is applied at the end of the algorithm. This second dilation has a much smaller neighbourhood (disk of radius 3) than the first one and it is used as a refined way of slightly expanding the borders of the segments detected and filling any small holes still present. This is especially helpful in larger vocalizations which are sometimes split into multiple smaller vocalizations since this dilation can connect two vocalizations if they are close enough (depending on the dilation neighbourhood) to each other. Compared to the original algorithm presented by Lasseck, this variation produces less noise segments and fewer, but larger, vocalization segments.

2.2. Classification

Following this segmentation-detection process, an instance based classification algorithm with no explicit training phase is implemented. In our approach, 'weakly' labelled recordings are used. Hence, the species present are the labels of that recording ($labels\_rec$), however, we have no further information as to the specific vocalizations. For each recording, the segments that derive from the segmentation-detection process (seg_i) are considered to be attributable to vocalizations from the bird species included in the weak labels.

For each segment, we create a list of possible labels ($labels\_seg_i$), initialized to the weak labels of the recording that contains the segment. The $labels\_seg_i$ list of a segment will later on be shortened to either one or multiple labels by the classification process via deductive elimination of the less possible labels for that segment. During classification, each segment in need of labelling is matched using normalized correlation (scikit-image's *match_template* function) to different recordings in order to obtain all the possible label matches. In *match_template*, normalized correlation is used to match a template (vocalization) to a 2D target image (spectrogram of a recording). The result is a response image of same size as the target image, with correlation coefficients between the template and target image of values between -1.0 and 1.0. The matching value between a segment and a specific recording is found by searching for the maximum peak in the response image. Due to the number of recordings and segments detected in each of them, this process is very time consuming. However, similar bird sounds should appear in similar frequencies, hence we reduce the computational load by only applying *match_template* to a smaller range of frequencies (5 frequency bins below and above the segment frequencies). Furthermore, since the weak labels of a recording and a segment are already known, we only need to search for a segment match in recordings that contain at least one of the segment labels ($labels\_seg_i$).

The proposed classification has no need for a separate training set of recordings as it can classify recordings by finding matches between them. The performance of the method increases as the number of recordings per each species increases. The chance of the classification process finding a match for a segment increases along the variation of each species' vocalizations. The classification process is implemented in three different procedures, namely the First-Pass, Second-Pass and Third-Pass. All three are applied to the recordings in order, as explained in the following subsections and illustrated in Figure 1.

2.2.1. First-Pass

In the First-Pass of the classification, in order to best utilize the information provided by the weak labels, we create groups of recordings $recs(c_i)$ for each segment seg_i to find matches with, where c_i denotes the different label combinations produced by the initialised $labels\_seg_i$ list. The record-

ings in $recs(c_i)$ have label(s) c_i present in their weak labels. For each segment in need of a label the matching process will search through the list of recordings $recs(c_i)$ increasing the number of weak labels (i.e. $|c_i| = 1, 2, 3, ...$) until a match is found or there are no more recordings remaining. Since *match_template* always returns a result (maximum peak in the response image), in our implementation, we consider that a match is found when the similarity rate returned by *match_template* is 0.4 or greater. The 0.4 threshold was obtained after preliminary experimentation. All the different values of the matches found in these recordings for each possible label combination c_i will be summed and the label(s) with the highest sum (C_i) will be assigned as the segments label(s). If no match is found in $recs(c_i)$ the Match Not Found (MNF) label is assigned to seg_i. Segments with the MNF label and segments that have more than one possible labels in $labels\_seg_i$ are classified as Unknown in our evaluation results (Section 3), even if the correct segment label is between the multiple possible labels.

2.2.2. Second-Pass

The Second-Pass of the process derived from the need to solve the issue of unclassified segments, MNF segments, produced through the First-Pass of the classification. Since we use only weakly labelled datasets, all the labels of a recording must be assigned to at least one segment. A trivial solution of reducing the MNF segments is: when there are MNF segments and labels with no corresponding segments in a recording, we assign the unallocated labels to all the MNF segments. This will solve the issue of unallocated labels and MNF segments in a recording but will not completely eliminate the Unknown segments (MNF segments and segments with multiple labels), since more than one label may be unallocated and thus assigned to a single segment. Case 1 in Figure 1 depicts what happens during the Second-Pass when there is an unallocated label (label B) and an MNF segment (segment 4). In this case, the unallocated label is be assigned to segment 4.

2.2.3. Third-Pass

After reducing the MNF segments, there may still be labels unallocated in some recordings. Hence, the Third-Pass of the classification process derived from the need for all labels of a recording to get assigned to at least one segment. More specifically, in a recording for which all segments have labels but some of the weak labels of the recording are not assigned to any segments, there must be some labels that are assigned, most likely incorrectly, to more than one segments. It is possible that more than one segment may have this label, but when a label is unallocated then we assume that one of the segments matched to the same label is falsely classified. We search for the best match for any unallocated label among the multiple segments of the rest of the labels. If a match is found, the label of the segment it derives from is changed to the unallocated label. An example of the Third-Pass is depicted in case 2 of Figure 1, where all segments have labels assigned to them, however label B is not assigned to any of them. The best match with label B is found within the segments that have the same label (segments 2, 3 and 4). Segment 4 has the max match of 0.57, thus label B is be assigned to it.

3. EVALUATION

To our knowledge, there is currently no public dataset with strong time-frequency labelling of each bird vocalization. Thus, in order to evaluate our proposed method, we created a synthetic dataset where the boundaries of each vocalization are known. The audio dataset provided during the Neural Information Processing Scaled for Bioacoustics (NIPS4B) bird song competition of 2013[1] contains recordings that have already been weakly labelled. Since there is no per unit annotation in it, we created a synthetic dataset of 50 recordings with vocalizations deriving from the single labelled recordings in the NIPS4B dataset. Out of the 87 labels of the NIPS4B dataset, 51 have recordings that are labelled with only one species. Each synthetic recording is 5 seconds long and it consists of one of the recordings of NIPS4B with no labels, hence containing only natural background noise in it. Each synthetic recording is allocated 2 to 4 randomly picked labels out of the above mentioned 51 labels. A source recording is randomly picked for each of the labels and from that recording one segment produced by our proposed segmentation-detection process is placed in the synthetic recording. Thus, each synthetic recording contains 2 to 4 segments. The resulting dataset consists of 50 recordings, with a total of 138 segments, hence a mean of 2.76 segments per recording. We use the remainder of the original NIPS4B dataset in order to search for the segment matches, hence providing our classification process with a broader variation of species' vocalizations than the one available by using only the synthetic dataset. The boundaries of each segment in the synthetic recordings are known, hence the following evaluation measures are only for the classification process and its different procedures.

In Table 1, the results of the segment classification using all three passes are depicted. The First-Pass produces a correct classification of 68.9% and 6.5% of Unknown segments, the latest one includes segments that are either not matched to any label (MNF label) or have more than one labels. After the Second-Pass of the algorithm, the percentage of Unknown segments is reduced to 4.4%, while the correctly classified segments are increased. Finally, after the Third-Pass, we have a slight increase to the number of correct classifications, namely 4.4%, which leads to the total result of 75.4% correctly classified segments.

Most of the misclassifications happen due to the fact that the segmentation-detection process produces a lot of smaller segments that usually contain very simple vocalizations, and in many cases, fragments of vocalizations, that can

[1] http://sabiod.univ-tln.fr/nips4b/challenge1.html

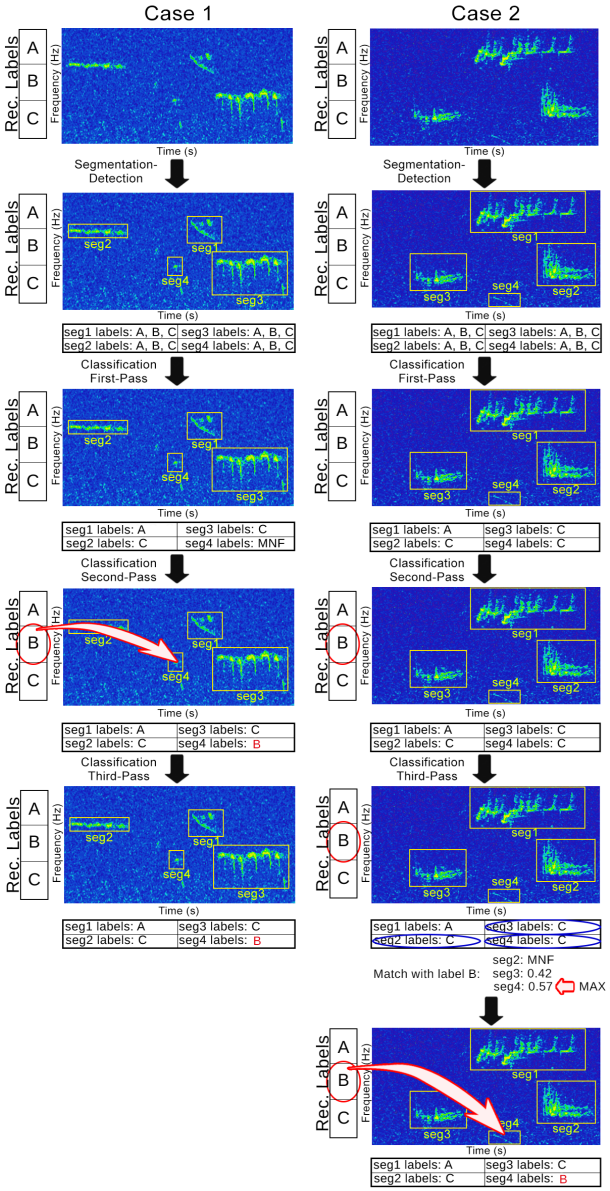

Fig. 1. Example of the proposed two step process. Case 1 describes what happens when there is an unallocated label and a segment with MNF label. Case 2 describes what happens when there is an unallocated label and multiple segments have one of the other labels.

be matched to multiple labels easily. In the event that the segments are part of vocalizations they are considered 'out of context'. When there are 'out of context' segments the classification results can be verified through a process of inverse matching. More explicitly, checking the recording where a match is found to see if it was matched to a single segment or a part of a bigger segment, by checking the area around where the match was found. If the segment is matched to part of a bigger vocalization then it must have the remaining of the vocalization at a close by area in order for it to be considered a correct classification. However, inverse matching cannot be applied in the synthetic dataset case, because the segments are chosen at random, so they are not placed together with the rest of the vocalization.

Table 1. Classification Results for D

	Correct	Wrong	Unknown
Chance	36.2%	63.8%	—
First-Pass	68.9%	24.6%	6.5%
Second-Pass	71%	24.6%	4.4%
Third-Pass	75.4%	20.2%	4.4%

Table 2. Classification Results for D_{1000}

	Correct	Wrong	Unknown
Chance	32.89%	67.11%	—
First-Pass	66.5%	21.7%	11.8%
Second-Pass	71%	22.4%	6.6%
Third-Pass	74.3%	19.1%	6.6%

In order to evaluate classification when the 'out of context' problem does not occur as often, we created a second synthetic dataset (D_{1000}) of 50 recordings, where segment size $\geq$ 1000 pixels. In this dataset, each recording contains 2 to 4 labels, and in total there are 152 segments, hence a mean of 3.04 segments per recording. The results produced by the different classification steps are shown in Table 2.

In the evaluation of the classification process using D_{1000} (Table 2) almost the same results as the one produced by dataset D can be noticed. This indicates that smaller segments are not the limiting factor in classification performance. In the D_{1000} results, even though the missclassifications of most of the smaller 'out of context' segments are not present, still there are segments with simple structure (e.g. a straight line in frequency or time), which can get matched to larger vocalizations. This can be solved with an inverse matching process which we will explore in future work.

4. CONCLUSIONS

Taking advantage of the good bird species classification results produced by image segmentation and event detection methods, we proposed a two step process that can be applied to 'weakly' labelled recordings. Our method is used to fully annotate recordings at a unit level instead of finding the species present. The first step of our approach implements a fully automatic way of extracting vocalizations from each recording using its corresponding spectrogram. In the second step of our proposed process, we are able to reduce the possible labels of each detected vocalization by utilizing the information provided by the weak labels of a set of recordings and using cross-correlation to find the best visible match of a vocalization. According to the assessment of correct classification, in our synthetic dataset, our two step process achieves up to 75.4% successful classification per vocalization.

5. REFERENCES

[1] D. Luther, "Signaller: Receiver coordination and the timing of communication in amazonian birds," *Biology Letters*, vol. 4, pp. 651–654, 2008.

[2] D. Luther and R. Wiley, "Production and perception of communicatory signals in a noisy environment," *Biology Letters*, vol. 5, pp. 183–187, 2009.

[3] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[4] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes, "Bayesian classification of flight calls with a novel dynamic time warping kernel," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, Dec 2010, pp. 424–429.

[5] F. Briggs, B. Lakshminarayanan, L. Neal, X. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustic Society of America*, vol. 131, pp. 4640–4650, 2014.

[6] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1541–1550, Nov 2008.

[7] K. Lee, D. Ellis, and A. Loui, "Detecting local semantic concept in environmental sounds using Markov model based clustering," in *Proc. IEEE ICASSP*, 2010.

[8] L. Lu, F. Ge, Q. Zhao, and Y. Yan, "A SVM-based audio event detection system," in *International Conference on Electronical and Control Engineering*, 2010.

[9] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words approach for multimedia event classification," in *Interspeech*, 2012.

[10] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[11] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proc. IEEE ICASSP*, 2008.

[12] Anurag Kumar and Bhiksha Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*, New York, NY, USA, 2016, MM '16, pp. 1038–1047, ACM.

[13] B. Lakshminarayanan, R. Raich, and X. Fern, "A syllable-level probabilistic framework for bird species identification," in *ICMLA*. 2009, IEEE.

[14] G. Fodor, "The Ninth Annual MLSP competition: First place," *International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–2, 2013.

[15] Mario Lasseck, "Towards automatic large-scale identification of birds in audio recordings," in *6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, 2015, pp. 364–375.

[16] Ilyas Potamitis, "Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity," *Ecological Informatics*, vol. 26, no. 3, pp. 6–17, 2015.

[17] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing (3rd Edition)*, Prentice-Hall, Inc., 2006.