Chapter 5

Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification

Olivier Dufour, Thierry Artieres, Hervé Glotin and Pascale Giraudet

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/56872

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. 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. 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 multilabeled.

2. Description of the method

We present now the main steps of our approach. The Figures 1 and 2 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 $\{(x^i, y^i) \mid i = 1..\ N\}$ where each input x^i is an audio recording and each y^i is a bird species $\forall\ i, y^i \in \{b_u \mid u = 1..\ K\}$ (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.

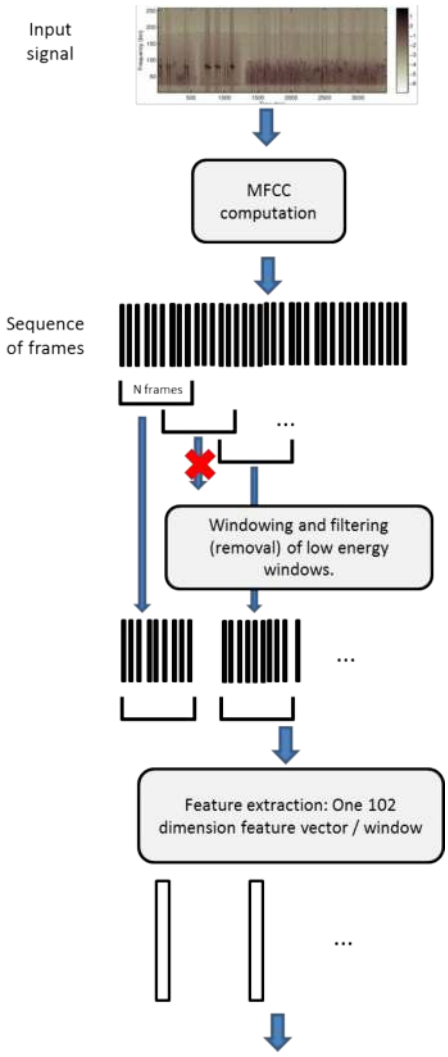

Figure 1. Main steps of the preprocesing and of feature extraction.

http://dx.doi.org/10.5772/56872

2.1. Preprocessing

Our preprocessing is based on MFCC cepstral coefficients, which have been proved useful for speech recognition [4, 11]. A signal is first transformed into a series of frames where each frame consists in 17 MFCC (mel-frequency cepstral coefficients) feature vectors, including energy. Each frame represents a short duration (e.g. 512 samples of a signal sampled at 44.1 kHz).

2.2. Windowing, silence removal and feature extraction

2.2.1. 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 syllable extraction step that is used in most of methods for bird identification [12, 2, 1], but is much simpler to implement. In our case we considered segments of about 0.5 second duration (i.e. n ~ few hundreds of frames) and used a sliding window with overlap (about 80%).

2.2.2. Silence removal

We first want to remove segments (windows) corresponding to silence since these would 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 clustering makes that the windows are clustered into silence segments on the one hand, and calls and songs segments on the other hand. Each window with low average energy is considered a silence window and removed from consideration. Our best results were achieved when performing a clustering in three clusters and removing all windows in the lowest energy cluster.

2.2.3. 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 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 six features we compute are defined as:

$$f_1 = \frac{\sum_{i=1}^{n}\left(\left|v\,i\right|\right)}{n} \tag{1}$$

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

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

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

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

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

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 duration of about 0.5 second with 80% overlap.

2.3. Training

Based on the feature extraction step we described above the simplest strategy to train a classifier (e.g. we used Support Vector Machines) on the feature vectors S_t which are long enough to include a syllable or a call, with the idea of aggregating 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 calls and songs of a particular species are completely different sounds [9] 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 x 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.

http://dx.doi.org/10.5772/56872

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), 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 x $2K$ of scores P (c | S_t) with $c \in \{b_u^j \mid u = 1..K,\ j = 1,2\}$ and t = 1..m.

We experimented few ways to aggregate all these scores into a set of K scores, one for each species, enabling ranking the species by decreasing probability of occurrence. Indeed this is the expected format of a challenge submission, from which an AUC (Area Under the Curve) score is computed. First we compute $2K$ scores, one for each class, then we aggregate 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 ultimate step consists in computing a score for each species b_u given the scores of the two corresponding classes b_u^1 and b_u^2. We used the following aggregation formulae:

$$P(b_u \mid x) = 1 - \left(1 - P\left(b_u^1 \mid x\right)\right) \times \left(1 - P\left(b_u^2 \mid x\right)\right) \tag{7}$$

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 were 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 (SNR) signals of a second bird species of bird. Moreover, according to circadian rhythm of each species, other acoustically actives species of animals can be present such as nocturnal and diurnal insects (Grylidae, Cicada).

Test data consisted in ninety 150-seconds audio recordings with possibly none or multiple species occurring 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 SNR 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 Acoustic 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 maximal acoustical bird-activity period.

3.2. Implementation details

3.2.1. Frames and overlapping sizes

We computed Mel-frequency cepstral coefficients (MFCC) with the *melfcc.m Matlab* function from ROSA laboratory of Columbia University [8]. This function proposes 17 different input parameters. We tested numerous possible configurations [7] 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.

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 transforms 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.

3.2.2. LIBSVM settings

We used a multiclass SVM algorithm based on LIBSVM [3]. 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.

http://dx.doi.org/10.5772/56872

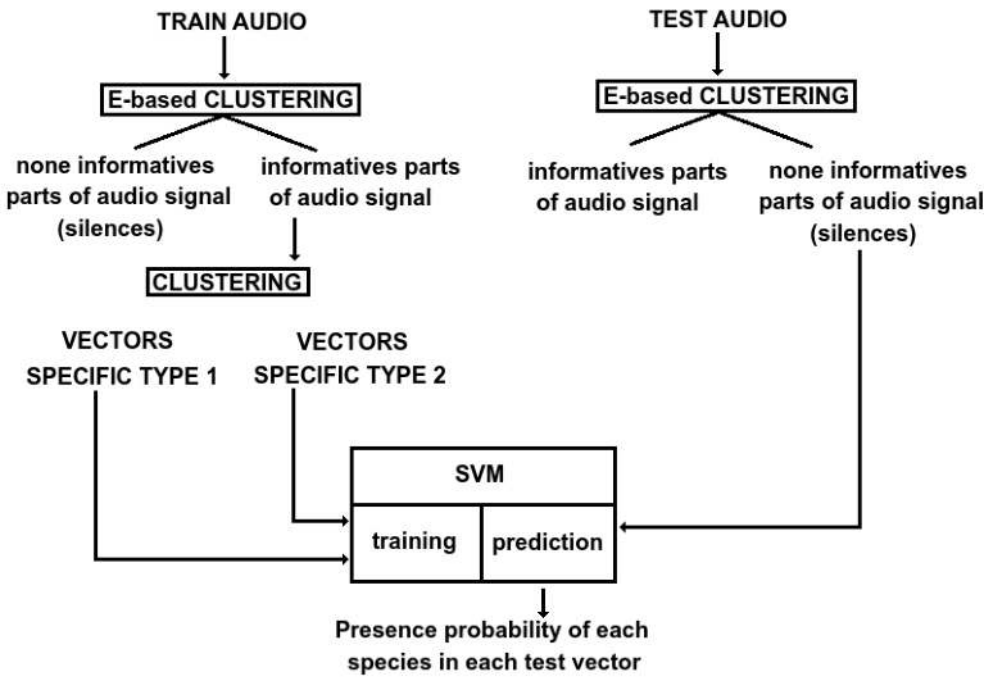

Figure 2. Technical principle of our best-scored run

4. Results

4.1. General results

We report only our best results that correspond to the method presented in this paper for various computations 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) and 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 (Private score) of all challengers was 0.694 (the corresponding public Leaderboard score was 0.743). See [13] for the best system, and [14] for the description of the other systems. 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 behavior and improved generalization ability.

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

Table 1. Score Kaggle icml (AUC) according to the way scores are aggregated. Public scores are calculated on a third of the test data, while private scores are calculated on the other part. Only the private scores are the official competition results. The best private score of all challengers is 0.694 [13].

4.2. Monospecific results

According to these scores for 7 species, we notice in Figure 3:

- Scores of our model are close to the best ones and evolve the same way for 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 compared to the presence probability of this same species in a half-second frame.
- All teams were not able to score high for *Columba palumbus* (Common Wood-pigeon), *Erithacus rubecula* (European Robin), *Parus caeruleus* (Blue Tit), *Parus palustris* (Marsh Tit), *Pavo cristatus* (Blue Peafowl) and *Turdus viscivorus* (Mistle Thrush).
    - In the Common Wood-pigeon (top of Figure 4) train recording, we can see a series of 5 syllables (around 500 Hz). Syllables are very stable and different. Their alternation in time domain is strict. Also, the train recording is highly corrupted by cicadas between 4 and 6 kHz and in the test recording, SNR is low. The series last 2.5 seconds (compared to 4 seconds in TRAIN) and are composed of 6 syllables well differentiated.
    - The European Robin (bottom of Figure 4) is typically bird species whose songs are diverse and rich in syllables. Frequency-domain variability between different songs and syllables is important. Song duration varies between 1.5 and 3 seconds. It is one of the rare species that can emit up to 8 kHz.
    - In Blue Tit train recording, other species of birds are present. Therefore, Blue Tit produces 5 different cries composed of 5 different syllables.
    - Mistle Thrush train recording songs vary a lot and are very different from songs in the test recordings.

MFCC compression has the property of lowering the weights of cepstral coefficients corresponding to higher frequencies of the spectrogram. As a result, MFCC can lead to losing a part of the signal that may be important in European Robin's case. Futhermore, the high variability of the cries or songs of the different species is difficult to manage by classifiers, especially when they are constrained to retain and learn only 2 types of emissions per species. Considering two types of emissions was particularly sub-optimal for these 3 cases.

http://dx.doi.org/10.5772/56872

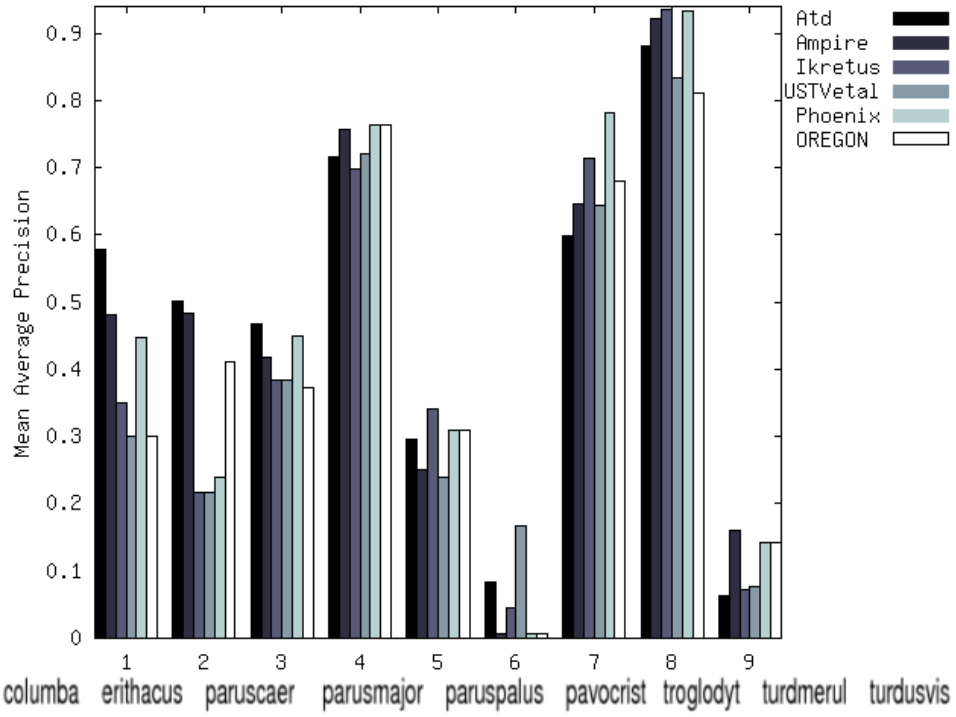

Figure 3. Mean Average Precision (MAP) scores on nine species (ordered in abscissa from left to right) of the 6 best teams of the challenge. The label 'USTVetal' refers to our team (MAP was not the official metrics of the challenge but give interesting comparisons).

- For all teams, scores were very satisfactory for *Parus major* (Great Tit), *Troglodytes troglodytes* (Winter Wren) and *Turdus merula* (Eurasian Blackbird).
  - Great Tit's signals (middle of Figure 4) are very simple and periodically repeated. A 500-hertz high-pass filter has been applied on the train recording.
  - Winter Wren's acoustic patterns are really stable. A 1000-hertz high-pass filter has been applied on the train recording.
  - Eurasian Blackbird's train recording has been filtered by a band pass filter from 1-6 kHz. Best Mean Average Precisions were obtained when low frequency and high-frequency noise was removed by filtering.

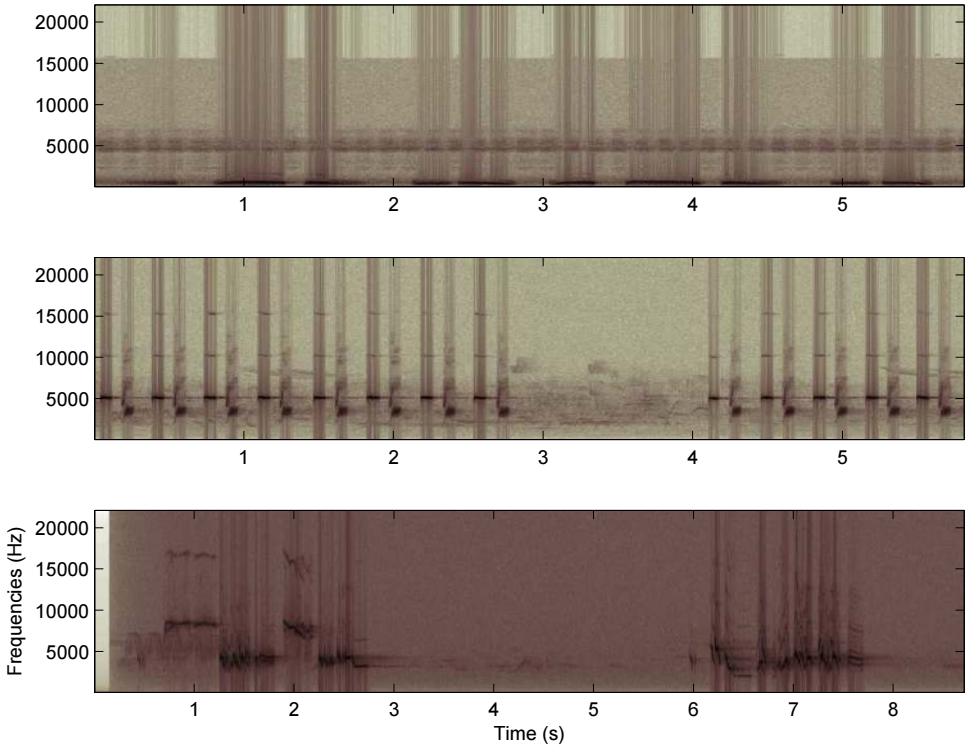

Figure 4. Time-Frequential spectrograms of train recording's extracts. From top to bottom: Common Wood-pigeon, Great Tit and European Robin.

We assume that congruence observed between the scores of the 6 best teams for these 9 species is the same considering each of the species. The fact that the scores of each species evolve the same way indicates that the Mean Average Precision (M.A.P) differences between species can be due to:

1. Some species produce sounds harder to characterize than others: strong variability in frequency and/or temporal domain.
2. Train recordings can't be compared to test recordings regarding SNR: filters, harmonic richness, source-microphone distances etc. differ a lot.
3. Signals of interest are easier to extract in some train recordings than in others because of data acquisition. Some filters have been applied to a part of the train recordings.

http://dx.doi.org/10.5772/56872

4. For a given species, the signals provided in the train recording may not include a global repertoire and this way not be part of the respective species test recordings.

5. For each species, frequency content of emissions and location of source in its environment differ widely. Each bird species uses the available space in an ecosystem differently. Obstacles between source and microphone depend on diet and customs of species (arboricol, walking, granivorous, insectivorous species etc). But all frequencies aren't affected the same way by transmission loss in the environment. For example, low frequencies are particularly well filtered by vegetation close from the ground. Common Wood-pigeon typically emits in low frequencies (see figure 4).

6. Natural (rain, wind, insects) or anthropic (motors etc) acoustic events are more diverse and strong (regarding energy) in test recordings than in train. In addition, these events vary much from one species to an other.

Hence, it seems reasonable to affirm that more complex syllables extraction methods (segmentation step) combined with the MFCC way constitute a better solution to improve our performance. They would allow us to retain intraspecific variability for each class and eliminate non-relevant information.

5. Conclusion and perspectives

Although the method that we presented is simple it performed well on the challenge and was 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 [10, 13, 15, 16].

Possible improvements would consist in the integration in the model of additional information such as syllables extraction, weather condition, or a taxonomia of species, allowing more accurate hierarchical classification schemes. Also the MFCC shall be replaced either by a scattering transform [17] or a deep convolutional network [18], that build invariant, stable and informative signal representations for classification.

Acknowledgements

We thank Dr. Xanadu Halkias for her useful comments on this paper. This work is supported by the MASTODONS CNRS project Scaled Acoustic Biodiversity SABIOD and the Institut Universitaire de France that supports the “Complex Scene Analysis” project. We thank F. Jiguet and J. Sueur and F. Deroussen [6, 5] 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).

Author details

Olivier Dufour[1,2], Thierry Artieres[3], Hervé Glotin[1,2,4] and Pascale Giraudet[1]

*Address all correspondence to: olivierlouis.dufour@gmail.com, thierry.artieres@lip6.fr, glotin@univ-tln.fr, giraudet@univ-tln.fr

1 Université du Sud Toulon Var, France

2 Aix-Marseille Université, CNRS, ENSAM, LSIS, Marseille, France

3 LIP6, Université Paris VI, France

4 Institut Universitaire de France, Paris, France, France

References

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

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

[3] C.-C. Chang. Libsvm. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2008.

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

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

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

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

[8] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

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

http://dx.doi.org/10.5772/56872

[10] H. Glotin and J. Sueur. Overview of the first international challenge on bird classification, 2013. online web resource.

[11] A. Michael Noll. 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.

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

[13] Rafael Hernandez Murcia, "Bird identification from continuous audio recordings", in Proc. of first int. wkp of Machine Learning for Bioacoustics ICML4B joint to ICML 2013, Ed. H. Glotin et al, Atlanta, ISBN 979-10-90821-02-6 http://sis.univ-tln.fr/~glotin/ICML4B2013_proceedings.pdf

[14] H. Glotin, Y. Lecun, P. Dugan, C. Clark, X. Halkias, Proceedings of the first int. wkp of Machine Learning for Bioacoustics ICML4B joint to ICML 2013, Ed. H. Glotin et al, Atlanta, ISBN 979-10-90821-02-6 http://sis.univ-tln.fr/~glotin/ICML4B2013_proceedings.pdf

[15] Briggs et al., "ICML 2013 Bird Challenge – Tech Report, in Proc. of first int. wkp of Machine Learning for Bioacoustics ICML4B joint to ICML 2013, Ed. H. Glotin et al, Atlanta, ISBN 979-10-90821-02-6 http://sis.univ-tln.fr/~glotin/ICML4B2013_proceedings.pdf

[16] Dan Stowell and Mark D. Plumbley, " Acoustic detection of multiple birds in environmental audio by Matching Pursuit",in Proc. of first int. wkp of Machine Learning for Bioacoustics ICML4B joint to ICML 2013, Ed. H. Glotin et al, Atlanta, ISBN 979-10-90821-02-6 http://sis.univ-tln.fr/~glotin/ICML4B2013_proceedings.pdf

[17] J. Andén and S. Mallat, "scattering transform applied to audio signals and musical classification" Proceedings of International Symposium on Music Information Retrieval (ISMIR'11), 2011

[18] Mikael Henaff, Kevin Jarrett, Koray Kavukcuoglu and Yann LeCun: Unsupervised Learning of Sparse Features for Scalable Audio Classification, Proceedings of International Symposium on Music Information Retrieval (ISMIR'11), (Best Student Paper Award), 2011