Clusterized Mel Filter cepstral coefficients and Support Vector Machines for bird song idenfication

Olivier Dufour OLIVIERLOUIS.DUFOUR@GMAIL.COM
LSIS, Université du Sud Toulon Var

Thierry Artieres THIERRY.ARTIERES@LIP6.FR
LIP6, Université Paris 6

Hervé GLOTIN GLOTIN@UNIV-TLN.FR
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
Institut Universitaire de France, Bd St Michel, 75005 Paris

Pascale Giraudet GIRAUDET@UNIV-TLN.FR
Université du Sud Toulon Var

1. Introduction

We present here our contribution to the "Machine Learning for Bioacoustics" workshop technical challenge of 30th International Conference on Machine Learning (ICML 2013). The aim is to build a classifier able to recognize bird species one can hear from a recording in the wild. The method we present here is a rather simple strategy for bird songs and calls classification. It builds on known and efficient technologies and ideas and must be considered as a baseline on this challenge[1]. The method we present is dedicated to the particular setting of the challenge. It relies in particular on the fact that training signals are monolabel, i.e. only one species may be heard, while test signals are multilabel.

2. Description of the method

We present now the main steps of our approach. Figure 1 illustrates the main steps of the preprocessing and of feature extraction. We consider we want to learn a multilabel classifier from a set of N monolabeled training samples $\left\{(x^i, y^i)|i = 1..N\right\}$ where each input x^i is a audio recording and each y^i is a bird species $\forall i, y^i \in \{b_u|u = 1..K\}$ (in our case there are 35 species, $K = 35$). The system should be able to infer the eventually multiple classes (presence of bird species) in a test recording x.

[1] As we are also co-organizing this challenge, our participation aimed at defining a baseline system, with raw features, that all other participants could compare too. We did not look for optimizing each parameter of our system, and as any other participant, we conducted all the modeling and experimentation applying strictly the rules of the challenge.

2.1. Preprocessing

Our preprocessing is based on mfcc cepstral coefficients which have been proved useful for speech recognition (Chang-Hsing et al., 2006; Michael Noll, 1964). A signal is first transformed into a series of frames where each frame consists in 17 mfcc (mel cepstra feature coefficients) feature vectors, including energy. Each frame represents a short duration (e.g. 512 samples of a signal sampled at 44kHz).

2.2. Windowing, silence removal and feature extraction

Windowing. We use windowing, i.e. computing a new feature vector on a window of n frames, to get new feature vectors that are representative of longer segments. The idea is close to the standard syllabe extraction step that is used in most of methods for bird identification (Neal et al., 2011; Briggs et al., 2012; 2009) but is much simpler to implement. In our case we considered segments of about 0.5 second duration (i.e. $n \approx$ few hundreds of frames) and used a sliding window with overlap (about 80%).

Silence removal. We first want to remove segments (windows) corresponding to silence since these would

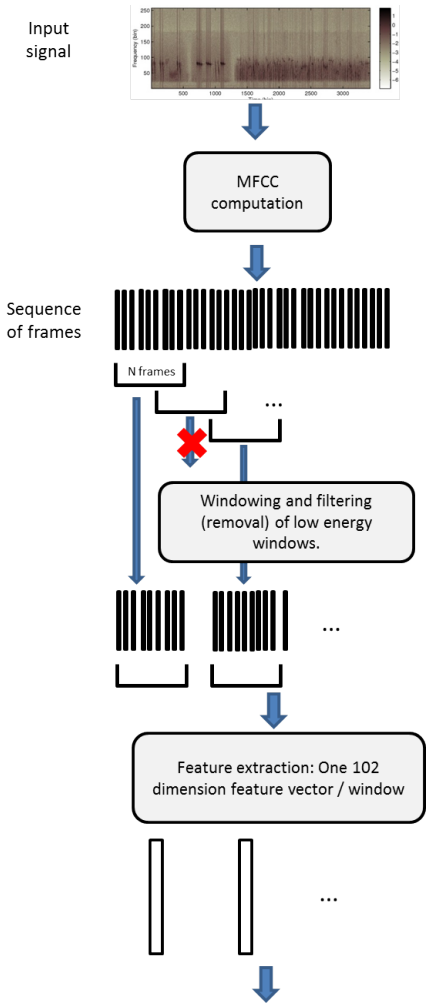

Figure 1. Main steps of the preprocesing and of feature extraction.

perturbate the training and test steps. This is performed with a clustering step (learnt on training signals) that only considers the average energy of the frames in a window. Ideally this cluster makes that the windows are clustered into silence segments on the one hand, and calls and song segments on the other hand. Each window with low average energy is considered a silence window and removed from consideration. Our best results were achived when performing a clustering in three clusters and removing all windows in the lowest energy cluster.

Feature extraction. The final step of the preprocessing consists in computing a reduced set of features for any remaining segment / window. Recall that each segment consists in a series of n 17-dimensional feature vectors (with n in the order of hundreds). Our feature extraction consists in computing 6 values for representing the series of n values for each of the 17 mfcc features. Let consider a particular mfcc feature v, let note $(v_i)_{i=1..n}$ the n values taken by this feature in the n frames of a window and let note $\bar{v_i}$ the mean value of v_i. Moreover let note d and D the velocity and the acceleration of v, which are approximated all along the sequences with $d_i = v_{i+1} - v_i$, and $D_i = d_{i+1} - d_i$. The 6 values we compute are defined as:

$$f_1 = \frac{\sum_{i=1}^{n}(|v_i|)}{n} \tag{1}$$

$$f_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v_i})^2} \tag{2}$$

$$f_3 = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(d_i - \bar{d_i})^2} \tag{3}$$

$$f_4 = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}(D_i - \bar{D_i})^2} \tag{4}$$

$$f_5 = \frac{\sum_{i=1}^{n-1}|d_i|}{n-1} \tag{5}$$

$$f_6 = \frac{\sum_{i=1}^{n-2}|D_i|}{n-2} \tag{6}$$

At the end a segment in a window is represented as the concatenation of the 6 above features for the 17 cepstral coefficients. It is then a new feature vector s_t (with t the number of the window) of dimension 102.

Each signal is finally represented as a sequence of feature vectors s_t, each representing a duration of about 0.5 second with 80% overlap.

2.3. Training

Based on the feature extraction step we described above the simplest strategy is to train a classifier (e.g. we used Support Vector Machines) on the feature vectors s_t which are long enough to include a syllabe or a call, with the idea of agregating all the results found on the windows of a test signal to decide which species are present (see section *Inference* below).

Yet we found that a better strategy was to first perform a clustering in order to split all samples (i.e. s_t) corresponding to a species into two different classes. The rationale behind this process is that call and song of a particular species are complety different sounds (Fagerlund, 2004) so that corresponding feature vectors s_t probably lie in different areas in the feature

space. It is then probably worth using this prior to design classifiers (hopefully linear) with two times the number of species rather than using non linear classifiers with as many classes as there are species.

We implemented this idea by clustering all the frames s_t for a given species into two or more clusters. The two clusters are now considered as two classes that correspond to a single species. At the end, a problem of recognizing K species in a signal turns into a classification problem with $2 \times K$ classes. Note also that since the setting of the challenge is such that there is only one species per training signal, all feature vectors s_t of all signal of a given bird species b_u that fall into cluster one are labeled as belonging to class b_u^1 and all that fall into cluster 2 are labeled as belonging to class b_u^2.

The final step is to learn a multiclass classifier (SVM) in a one-versus-all fashion, i.e. learning one SVM to classify between the samples from one class and the samples from all other classes. This is a standard approach (named Binary Relevance) for dealing with multilabel classification problem where one sample may belong to multiple classes. It is the optimal method with respect to the Hamming Loss, i.e. the number of class prediction errors (either false positive and false negative).

2.4. Inference

At test time an incoming signal is first preprocessed as explained before in section 2.1, silence windows are removed (using clusters found on the train dataset), and feature extraction is performed for all remaining segments. This yields that an input signal is represented as a series of m feature vectors s_t.

All these feature vectors are processed by all $2K$ binary SVMs which provide scores that are interpreted as class posterior probabilities (we use a probabilistic version of SVM), we then get a matrix $m \times 2K$ of scores $P(c|s_t)$ with $c \in \left\{b_u^j | u = 1..K, j = 1, 2\right\}$ and $t = 1..m$.

We experimented few ways to agregate all these scores into a set of K scores, one for each species, enabling ranking the species by decreasing probability of occurence. Indeed this is the expected format of a challenge submission, from which a AUC (Area Under the Curve) score is computed. First we compute $2K$ scores, one for each class, then we agregate the scores of the two classes of a given species.

Our best results were obtained by computing mean probabilities of all scores $\{P(c|s_t)|t = 1..m\}$ for each class c, using harmonic mean or trimmed harmonic mean (where a percentage of the lowest scores are discarded before computing the mean). This yields scores that we consider as class posterior probabilities of classes given the input signal x, $P(c|x)$.

The utimate step consists in computing a score for each species b_u given the scores of the the two corresponding classes b_u^1 and b_u^2. We used the following agregation formulae:

$$P(b_u|x) = 1 - (1 - P(b_u^1|x)) \times (1 - P(b_u^2|x)) \quad (7)$$

3. Experiments

3.1. Dataset

We describe now the data used for the "Machine Learning for Bioacoustics" technical challenge. Note that the training dataset (signals with corresponding ground truth) was available for learning systems all along the challenge together with the test set, without ground truth. Participants were able to design their methods and select their best models by submitting predictions on the test set which were scores on a subset only of the test set (33%). The final evaluation and the ranking of participants was performed on the full test set once all participants have selected 5 of all their systems submitted.

Training data consisted in thirty five 30-seconds audio recordings labeled with a single species, there was one recording per species (35 species overall). Yet, some train recording can include low signal-to-noise ratio (S.N.R) signals of a second bird species of bird. Moreover, according to circadian rythm of each species, other acoustically actives species of animals can be present such as nocturnal and diurnal insects.

Test data consisted in ninety 150-seconds audio recordings with possibly none or multiple species occuring in each signal.

The training and test data recordings have been performed with various devices in various geographical and climatological settings. In particular background and S.N.R are very different between training and test. All wav audio recordings have been sampled at 44 100 Hz with a 16-bits quantification resolution. Recordings were performed with 3 Song Meter SM2+ (Wildlife Acoutic recording device). Each SM2+ has been installed in a different sector (A,B and C) of a Regional Park of the Upper Chevreuse Valley. Every SM2+ recorded, at the same dates and hours (between 24 03 2009 and 22 05 2009), one 150-seconds recording per day between 04h48m00s a.m. and 06h31m00s a.m., which correspond to the acoustical maximum bird-activity period.

3.2. Implementation details

Frames and overlapping sizes. We computed Mel-frequency cepstral coefficients (MFCC) with the *melfcc.m Matlab* function from ROSA laboratory of Columbia university (Ellis, 2005). This function propose 17 different input parameters. We tested numerous possible configurations (Dufour et al., 2012) and measured for each one the difference of energy contained in a given TRAIN file and a reconstructed signal of this recording based on cepstral coefficients. Other details are given on the ICML bird challenge official web page : `http://sabiod.univ-tln.fr/icml2013/BIRD_SAMPLES/`. The difference was minimal with following parameters values:

window=512, *fbtype=mel*, *broaden=0*, *maxfreq=sr/2*, *minfreq*=0, *wintime=window/sr*, *hoptime= wintime/3*, *numcep=16*, *usecmp=0*, *dcttype=3*, *nbands=32*, *dither=0*, *lifterexp=0*, *sumpower=1*, *preemph=0*, *modelorder=0*, *bwidth=1*, *useenergy=1*

This process transformes a 30-seconds train audio recording (at 44 kHz sampling rate) into about 7 700 frames of 16 cepstral coefficients which we augmented with the energy computed by setting *useenergy=0*.

Next we computed feature vector s_t on 0.5 second windows with 80% overlap, which yields about $n = 300$ feature vectors per training signal (hence per species since there is only one training recording per species) and about $m =$ feature vectors per test signal.

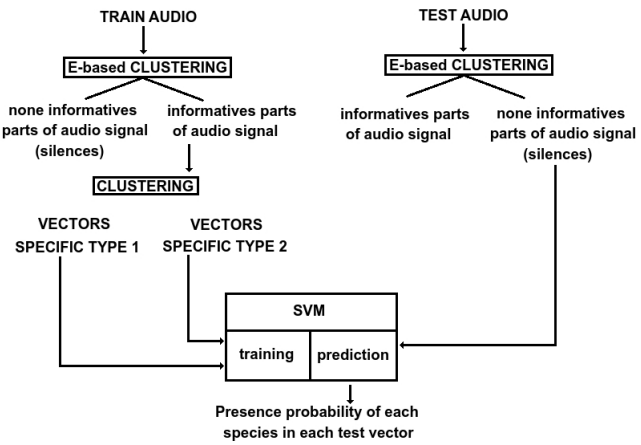

Figure 2. Technical principe of our best scored run

Table 1. Score Kaggle icml (AUC) according to the way scores are aggregated. Public scores are calculated on approximately 33% of the test data. Private scores are based on the other 67%.

mean aggregation	Private score	Public score
arithmetic mean	0.61362	0.63974
harmonic mean	0.64234	0.67344
trimmed mean 10%	0.64158	0.68612
trimmed mean 20%	0.64639	0.69163
trimmed mean 24%	0.64699	0.69103
trimmed mean 30%	0.64614	0.68881

LIBSVM settings. We used a multiclass S.V.M algorithm based on LIBSVM (Chang, 2008). We selected model parameters (kernel type etc) through two fold cross validation. Best scores have been obtained with C-SVC SVM type and linear kernel function.

3.3. Results

3.3.1. General results

We report only our best results that correspond to the method presented in this paper for various computation for the class score at inference time. Table 1 shows how the way the mean score of a class is computed on the test set (see section 2.4) influences the final result. The table compares arithmetic mean, harmonic mean, and trimmed arithmetic mean (at 10, 20 et 30%). A trimmed mean at p% is the arithmetic mean computed after discarding p% extreme values, i.e. the $p/2$% lowest values and the $p/2$% largest values.

Although our method is simple it reached the fourth rank over more than 77 participating teams at the Kaggle ICML Bird challenge with a score of 0.64639 while the best score was 0.69454. It is also worth noting that our system ranked about fifteen only on the validation set (one third of the total test set). This probably shows that our system being maybe simpler than other methods exhibits at the end a more robust behaviour and improved generalization ability.

3.3.2. Monospecific results

According to these scores for 7 species, we notice:

- Scores of our model are close from best ones and evolve the same way according the concerned species. The slight difference is probably due to the way we calculate (trimmed mean) the presence probability of one given species in a 150-seconds recording from presence probability of this same species in a half-second frame.

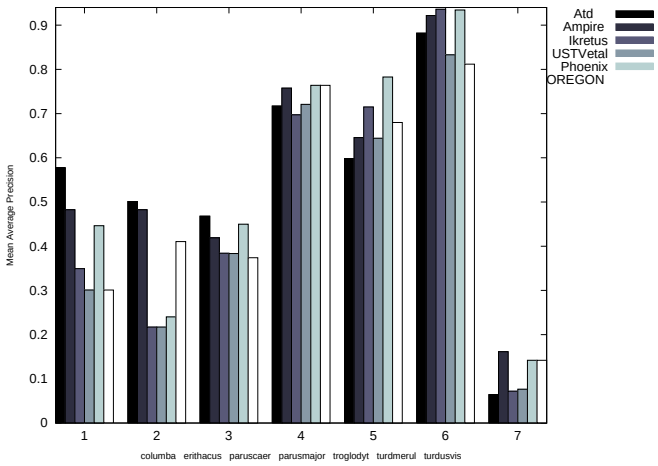

Figure 3. Mean average precision scores for 7 species reached by 6 best teams

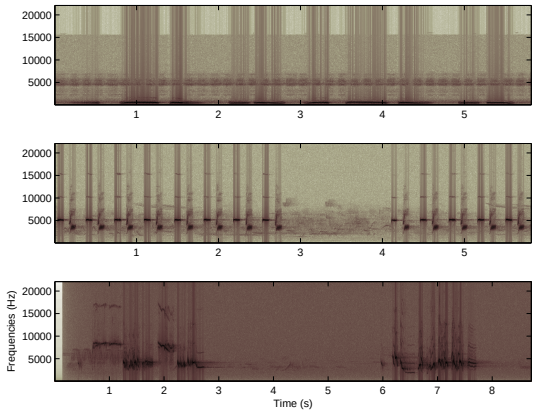

Figure 4. Time-Frequential spectrograms of train recording's extracts. From top to bottom: Common Wood-pigeon, Great Tit and European Robin.

- For all teams, scores are less adequate for *Columba palumbus* (Common Wood-pigeon), *Erithacus rubecula* (European Robin), *Parus caeruleus* (Blue Tit) et *Turdus viscivorus* (Mistle Thrush).
  - In Common Wood-pigeon (figure 4) train recording, this species emits series of 5 syllables (around 500 Hz). Syllables are very stable and different. Temporal construction of series is strict. Plus, train recording is highly corrupted by cicades between 4 and 6 kHz and in TEST recording, S.N.R. is low, series last 2.5 seconds (against 4 seconds in TRAIN) and are composed of 6 syllables well differentiate.
  - European Robin (figure 4) is typically a species of bird whose sings are rich in syllables and variable. Frequential-domain variability between different sings and syllables is important. Sing duration varies between 1.5 and 3 seconds. It is one of rare species emeting syllables up to 8 kHz.
  - In Blue Tit train recording, othes species of birds are present. Therefore, Blue Tit produces 5 different cries composed of 5 different syllables.
  - Mistle Thrush train recording sings vary much and are very different from sings in test recordings.

  MFCC compression has property to lower weights cepstral coefficients corresponding to higher frequencies of spectrogram. As a result, MFCC leads to lose a part of the signals may be important in Blue Tit's case. Futhermore, high variability of cries or sings of this is hard to manage by classifiers, especially when they are constrained to retain and learn only 2 types of emission per species. Considering two types of emissions was particularly inappropriate for these 3 cases.

- For all teams, scores are very satisfactory for *Parus major* (Great Tit), *Troglodytes troglodytes* (Winter Wren), *Turdus merula* (Eurasian Blackbird).
  - Great Tit's signals are very simple and perfectly repeated. A 500-hertz high-pass filter has been apply during train recording.
  - Winter Wren's acoustic patterns are really stable. A 1000-hertz high-pass filter has been applied during train recording (figure 4).
  - Eurasian Blackbird's train recording has been filtered by 1000-hertz high-pass filter and 6000-hertz low-pass filter.

  Best Mean Average Precisions have been obtain when low-frequency and high-frequency noises had been cleared by filter.

We assume that congruence observed between scores of 6 best teams for this 7 species is the same considering every species. The fact that scores of each species evolve the same way indicates that M.A.P (Mean Average Precision) differences between species can be due to:

1. some species produce sounds harder to characterize than others : strong variability in frequential and/or temporal domain.
2. Train recordings can't be compared to test recordings regarding S.N.R: frequential filters, harmonic

richness, source-microphone distances ... differ a lot.

3. Signals of interest are easier to extract in some train recordings than in others because of data acquisition. Some frequential filters have been applied to a part of train recordings.

4. For a given species, signals provided by corresponding train recording may be not exhaustive in comparison with signals present in test audios.

5. According to species, frequential content of emissions and location of source in its environment vary much. Each species of bird use differently available space in an ecosystem. Obstacles between source and microphone depend on diet and customs of species (arboricol, walking, granivorous, insectivorous species ...). But all frequencies aren't affected the same way by transmission loss in the environment. By instance, low frequencies are particularly well filtered by vegetal covert close from the ground. Common Woodpigeon typically emits in low frequencies.

6. Natural (rain, wind, insects) or anthropic (motors ...) acoustic parasites originated from are more diversified and strong (regarding energy) in test audios than in train. In addition, nature and acoustic presence of parasites vary much from one species to an other.

Hence, it seems reasonable to affirm that more complex syllable extraction methods (segmentation step) combined to MFCC constitute a good solution to improve our performances. They would allow to retain intraspecific variability of each class and eliminate none-relevant information.

4. Conclusion and perspectives

Although the method that we presented is simple it was to perform well on the challenge and to be much robust between validation step and test set. We believe this robustness comes from the simplicity of the method that do not rely on complex processing steps (like identifying syllables) that other participants could have used (Glotin & Sueur, 2013).

Possible improvements would consist in the integration in the model of additional information such as weather condition, or a taxonomia of species, allowing for more accurate hierarchical classification schemes.

5. Acknowledgments

This work is supported by the MASTODONS CNRS project Scaled Acoustic Biodiversity SABIOD and the Institut Universitaire de France which supports the "Complex Scene Analysis" project. We thank F. Jiguet and J. Sueur and F. Deroussen (Deroussen & Jiguet, 2011; Deroussen, 2001) who provided the challenge data.

PhD funds of 1st author are provided by Agence De l'Environnement et de la Maîtrise de l'Energie (mila.galiano@ademe.fr) and by BIOTOPE company (Dr Lagrange, hlagrange@biotope.fr, R&D Manager).

References

Briggs, F., Fern, X., and Raich, R. Acoustic classification of bird species from syllables: an empirical study. Technical report, Oregon State University, 2009.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X., Raich, R., Betts, M., Frey, S., and Hadley, A. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *Journal of the Acoustical Society of America*, 2012.

Chang, Chih-Chung. Libsvm. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chang-Hsing, L., Yeuan-Kuen, L., and Ren-Zhuang, H. Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications Vol. 1, pp.17-23*, 2006.

Deroussen, F. Oiseaux des jardins de france. Nashvert Production, Charenton, France, 2001. naturophonia.fr.

Deroussen, F. and Jiguet, F. Oiseaux de france, les passereaux, 2011.

Dufour, O., Glotin, H., Artières, T., and Giraudet, P. Classification de signaux acoustiques : Recherche des valeurs optimales des 17 paramètres dentrée de la fonction melfcc. Technical report, Laboratoire Sciences de l'Information et des Systèmes, Université du Sud Toulon Var, 2012.

Ellis, Daniel P. W. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. URL http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/. online web resource.

Fagerlund, S. Acoustics and physical models of bird sounds. In *Seminar in acoustics, HUT, Laboratory of Acoustics and Audio Signal Processing*, 2004.

Glotin, H. and Sueur, J. Overview of the first international challenge on bird classification, 2013. URL http://sabiod.univ-tln.fr. online web resource.

Michael Noll, A. Short-time spectrum and cepstrum techniques for vocal-pitch detection. *Journal of the Acoustical Society of America, Vol. 36, No. 2, pp. 296-302*, 1964.

Neal, L., Briggs, F., Raich, R., and Fern, X. Time-frequency segmentation of bird song in noisy acoustic environments. In *International Conference on Acoustics, Speech and Signal Processing*, 2011.