Bag of MFCC-based words for bird identification

Julien Ricard and Hervé Glotin

LSIS/DYNI
University of Toulon, France
http://www.lsis.org/dyni

Abstract. The algorithm used by the authors in the bird identification task of LifeCLEF 2016 consists in creating a dictionary of MFCC-based words using k-means clustering, computing histograms of these words over short audio segments and feeding them to a random forest classifier. The official score achieved is 0.15 MAP.

Keywords: bird identification, MFCC, k-means, bag-of-words, random forest

1 Foreword

The algorithm presented here is quite standard and was initially used on smaller datasets to improve, in a late fusion scheme, a classifier based on pairs of spectrogram peaks, described in the context of audio fingerprinting in [1]. Because of some memory issues we have not been able to run this latter algorithm on the challenge dataset. We propose in the discussion some options to possibly overcome this issue.

2 Algorithm

The method is based on the *bag-of-words* approach, initially used in text analysis to model long-term distribution of words [2] and more recently in audio signal analysis for tasks including spoken language identification [3] or urban soundscape identification [4]. The different steps of the algorithm are:

1. The original 44.1 kHz audio files were split in 0.2s segments with 50% overlap.
2. Only the segments having energy values higher than a relative (to the whole audio file) value and spectral flatness values smaller than an absolute threshold were kept. This method assumes that a segment containing bird vocalizations is *voiced* (i.e. is made of stable time-frequency components, or partials) and has high energy compared to the environmental noise. Even though not all bird vocalizations are voiced, this simple technique has proven to give high precision (close to 1) in a bird presence/absence classification[1].

[1] The recall was not great (about 0.5), but we were more interested in making sure the detected segments actually contained bird vocalizations than in detecting all these segments.

3. The Mel-Frequency Cepstral Coefficients (MFCC) were computed, with the following parameters:
   - analysis window size: 11.6ms
   - analysis window overlap: 50%
   - min frequency: 0Hz
   - max frequency: 22050Hz
   - number of mel bands: 32
   - number of MFCC: 15 (the first coefficient, related to the energy, was removed)
4. A k-means clustering was performed on all the MFCC and their derivatives, with k=500.
5. For every files the normalized histogram of MFCC-based words (i.e. the 500 clusters) was computed (using only segments kept in step 2).
6. The resulting feature vectors were then fed to a random forest classifier with the following parameters:
   - number of trees: 400
   - minimum number of samples required to split an internal node: 10

The algorithm was implemented in Python, using Numpy, PySoundFile[2], librosa[3] and scikit-learn[4].

3 Results

Our method achieved the following results (MAP):

- with background species: 0.149
- only species: 0.183
- soundscape: 0.037

The full list of results is given in http://www.imageclef.org/lifeclef/2016/bird.

4 Discussion

The proposed method is quite naive and achieve low performance compared to the other participants.

The implementation of the random forest algorithm used[5] required to be fed with the whole training dataset, which, because of the large amount of data, lead to memory issues. We had to limit the number of trees and branches, which probably decreased the performance of the models. An online (mini-batch) implementation, such as the ones proposed in [5] or [6] could be used to overcome this problem.

[2] https://github.com/bastibe/PySoundFile

[3] https://github.com/librosa/librosa

[4] http://scikit-learn.org

[5] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

As mentioned in the foreword, the predictions were supposed to be fused with those of another algorithm. Some preliminary tests on the training set (splitting the training set in a new 70% training/30% test set) showed that this fusion improved the performance from 0.13 to 0.2 MAP, using the same random forest parameters as shown earlier. Combining this with a mini-batch implementation of random forest to use a larger number of trees and larger trees would possibly have improved further the performance.

We believe that the weaknesses of our method lie in two main aspects. First, apart from the MFCC derivatives, which describe only very short-term changes, we did not encode any temporal information. It has been shown in [7] that extracting some signal modulations helped in automatic bird identification, and we assume [6] that the analysis of acoustic sequences, as described for instance in [8], might also help. Secondly, while we focused here on the voiced parts of the signal, some bird vocalizations are unvoiced and should also be taken into account, by adding some features that could describe them and modifying our detection of the segments of interest accordingly.

References

1. Wang, A.: An Industrial Strength Audio Search Algorithm. In: ISMIR, pp. 7–13 (2003)
2. Sebastiani, F.. Machine learning in automated text categorization. In: ACM Computing Surveys 34, pp 1-47 (2002).
3. Ma, B. and Haizhou L. : Spoken language identification using bag-of-sounds (2005).
4. Aucouturier, J.-J., Defreville, B. and Pachet F.: The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. In: Journal of the Acoustical Society of America 122.2, pp 881–891 (2007).
5. Saffari, A. et al.: Online Random Forests. In 3rd IEEE ICCV Workshop on On-line Computer Vision (2009). Code: https://github.com/amirsaffari/online-multiclass-lpboost
6. Lakshminarayanan, B. et al.: Mondrian forests: Efficient online random forests. In Advances in Neural Information Processing Systems (2014). Code: https://github.com/balajiln/mondrianforest
7. Stowell, D. and Plumbley, M. D.: Largescale analysis of frequency modulation in birdsong data bases. In Methods in Ecology and Evolution 5.9, pp. 901–912 (2014).
8. Kershenbaum, A. et al.: Acoustic sequences in nonhuman animals: a tutorial review and prospectus. In Biological Reviews (2014).
9. Katahira, K. et al.: Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. In PloS one 6.9, e24516 (2011).

[6] We have not yet run any experiment to prove that sequence analysis helped in automatic bird identification, and we have not found any studies showing it. However, it has been shown that some acoustic sequences in bird songs can be explained by simple hidden Markov processes [9], and while it does not prove that it can help in our task, it proves that some sequencing rules exist, and we assume that they might be helpful.