Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition*

Mario Lasseck
Animal Sound Archive
Museum für Naturkunde Berlin
Mario.Lasseck@mfn-berlin.de

Abstract

The challenge of the NIPS4B competition is to identify 87 sound classes of birds and other animals present in 1000 audio recordings, collected in the field. The difficulty of this task lies in the large number of species and sounds that have to be identified in various contexts dealing with different levels of background noise and simultaneously vocalizing animals. The solution presented here ranks first place on the kaggle private leaderboard and achieves an Area Under the Curve of 91.7% (AUC).

1 Introduction

The audio data was recorded at different places in Provence France and is provided by the BIOTOPE society, having one of the largest collections of wildlife recordings of birds in Europe. The nearly 2 hours of recordings are split into smaller clips ranging from 0.25 to 5.75 seconds. The recordings were done with Wildlife Acoustics SM2 and are presented in uncompressed WAV format with a sample rate of 44.1 kHz. The 87 individual sound classes within these recordings represent different bird species and their songs, calls and drumming. Other animal species living in the same environment like insects and one amphibian are also included. The training set consists of 687 audio files. Each file is paired with the subset of sound classes present in that recording. Some recordings are empty, containing only background noise, others contain up to 6 different simultaneously vocalizing birds or insects. Each species is represented by nearly 10 training files within various contexts, different background noises and an arbitrary number of other species. The goal of the competition is to identify which of the 87 sound classes of birds and amphibians are present in 1000 continuous wildlife recordings, using only the provided audio files and machine learning algorithms for automatic pattern recognition.

2 Preprocessing and Segmentation

The method of segmentation has a big influence on classification results. Several different approaches were tested. The one that works best regarding leaderboard score is surprisingly simple. Audio files are first resampled to 22050 Hz. After applying the STFT using a hanning window with a size of 512 samples and 75% overlap the resulting spectrogram is normalized to a maximum of 1.0. The 4 lowest and 24 highest frequency bins are removed,

* In proc. of 'Neural Information Scaled for Bioacoustics' joint to NIPS, http://sabiod.org/nips4b, Nevada, dec. 2013, Ed. Glotin H. et al.

leaving 228 frequency bins or spectrogram rows representing the relevant frequency range of approximately 170 to 10000 Hz. The narrowed spectrogram of each audio file is treated as grayscale image and further processed for noise reduction and segmentation.

To reduce background noise each pixel value is set to 1 if it is above 3 times the median of its corresponding row (frequency band) AND 3 times the median of its corresponding column (time frame), otherwise it is set to 0. This *Median Clipping* per frequency band and time frame removes already most of the background noise. Variable noise levels in different frequency regions are compensated and short, broadband distortions coming from rain, wind or microphone handling are attenuated.

The resulting binary image is further processed using standard image processing techniques (e.g. closing, dilation, median filter). Finally, all connected pixels exceeding a certain spatial extension are labeled as a segment and a rectangle with a small area added to each direction is used to define its size and position. Figure 1 gives an example of the preprocessing steps involved and Figure 2 shows the outcome of a complete segmentation process.

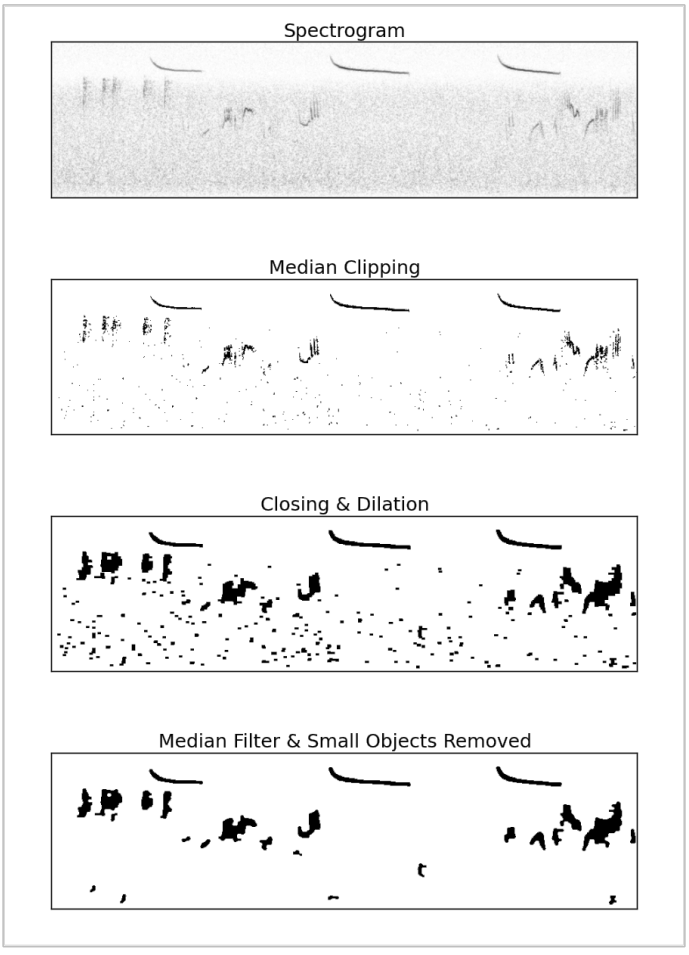


Figure 1: Preprocessing of Spectrogram Image

Preprocessing the spectrograms extracts 9198 segments from the training data and 16726 segments from the test recordings.

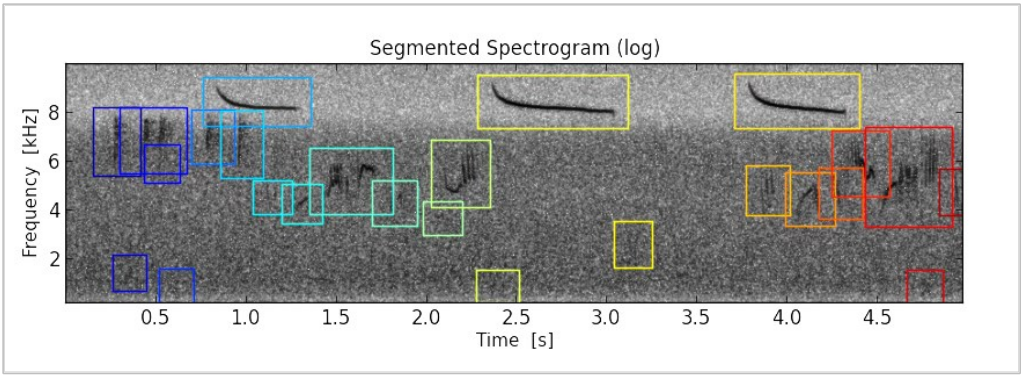


Figure 2: Spectrogram Image with marked Segments

3 Feature Extraction

Features are calculated for both, training and test files, coming from three different sources: File-Statistics, Segment-Statistics and Segment-Probabilities.

File-Statistics include minimum, maximum, mean and standard deviation taken from all values of the unprocessed spectrogram. Additionally the spectrogram is divided into 16 equally sized and distributed frequency bands and their minima, maxima, means and standard deviations are also included.

For Segment-Statistics the number of segments per file plus minimum, maximum, mean and standard deviation for width, height and frequency position of all segments per file are calculated.

In order to find Segment-Probabilities a variation of Fodor's method [1] is used which was already successfully applied in the MLSP 2013 Competition. The highest matching probability of all segments extracted from training files associated with one or more sound class is determined in all files by template matching using normalized cross-correlation [2]. A Gaussian blur with a sigma of 1.5 is applied to segment and target spectrogram before matching. Best matches are only searched for within the frequency range of the segment (± a small tolerance of 4 pixels). Unlike Fodor, the template matching uses only absolute-intensity spectrograms and for better performance the OpenCV library [4] is used.

File- and Segment-Statistics produce 81 features per file scaled to the range [0 1]. Segment-Probabilities create, corresponding with the number of extracted segments from the training set, 9198 features per file.

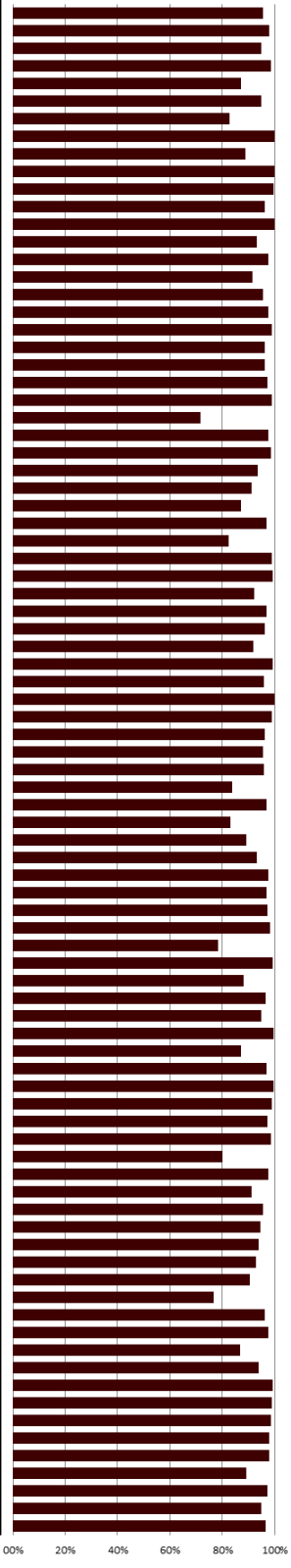
4 Feature Selection

As in [1] already suggested the multi-instance multi-label classification problem is turned into 87 individual classification problems. In that way the probability for each target sound class is calculated separately for all files. Each of the 87 classifiers uses all File- and Segment-Statistics. But as for the Segment-Probabilities, only those belonging to segments extracted from training files associated with the corresponding target sound class are included. This way the number of features to be taken into account for learning and predicting a particular sound class can be reduced significantly which produces much better classification results.

To give an example: Sound class 86 appears in 8 training files (107, 172, 264, 353, 387, 504, 510 and 596). For learning and predicting this particular sound class, only matching probabilities belonging to segments extracted from these 8 files are included as features. The number of relevant features selected from Segment-Probabilities for each sound class is listed in Table 1.

Table 1: Number of selected features and estimators per sound class plus AUC scores

Class			Features				
No.	Name	Scientific Name	File Statistics	Segment Statistics	Segment Probilities	Estimators	AUC
1	Aegcau_call	Aegithalos caudatus	68	13	148	458	95.6%
2	Alaarv_song	Alauda arvensis	68	13	213	500	97.8%
3	Anttri_song	Anthus trivialis	68	13	271	500	95.0%
4	Butbut_call	Buteo buteo	68	13	131	424	98.5%
5	Carcan_call	Linaria cannabina	68	13	244	500	87.1%
6	Carcan_song	Linaria cannabina	68	13	137	436	94.9%
7	Carcar_call	Carduelis carduelis	68	13	192	500	82.6%
8	Carcar_song	Carduelis carduelis	68	13	301	500	99.8%
9	Cerbra_call	Certhia brachydactyla	68	13	183	500	88.7%
10	Cerbra_song	Certhia brachydactyla	68	13	298	500	99.8%
11	Cetcet_song	Cettia cetti	68	13	226	500	99.6%
12	Chlchl_call	Chloris chloris	68	13	203	500	96.4%
13	Cicatr_song	Cicadatra atra	68	13	240	500	100.0%
14	Cicorn_song	Cicada orni	68	13	98	358	93.3%
15	Cisjun_song	Cisticola juncidis	68	13	118	398	97.5%
16	Colpal_song	Columba palumbus	68	13	103	368	91.6%
17	Corcor_call	Corvus corone	68	13	158	478	95.7%
18	Denmaj_call	Dendrocopos major	68	13	157	476	97.7%
19	Denmaj_drum	Dendrocopos major	68	13	392	500	98.8%
20	Embcir_call	Emberiza cirlus	68	13	208	500	96.1%
21	Embcir_song	Emberiza cirlus	68	13	283	500	96.2%
22	Erirub_call	Erithacus rubecula	68	13	325	500	97.3%
23	Erirub_song	Erithacus rubecula	68	13	275	500	98.8%
24	Fricoe_call	Fringilla coelebs	68	13	361	500	71.6%
25	Fricoe_song	Fringilla coelebs	68	13	230	500	97.6%
26	Galcri_call	Galerida cristata	68	13	331	500	98.5%
27	Galcri_song	Galerida cristata	68	13	152	466	93.4%
28	Galthe_call	Galerida theklae	68	13	117	396	91.1%
29	Galthe_song	Galerida theklae	68	13	155	472	87.0%
30	Gargla_call	Garrulus glandarius	68	13	235	500	97.1%
31	Hirrus_call	Hirundo rustica	68	13	76	314	82.5%
32	Jyntor_song	Jynx torquilla	68	13	166	494	98.9%
33	Lopcri_call	Lophophanes cristatus	68	13	252	500	99.3%
34	Loxcur_call	Loxia curvirostra	68	13	377	500	92.1%
35	Lularb_song	lullula arborea	68	13	323	500	97.1%
36	Lusmeg_call	Luscinia megarhynchos	68	13	211	500	96.3%
37	Lusmeg_song	Luscinia megarhynchos	68	13	307	500	91.7%
38	Lyrple_song	Lyristes plebejus	68	13	235	500	99.3%
39	Motcin_call	Motacilla cinerea	68	13	176	500	95.8%
40	Musstr_call	Muscicapa striata	68	13	154	470	99.9%
41	Oriori_call	Oriolus oriolus	68	13	96	354	99.0%
42	Oriori_song	Oriolus oriolus	68	13	255	500	96.4%
43	Parate_call	Periparus ater	68	13	446	500	95.7%
44	Parate_song	Periparus ater	68	13	609	500	96.0%
45	Parcae_call	Cyanistes caeruleus	68	13	253	500	83.8%
46	Parcae_song	Cyanistes caeruleus	68	13	351	500	96.9%
47	Parmaj_call	Parus major	68	13	312	500	83.3%
48	Parmaj_song	Parus major	68	13	532	500	89.3%
49	Pasdom_call	Passer domesticus	68	13	308	500	93.1%
50	Pelgra_call	Pelophylax kl. grafi	68	13	101	364	97.7%
51	Petpet_call	Petronia petronia	68	13	202	500	96.8%
52	Petpet_song	Petronia petronia	68	13	176	500	97.3%
53	Phofem_song	Pholidoptera femorata	68	13	279	500	98.2%
54	Phycol_call	Phylloscopus collybita	68	13	168	498	78.4%
55	Phycol_song	Phylloscopus collybita	68	13	500	500	99.4%
56	Picpic_call	Pica pica	68	13	185	500	88.1%
57	Plaaff_song	Platycleis affinis	68	13	287	500	96.7%
58	Plasab_song	Platycleis sabulosa	68	13	324	500	94.9%
59	Poepal_call	Poecile palustris	68	13	335	500	99.7%
60	Poepal_song	Poecile palustris	68	13	253	500	87.2%
61	Prumod_song	Prunella modularis	68	13	304	500	96.8%
62	Ptehey_song	Pteronemobius heydenii	68	13	132	426	99.7%
63	Pyrpyr_call	Pyrrhula pyrrhula	68	13	120	402	99.0%
64	Regign_call	Regulus ignicapillus	68	13	264	500	97.2%
65	Regign_song	Regulus ignicapillus	68	13	375	500	98.7%
66	Serser_call	Serinus serinus	68	13	178	500	80.1%
67	Serser_song	Serinus serinus	68	13	465	500	97.5%
68	Siteur_call	Sitta europaea	68	13	147	456	91.3%
69	Siteur_song	Sitta europaea	68	13	466	500	95.6%
70	Strdec_song	Streptopelia decaocto	68	13	66	294	94.7%
71	Strtur_song	Streptopelia turtur	68	13	118	398	94.0%
72	Stuvul_call	Sturnus vulgaris	68	13	155	472	92.9%
73	Sylatr_call	Sylvia atricapilla	68	13	178	500	90.7%
74	Sylatr_song	Sylvia atricapilla	68	13	152	466	76.7%
75	Sylcan_call	Sylvia cantillans	68	13	227	500	96.2%
76	Sylcan_song	Sylvia cantillans	68	13	344	500	97.4%
77	Sylmel_call	Sylvia melanocephala	68	13	287	500	86.8%
78	Sylmel_song	Sylvia melanocephala	68	13	224	500	93.9%
79	Sylund_call	Sylvia undata	68	13	48	258	99.4%
80	Sylund_song	Sylvia undata	68	13	189	500	98.8%
81	Tetpyg_song	Tettigettula pygmea	68	13	222	500	98.7%
82	Tibtom_song	Tibicina tomentosa	68	13	159	480	98.1%
83	Trotro_song	Troglodytes troglodytes	68	13	400	500	97.9%
84	Turmer_call	Turdus merula	68	13	358	500	89.0%
85	Turmer_song	Turdus merula	68	13	441	500	97.1%
86	Turphi_call	Turdus philomelos	68	13	22	206	94.8%
87	Turphi_song	Turdus philomelos	68	13	386	500	96.6%



5 Classification

The scikit-learn library is used for classification [3]. For each sound class an ensemble of randomized decision trees (sklearn.ensemble.ExtraTreesRegressor) is applied. The number of estimators is chosen to be twice the number of selected features per class but not greater than 500. The winning solution considers 4 features when looking for the best split and requires a minimum of 3 samples to split an internal node. During 12-fold cross validation the probability of each sound class in all test files is predicted and at the end, after removing the lowest and highest value, averaged.

Good classification results are possible even without calculating File- and Segment-Statistics and therefor without the need to segment the test recordings. Just with Segment-Probabilities, using the same parameter settings as mentioned above, a score of 91.6% AUC on the private leaderboard can be achieved. A score around 84% is achievable using File- and Segment-Statistics exclusively.

By ranking feature importance returned from the decision trees during training one can find important segments to identify each sound class. Figure 3 and 4 show the ten most important segments to identify the songs of Cetti's Warbler (sound class 11) and Common Chiffchaff (sound class 55). Both sound classes achieve very good classification results with a score close to 100%. Figure 5 gives an example of a sound class with poor classification results. The feature ranking returned from decision trees to identify the call of the European Serin (sound class 66) is partly incorrect and segments are not properly assigned.

To give an idea how well individual species can be identified, a score per sound class is calculated on one third of the training data during 3-fold cross validation. The average of this score is listed and visualized in Table 1.

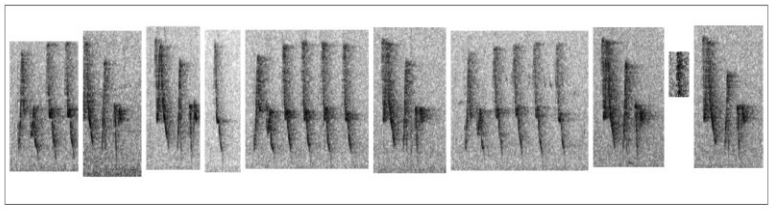


Figure 3: Important segments to identify the song of Cettia cetti (Cetti's Warbler)

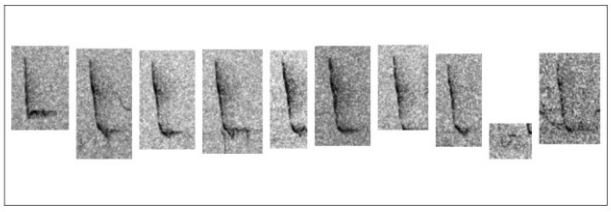


Figure 4: Important segments to identify the song of Phylloscopus collybita (Common Chiffchaff)

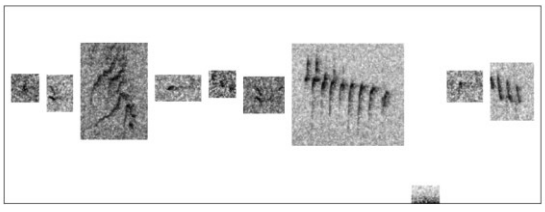


Figure 5: Important segments to identify the call of Serinus serinus (European Serin)

6 Conclusion

This working note describes the winning solution of the NIPS4B 2013 multi-label Bird Species Classification Challenge. The solution of the MLSP 2013 Competition, implemented and described by Fodor, was used as a starting point for further development. The here proposed method includes an efficient way of extracting single sound events and connected sequences of bird calls and syllables in complex acoustic scenes and noisy environments. An ensemble of randomized decision trees is used to learn and predict the binary relevance of each sound class separately with individually selected features per class. The complete source code to reproduce the classification results and additional figures are available at www.animalsoundarchive.org/RefSys/Nips4b2013.php.

Acknowledgments

I would like to thank Prof. Hervé Glotin for organizing this competition, BIOTOPE and ADEME for financing the corpus constitution and kaggle for providing the competition platform. I especially thank Gabor Fodor for documenting his approach and publishing his code for the 2013 MLSP Challenge. I also want to thank Dr. Karl-Heinz Frommolt for supporting my work, sharing his knowledge and providing me with the access to the resources of the Animal Sound Archive [5] at the Museum für Naturkunde Berlin.

References

[1] Fodor G. (2013) The Ninth Annual MLSP Competition: First place. Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on, Digital Object Identifier: 10.1109/MLSP.2013.6661932 Publication Year: 2013, Page(s): 1 - 2

[2] Lewis J.P. (1995) Fast Normalized Cross-Correlation, Industrial Light and Magic

[3] Pedregosa F. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: 2825-2830

[4] Bradski G. (2000) The OpenCV Library. Dr. Dobb's Journal of Software Tools, http://docs.opencv.org/modules/imgproc/doc/object_detection.html

[5] Animal Sound Archive URL: http://www.animalsoundarchive.org/ [25. 11. 2013]