A comparative study on filtering and classification of bird songs

Nicolas Figueiredo, Felipe Felix
University of São Paulo
nsf,fsfelix@ime.usp.br

Carolina Brum Medeiros
fliprl
carolina.medeiros@mail.mcgill.ca

Marcelo Queiroz
University of São Paulo
mqz@ime.usp.br

ABSTRACT

This paper presents a combination of signal processing and machine learning techniques for classification of bird song recordings. Our pipeline consists of filters to enhance the bird song signal with respect to environmental noise, followed by machine learning algorithms that exploits various acoustic features. The filtering stage is based on the assumptions that bird songs are tonal and sporadic, and that noise, present along the entire recording, has large bandwidth. We present and discuss the results of an experiment on a dataset containing recordings of bird songs from species in the Southern Atlantic Coast of South America. This experiment compares the use of several acoustic features (RMD, ZCR, MFCC, spectral centroid/bandwidth/rolloff and syllable duration), extracted from pre-filtered recordings using three proposed filters, combined with traditional classification strategies (KNN, NB and SVM), in order to identify useful filter/feature/classifier combinations for this bird song classification task. This strategy produces improved classification results with respect to those reported in a previous study using the same dataset.

1. INTRODUCTION

Bird song classification is an important task for ornithologists and biologists in general. Due to their relative easy detection, responsiveness to change, and relationships with lower trophic levels, birds have been widely used as indicators of biodiversity trends [1]. Their vocalizations have been used to monitor the abundance and composition of bird communities in several different habitats [2] [3]. However, manual classification of songs by ornithologists can be an expensive field work task. Recently, biologists have introduced the use of autonomous recording units (ARU's) for collection of environmental audio recordings, but the quantity of data generated by these devices makes manual inspection prohibitive [4]. Automation of this process is hindered by the acoustic diversity of bird songs, quality of recordings, noisy environments, and simultaneity of vocalizations of different species. These challenges and possibilities motivate the development of an automatic bird song classification system. The classification could then control the ARUs recording or guide the segmentation of continuous recordings, discriminating quiet frames or sounds that do not belong to a vocabulary.

Bird vocalizations are divided into calls and songs. In general, songs are spontaneous vocalizations produced by males in the breeding season and tend to be longer and more complex than calls. Calls are usually related to specific functions such as flight or threat. They are acoustically simpler and shorter, and produced by both sexes throughout the year. It is important to note that these are general definitions with plenty of exceptions to every characteristic presented [5]. The present work focuses solely on songs.

Each different bird song can be divided in the following descending hierarchical levels: phrases, syllables and elements. Elements can be defined as the inseparable components (straight lines, for example) of a song's spectrogram. Elements are the building-blocks of syllables, which can be composed of one or more elements. Finally, a phrase is defined as a group of syllables, and a song is composed of a group of phrases [5]. Ornithologists extract the durations of these structures and employ them as metrics to distinguish different bird songs, leading to a manual classification strategy which is very time-consuming [6].

Most automatic bird song classification systems can be divided in four stages: pre-processing, segmentation, feature extraction and classification. In the pre-processing stage, recordings are filtered as to enhance the bird songs present. Segmentation can then be applied to slice a song into its syllables or phrases. From individual syllables or full-length recordings some acoustic features such as MFCCs are then extracted and used to classify the song as pertaining to a certain species.

Lopes et al. [7] developed a bird song classification system based on SVM and Naive-Bayes classifiers and standard acoustic features such as MFCC, spectral centroid and spectral rolloff. The study presented here extends their work by considering other features, including melodic features that are considered state-of-the-art in the type of manual classification still performed by ornithologists [5, 6], and by developing a more robust pre-processing stage, developing specific filtering strategies that enhance the bird song and thus ease the classification task. We use the same dataset and the same methodology of [7], in order to provide a clear comparison between our results and theirs.

The design of our study, including the choice of a particular methodology and dataset, is not meant to imply that other approaches to bird song classification are not acknowledged. To cite a few, other classification strategies used in bird song classification include Hidden Markov

Models [4] and transfer learning from music genres [8], and other features based on wavelet decomposition have been used in the classification of inharmonic and transient bird songs [9]. Other recent approaches employed deep learning for the detection of bird songs using binary masks [10], and convolutional neural networks to classify spectrogram segments [11].

One of our main goals is to develop strategies that establish a clear dialogue with the manual classification techniques employed by ornithologists for the differentiation of bird species. Because of that, we focused on acoustic features with a clear physical interpretation and relationship to human hearing, and classification strategies which are explainable in terms of the perceptual representations employed, so that the whole process can be accompanied and iteratively improved in collaboration with ornithologists, and benefit from their expertise. Another motivation for preferring less computationally intensive feature extraction and classification schemes is the perspective of embedding them in ARUs as previously mentioned.

Regarding the pre-processing stage, several techniques for filtering environmental recordings have also been explored in the context of bird song recognition. Due to the diversity in types of bird songs, some techniques take advantage of specific traits of the bird songs being analyzed, such as their periodicity or frequency range. In [12], the periodicity of Antbird chirps is leveraged in the development of a correlation-maximization denoising filter that enhances the target call while suppressing other bird calls that don't follow the same structure. Lasseck [13] presents a filtering method based on image processing of grey-scale spectrograms: a binary mask is built where each pixel exceeding 3 times the median value of its corresponding row (frequency band) and 3 times the median of its corresponding column (time frame) in the spectrogram is set to 1, while the remaining pixels are set to 0. Another example of a generalistic approach is shown in [14], where a whitening filter utilizes low energy spectrogram frames to estimate the frequency profile of noise and attenuate each row of the spectrogram accordingly.

In this work, we present novel filtering techniques for the enhancement of bird song recordings with a prominent tonal quality, followed by a comparative classification study using several individual acoustic features and classifiers. Our goal is to identify useful combinations of filters, acoustic features and classifiers that produce best classification results for a dataset of tropical birds from the Southern Atlantic Coast of South America, created, labeled, and shared by Lopes et al. [7].

The paper is organized as follows: section 2 introduces the developed pre-processing filtering techniques based on spectral contrast and temporal variance masks. Section 3 presents the features and classifiers used, including an energy-based segmentation algorithm. Section 4 shows the obtained classification F-scores for the different features, classifiers and filtering techniques of our system. The results are discussed in the closing section 5.

2. FILTERING STRATEGIES FOR BIRD SONGS

2.1 Bird songs and environmental noise

As recordings are registered in the presence of environmental noise, a strategy for increasing the signal-to-noise ratio is needed. In order to do so, we first need to properly characterize our signal of interest and the noise usually present in such recordings. Bird songs are difficult to characterize because they can assume very different spectral characteristics, from melodic sequences of notes to periodic repetition of broadband noise-like chirps or screeches. Most bird songs are classified as tonal, consisting of a single fundamental frequency or several harmonically or non-harmonically related frequencies [15].

In this context, we define noise as anything but the signal of interest: a bird song. Environmental recordings often present insect sounds, rain, sounds from microphone manipulation, wind, and other animal utterances. The most common type of noise present in the analyzed recordings were insect sounds that were persistent throughout most of the recordings' duration and that occupy a broad frequency range.

The proposed filter intends to be as flexible as possible, while noting the difficulty being flexible because of the great variety in bird songs and types of noise existent in the environment. Because of the considerations above, we limit the development of our filter to a scenario where the following assumptions hold, regarding the signal of interest and the types of environmental noise present:

- Bird songs are tonal;
- Noise usually has a large frequency bandwidth;
- Bird songs are sporadic (they appear intermittently throughout a recording);
- Noise is usually present throughout a recording.

In the sequel the technical details of the filtering strategies will be presented.

2.2 The spectral mask approach

Mask filtering has been used in source separation problems such as single-channel speech separation [16], source separation in reverberant two-channel recordings [17] and speech music separation [18].

In our approach, the recording to be filtered is analyzed and a soft mask represented by a matrix of coefficients with values between 0 and 1 is composed. This matrix is developed so as to have the same dimensions as the spectrogram of the original recording. Filtering is then performed as an element-wise multiplication of both matrices. Such mask has to assume values close to 1 for bins containing bird songs and values close to 0 for the rest of the STFT bins. The filtering stage is concluded with the ISTFT of the resulting matrix.

The filtering mask can be thought of as a matrix of confidence scores for the presence of a bird song in each of the STFT bins. In the approach presented here, the final

filtering mask is composed of two different masks that intend to measure confidence scores for the presence of bird songs along the frequency and time axis separately. We then present and evaluate three different ways to combine these two partial masks.

2.3 Spectral contrast mask

Octave-based spectral contrast is an audio feature presented in [19] designed to capture the relative distribution of the harmonic and non-harmonic components of a signal. This feature is a descriptor of a spectrum's variability inside each octave, and its extraction is implemented in the libROSA Python library [20] as follows: a spectrogram is received as input, and its frequency axis is split into octaves. For each time frame, the linear difference between the spectrogram's energy peak and valley inside each octave is computed. This results in an S x T matrix, where S is the (user-defined) number of octaves and T is the number of time frames of the original spectrogram.

Higher values in the spectral contrast matrix are thus associated to octaves which contain strong spectral peaks, whereas lower values represent octaves with relatively constant energy within its frequency range. Based on our assumption that bird songs are tonal, they would appear in the spectrogram as clear and isolate spectral peaks, while noise would occupy a large bandwidth, and so the spectral contrast of octaves containing bird songs will be higher than those containing noise. Fig. 1 shows the spectral contrast matrix for a snippet of a recording taken from [7]. This snippet contains a *Trogon surrucura* song concentrated around 1250 Hz and (undesired) cicada vocalizations that span from about 2500 Hz to 15000 Hz.

In order to transform the spectral contrast matrix into a confidence mask, we first expand the matrix delivered by libROSA to the same dimensions of the original spectrogram, replicating every octave row for the appropriate number of frequency bins. After that, we normalize this M matrix producing a soft mask $\bar{M}$ such that

$$\bar{M}_{ij} = \frac{M_{ij} - min(M)}{max(M) - min(M)}, \forall i, j.$$

This way, areas that have a higher spectral contrast (assumed to be areas containing a bird song) will have $\bar{M}_{ij}$ values close to 1, and areas with lower spectral contrast (presumably noise) will have $\bar{M}_{ij}$ values close to 0.

2.4 Temporal variance mask

The second confidence mask developed explores the difference between the assumed temporal distributions of noise and signal. This difference can be illustrated in the following example: in a typical 10-second excerpt of a bird song recording, we would expect the bird to chirp intermittently between 2 and 4 times, while the environmental noise will remain unaltered during the whole excerpt. In order to explore this difference in the statistical distributions of signal and noise, we analyze the energy variance over time for each frequency band given by the STFT spectrogram. Bands containing a bird song will present large

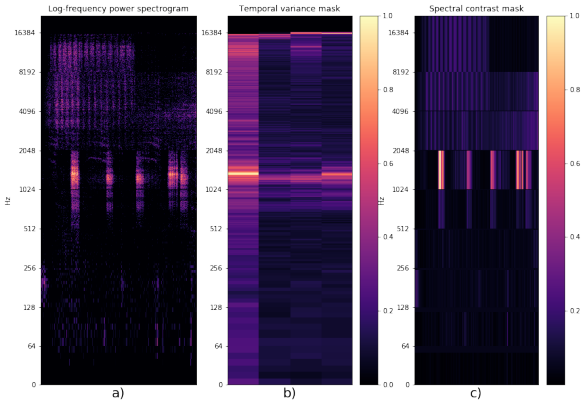

Figure 1: a) The power spectrogram of a recording taken from [7] containing a *Trogon surrucura* song and cicada vocalizations. b) The composed temporal variance mask. c) The composed spectral contrast mask.

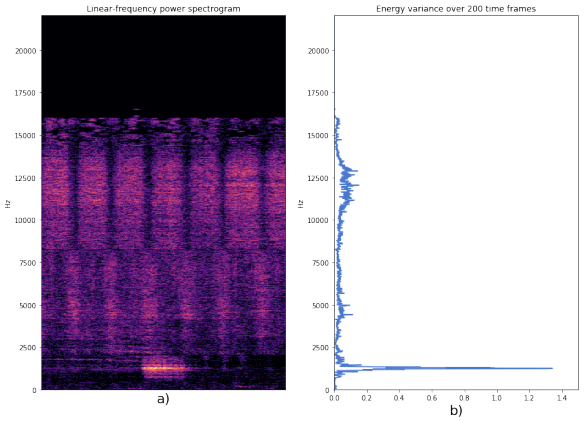

Figure 2: Temporal variance mask: a) Power spectrogram of a snippet of a recording taken from [7] containing a *Trogon surrucura* song and cicada vocalizations. b) Plotted standard deviation for every frequency line given by the STFT.

variations in energy between song frames and background noise frames, whereas frequency bands dominated by environmental noise would not display such large variations. Figure 2 shows this variation as a function of frequency in a 200-frame snippet of the same recording shown in Figure 1.

The following procedure is used to transform this analysis into a confidence mask: for each frequency band given by the STFT, the energy variance is computed for snippets of L consecutive time frames, resulting in an F x 1 vector where F is the number of rows of the original spectrogram. This vector is used as the confidence score for all L columns of the corresponding snippet. This process is repeated for all snippets of L frames until we reach the end of the spectrogram, so that we have an energy variance matrix with the same dimensions as the original spectrogram matrix. Finally, we normalize this matrix in the same way as the spectral contrast mask. This way, high energy variance areas (assumed to be areas containing a bird song)

will have values close to 1. Figure 1 shows the temporal variance mask for the presented recording snippet.

The choice of the default value for L=200 takes into account the STFT parameters used with libROSA, i.e. sampling rate = 44100 Hz, and 4096-sample FFTs with 512-sample hops, so that the L consecutive frames span around 2.32 seconds. This duration was chosen based on the average note duration in bird songs of the species contained in the dataset [7], to correspond to roughly two to three times the duration of the average note, so that whenever a note is present some silence is also included in the snippet where the energy variance is computed, increasing the variance for snippets containing the bird song.

2.5 Composition of the final filtering mask

Finally, the overall filtering mask is composed from the temporal variance and spectral contrast masks, using three different approaches:

- Filter A: the filtering mask is computed by a simple element-wise multiplication with equal weights given to the two partial masks;
- Filter B: the original spectrogram is pre-filtered with the variance mask via element-wise multiplication. Then, the spectral contrast mask of this pre-filtered spectrogram is computed. Finally, the spectral contrast mask is used to further filter the signal.
- Filter C: we first set to 1 all bins pertaining to the 95th percentile of each mask, resulting in two binary masks. We then compose the final mask as the intersection (logical AND) between both masks. Thus, Filter C corresponds to a binary filtering mask.

3. REPRESENTATION AND CLASSIFICATION OF BIRD SONGS

In an audio classification system, after the filtering stage we need to generate relevant parameters for the classification algorithm. In this work, we employed well-known acoustic features and a custom feature to capture a domain-specific parameter: the syllable duration of a bird song. We now present each feature used and an interpretation of which characteristics of a bird song they may capture.

- Mel Frequency Cepstral Coefficients (MFCC): MFCCs capture properties related to a signal's timbre [21]. We can think of it as a general picture or fingerprint of the characteristics of a bird song's spectrogram. Because of that, it is a promising differentiator of bird songs.
- Spectral centroid: usually, spectral centroids are related to the perception of "brightness" of a given sound, and so it is a potentially good distinguisher of bird songs that span different spectral regions.
- Spectral bandwidth: it captures the variability of frequencies in a given spectrogram. Although not a noisiness measure itself, it allows the differentiation of broadband noise-types from simple melodic bird songs, which typically have few prominent overtones.
- Spectral roll-off: measures the rate of spectral energy decay, serving as a complementary feature to both spectral centroid and bandwidth.
- Zero-crossing rate (ZCR): for tonal signals, ZCR is a pitch-related measure which, like spectral centroid, may be used to distinguish bird songs with different fundamental frequencies. Noisy signals tend to display a much higher ZCR than tonal signals, which is also useful to distinguish bird songs from environmental noise.
- Root mean square (RMS): RMS captures the mean energy of a signal over a given time window. Although not a reliable distinguisher for bird species (since energy also reflects the relative position of the microphone), this feature is useful for detection of bird presence and corresponding segmentation of audio signals.
- Syllable duration: it is a direct temporal measure of constituting elements of a bird song, frequently used by ornithologists in manual classification of recordings. We developed a method to extract this feature for general bird song recordings, based on the `auto_detec` function available in the `warbleR` library [22]. The algorithm is implemented as follows: first the mean RMS energy σ over the entire duration of the signal is obtained; then each frame n is selected as part of the bird song according to the conditions RMS[n]¿σ and RMS[$n+1$]¿σ, producing a binary frame-mask $\phi[n]$; next, we apply a morphological dilation operator to ϕ to link together close but separated bird song frames; finally, we estimate the duration of each syllable from contiguous frames with $\phi[n] = 1$.

It is common in ornithology datasets to associate each recording to a particular species; bird presence annotations on a frame-by-frame basis are usually not available. This prompted us to generate global features to represent each recording, by applying statistical operations (mean, standard deviation, maximum and minimum) to summarize local (frame-based) features. Classification is then performed using global feature vectors that describe each one of the above individual features by its statistical summary.

In order to increase both the computational efficiency and the classification performance of the classification methods used, we propose an energy-based frame selection strategy that aims to eliminate from the representation frames that are too weak to contain useful bird-related information. This frame selection strategy simply rejects frames with energy lower than the mean energy of each recording. Global features obtained by statistical summaries are thus computed by considering only selected frames. This simple strategy yielded better results in all experiments.

We considered the following algorithms for the classification stage: k-Nearest Neighbors (kNN), Naive-Bayes

(NB), and Support Vector Machines (SVM). These are well-known and widely used strategies for machine learning in general and for classification of bird songs in particular [7, 23]

4. EXPERIMENTS AND DISCUSSION

The dataset used in the experiments consists of bird songs of species in the Southern Atlantic Coast of South America. This dataset was created and shared by Lopes et al. [7], and contains 1631 song recordings and 674 calls of 77 species. Some species have a small number of recordings in the dataset, so in our experiments we selected the n species with the highest number of recordings, for several n values. Experiments were conducted for every combination of the following variables:

- filtering techniques: filters A, B, and C;
- local features: RMS, MFCC, spectral centroid, spectral bandwidth, spectral roll-off, zero-crossing rate and syllable duration;
- global features: mean, std, min and max of all local features;
- number of species: 3, 5, 8, 12, 20;
- classifiers: kNN, NB, SVM.

The results reported below were obtained with the following parameters: 5-fold cross-validation; k = 3 for the kNN algorithm; linear kernel for SVM; Gaussian distribution for NB; p = 2 for the spectral bandwidth and $p = 85\%$ for the spectral roll-off.

In order to demonstrate the effect of the filters in the classification, Figures 3 through 6 depict the variability on performance of various pairs of features and filtering stage, for a dataset containing 5 species. Some of the features are significantly affected by the usage of filtering as preprocessing technique, prior to training, while others are less sensitive to it.

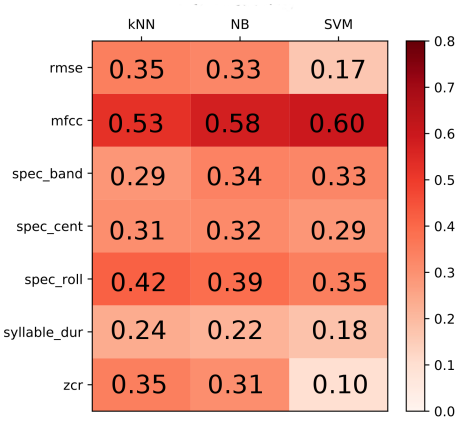

Figure 3: F-scores for each classifier/feature pair tested and 5 species, using non-filtered recordings.

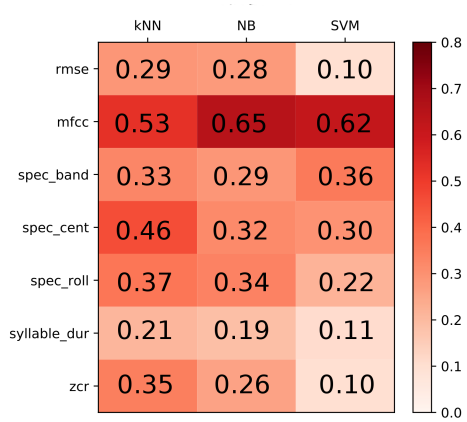

Figure 4: F-scores for each classifier/feature pair tested and 5 species, using recordings processed with filter A.

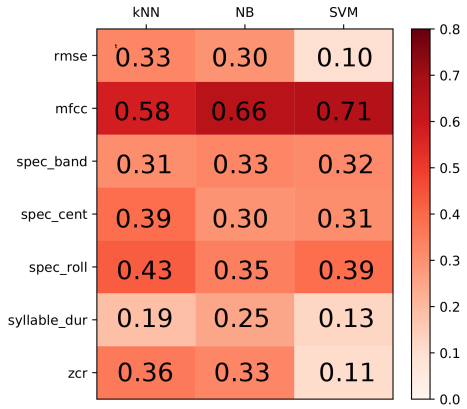

Figure 5: F-scores for each classifier/feature pair tested and 5 species, using recordings processed with filter B.

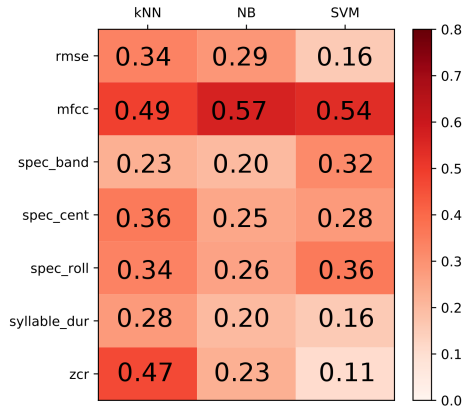

Figure 6: F-scores for each classifier/feature pair tested and 5 species, using recordings processed with filter C.

Figures 7 through 10 present in a compact way the F-scores for all pairs of individual features used and classifiers. In order to save space only the filter options that achieved the highest results have been shown, which in all cases correspond to Filter B. Results with Filter A have in

general reached close but slightly lower highest values, and Filter C have performed poorly in general.

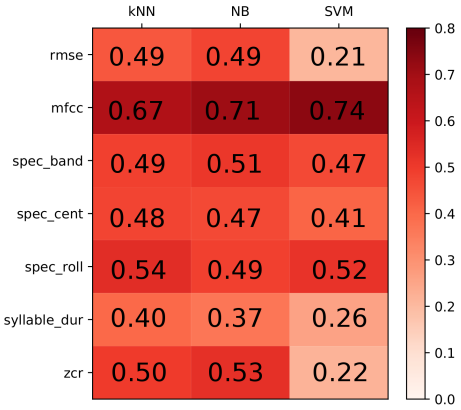

Figure 7: F-scores for each classifier/feature pair tested and 3 species.

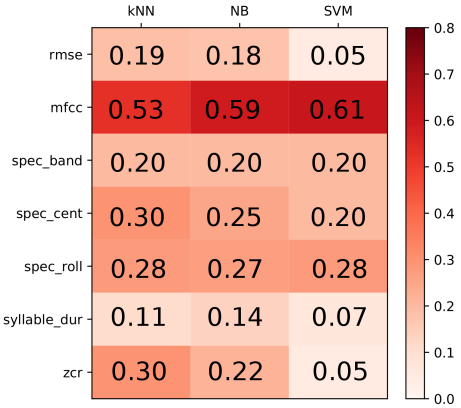

Figure 8: F-scores for each classifier/feature pair tested and 8 species.

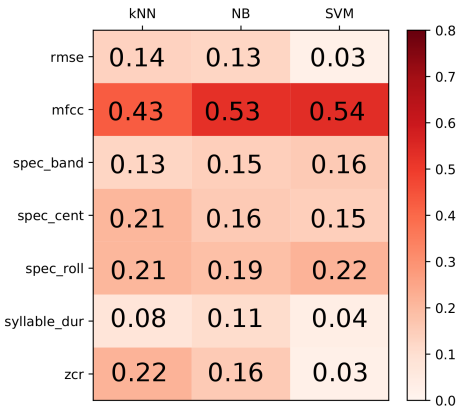

Figure 9: F-scores for each classifier/feature pair tested and 12 species.

Figure 10: F-scores for each classifier/feature pair tested and 20 species.

These results clearly point to MFCC being the best feature for this application, with all classifiers, and its advantage over the other features increases with the number of species being classified. This possibly indicates that the bird's timbre is an acoustic feature that distinguishes them well. On the other hand, the worst results were obtained by the syllable duration feature. We have two hypotheses for this: the syllable duration is not a good discriminator of bird songs of the specific species that were selected in this experiment, and/or our method is not robust enough, in the sense that the method employs heuristics that reflect certain assumptions about the song structure, which may not hold for this data.

Regarding the other tested features, it is hard to say which classifier works best in this context because of the generally small variance of the F-measure results. Classifier performance was highly dependent on the feature being used. As an example, spectral bandwidth and centroid paired best with k-NN while MFCC paired best with SVM. Even so, this best pairing sometimes changed with the number of species being classified and the filter used as the pre-processing stage. The only clear and consistent results were the poor pairing of SVM with RMS, syllable duration and zero-crossing rate. Because of its good pairing with MFCC, SVM obtained the best F-measure results amongst those tested. In fact, this classifier/feature pairing used along with Filter B as the pre-processing algorithm attained the best F-measure for every n species number tested.

Filter performance was also highly dependent on the acoustic feature being used. Classification based on spectral centroids and spectral bandwidth was highest when paired with Filter A, while zero-crossing rate and syllable duration paired best with Filter C. Root mean square actually showed its best results when applied to the unfiltered recordings, and the remaining features paired best with Filter B. The most notable F-measure gains attributed solely to filtering for classification amongst 5 species were: 8% for MFCC and NB (filter A), 9% for zero-crossing rate and k-NN (filter C), and 10% for MFCC and NB (filter B). One interesting result is that, while using MFCC as the descrip-

tor, Filter C actually performed worst than the unfiltered recordings. This is the most "agressive" filter tested, in that, being a binary mask, it will completely filter out most of the recording's information. A hypothesis for its poor pairing with MFCC is that in most cases it will filter out important components of a bird song timbre.

For each n tested, we report below the names of the species included, the best classifier/feature pair for unfiltered recordings and also the best classifier/feature/filter combination:

- $n = 3$ (species classified: *Gnorimopsar chopi, Sittasomus griseicapillus, Lathrotriccus euleri*)

  Best unfiltered result: SVM and MFCC. F-measure = 0.71 (+/- 0.11)

  Best filtered result: SVM, MFCC and Filter B. F-measure = 0.74 (+/- 0.14)

- $n = 5$ (species classified: *Pseudoleistes guirahuro, Saltator similis, Gnorimopsar chopi, Sittasomus griseicapillus, Lathrotriccus euleri*)

  Best unfiltered result: SVM and MFCC. F-measure = 0.62 (+/- 0.09)

  Best filtered result: SVM, MFCC and Filter B. F-measure = 0.67 (+/- 0.14)

- $n = 8$ (species classified: *Chiroxiphia caudata, Dysithamnus mentalis, Mimus saturninus, Pseudoleistes guirahuro, Saltator similis, Gnorimopsar chopi, Sittasomus griseicapillus, Lathrotriccus euleri*)

  Best unfiltered result: SVM and MFCC. F-measure = 0.57 (+/- 0.10)

  Best filtered result: SVM, MFCC and Filter B. F-measure = 0.59 (+/- 0.08)

- $n = 12$ (species classified: *Xiphorhynchus fuscus, Vanellus chilensis, Batara cinerea, Camptostoma obsoletum, Chiroxiphia caudata, Dysithamnus mentalis, Mimus saturninus, Pseudoleistes guirahuro, Saltator similis, Gnorimopsar chopi, Sittasomus griseicapillus, Lathrotriccus euleri*)

  Best unfiltered result: SVM and MFCC. F-measure = 0.48 (+/- 0.08)

  Best filtered result: SVM, MFCC and Filter B. F-measure = 0.54 (+/- 0.06)

- $n = 20$ (species classified: *Certhiaxis cinnamomeus, Leucochloris albicollis, Thamnophilus ruficapillus, Phleocryptes melanops, Piprites chloris, Myiophobus fasciatus, Poospiza nigrorufa, Pyriglena leucoptera, Xiphorhynchus fuscus, Vanellus chilensis, Batara cinerea, Camptostoma obsoletum, Chiroxiphia caudata, Dysithamnus mentalis, Mimus saturninus, Pseudoleistes guirahuro, Saltator similis, Gnorimopsar chopi, Sittasomus griseicapillus, Lathrotriccus euleri*)

  Best unfiltered result: SVM and MFCC. F-measure = 0.47 (+/- 0.04)

  Best filtered result: NB, MFCC and Filter B. F-measure = 0.53 (+/- 0.04)

No. of classes	Lopes (2011)	Figueiredo et al. (2018)
3	0.73	0.74
5	0.73	0.66
8	0.57	0.61
12	0.48	0.54
20	0.47	0.54

Table 1: Comparison between F-measures obtained in our experiment and [7]

Table 1 compares our best results with those in [7]. We can see an improvement in most results, especially for higher numbers of species.

5. CONCLUSIONS

This work dealt with the problem of bird species classification, approaching all stages of the classification pipeline, from pre-filtering through feature extraction to automatic classification. We obtained improved results with respect to those previously obtained by Lopes et al. [7] by combining new filtering techniques with standard acoustic features and classifiers. Our experiments indicate that a good feature/classifier pair for this problem is MFCC and SVM. The classification improvements obtained by the proposed filters are indicative of the validity of the spectral mask approach in this context, and of our initial assumptions about the nature of bird song signals and environmental noise. Further exploration and refining of these assumptions are an encouraging route for the development of better filtering techniques, leading to improved classification strategies.

We have shown that performance in bird song classification is highly variable according to number of classes, acoustic features, classification algorithms, and pre-processing filters. Using filters as enhancers for the underlying signals proved to be of vital importance in improving performance of the overall classification task. The results of our experiment suggest that there are many possible couplings between filtering technique and type of feature used, the quality of which may depend both on the quality of the recordings and on the specific species and types of environmental noise they contain.

Future work should address other domain-specific features that have been used consistently by ornithologists in manual bird song classification. It came as a surprise that syllable duration performed so poorly on the available data, but this may have several reasons unrelated to the importance of this particular feature in manual classification. In order to better understand and possibly overcome this difficulty, it would be useful to work with manually annotated domain-specific features and develop filters that preserve related acoustic characteristics. A promising avenue for exploration is the use of detection techniques such as CFAR for a segmentation phase following the filtering phase. This would facilitate the extraction of many domain-specific temporal features (such as durations of notes, syllables and phrases), while ensuring that only the relevant parts of the recordings are used. Lastly, employing a multi-feature approach would be a great step for-

ward in a more robust classification system.

Acknowledgments

Authors would like to thank Vitor Piacentini for his expert support on ornithological matters. The first author acknowledges the support from a scholarship by CAPES, and the fourth author acknowledges the support of FAPESP grant 2018/09373-8 and CNPq grant 309645/2016-6.

6. REFERENCES

[1] Ç. H. Şekercioğlu, G. C. Daily, and P. R. Ehrlich, "Ecosystem consequences of bird declines," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 18 042–18 047, 2004.

[2] N. T. Boelman, G. P. Asner *et al.*, "Multi-trophic invasion resistance in hawaii: bioacoustics, field surveys, and airborne remote sensing," *Ecological Applications*, vol. 17, no. 8, pp. 2137–2144, 2007.

[3] M. Depraetere, S. Pavoine *et al.*, "Monitoring animal diversity using acoustic indices: implementation in a temperate woodland," *Ecological Indicators*, vol. 13, no. 1, pp. 46–54, 2012.

[4] I. Potamitis, S. Ntalampiras *et al.*, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1–9, 2014.

[5] C. K. Catchpole and P. J. Slater, *Bird song: biological themes and variations*. Cambridge university press, 2003.

[6] L. S. Carneiro, L. P. Gonzaga *et al.*, "Systematic revision of the spotted antpitta (grallariidae: Hylopezus macularius), with description of a cryptic new species from brazilian amazonia," *The Auk*, vol. 129, no. 2, pp. 338–351, 2012.

[7] M. T. Lopes, L. L. Gioppo *et al.*, "Automatic bird species identification for large number of species," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 117–122.

[8] S. Ntalampiras, "Bird species identification via transfer learning from music genres," *Ecological Informatics*, vol. 44, pp. 76–81, 2018.

[9] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in recognition of bird sounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 051806, 2006.

[10] I. Potamitis, "Deep learning for detection of bird vocalisations," *arXiv preprint arXiv:1609.08408*, 2016.

[11] B. P. Tóth and B. Czeba, "Convolutional neural networks for large-scale bird song classification in noisy environment." in *CLEF (Working Notes)*, 2016, pp. 560–568.

[12] W. Chu and A. Alwan, "A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[13] M. Lasseck, "Bird song classification in field recordings: winning solution for nips4b 2013 competition," in *Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada*, 2013, pp. 176–181.

[14] F. Briggs, B. Lakshminarayanan *et al.*, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[15] N. Fletcher, "A class of chaotic bird calls?" *The Journal of the Acoustical Society of America*, vol. 108, no. 2, pp. 821–826, 2000.

[16] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.

[17] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[18] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Digital Signal Processing (DSP), 2011 17th International Conference on*. IEEE, 2011, pp. 1–6.

[19] D.-N. Jiang, L. Lu *et al.*, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 113–116.

[20] B. McFee, M. McVicar *et al.*, "librosa 0.5.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[21] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[22] M. Araya-Salas and G. Smith-Vidaurre, "warbler: An r package to streamline analysis of animal acoustic signals," *Methods in Ecology and Evolution*, vol. 8, no. 2, pp. 184–191, 2017.

[23] S. Fagerlund, "Automatic recognition of bird species by their sounds," *Finlandia: Helsinki University Of Technology*, 2004.