sciendo

Ornis Hungarica 2019. 27(2): 59–66.
DOI: 10.2478/orhu-2019-0015

Automatic bird song and syllable segmentation with an open-source deep-learning object detection method – a case study in the Collared Flycatcher *(Ficedula albicollis)*

Sándor Zsebők[1*], Máté Ferenc Nagy-Egri[2], Gergely Gábor Barnaföldi[2], Miklós Laczi[1,3], Gergely Nagy[1], Éva Vaskuti[1] & László Zsolt Garamszegi[1,4,5]

Received: September 12, 2019 – Revised: October 10, 2019 – Accepted: October 21, 2019

Zsebők, S., Nagy-Egri, M. F., Barnaföldi, G. G., Laczi, M., Nagy, G., Vaskuti, É. & Garamszegi, L. Zs. 2019. Automatic bird song and syllable segmentation with an open-source deep-learning object detection method – a case study in the Collared Flycatcher *(Ficedula albicollis)*. – Ornis Hungarica 2019(2): 59–66. DOI: 10.2478/orhu-2019-0015

Abstract The bioacoustic analyses of animal sounds result in an enormous amount of digitized acoustic data, and we need effective automatic processing to extract the information content of the recordings. Our research focuses on the song of Collared Flycatcher *(Ficedula albicollis)* and we are interested in the evolution of acoustic signals. During the last 20 years, we obtained hundreds of hours of recordings of bird songs collected in natural environment, and there is a permanent need for the automatic process of recordings. In this study, we chose an open-source, deep-learning image detection system to (1) find the species-specific songs of the Collared Flycatcher on the recordings and (2) to detect the small, discrete elements so-called syllables within the song. For these tasks, we first transformed the acoustic data into spectrogram images, then we trained two deep-learning models separately on our manually segmented database. The resulted models detect the songs with an intersection of union higher than 0.8 and the syllables higher than 0.7. This technique anticipates an order of magnitude less human effort in the acoustic processing than the manual method used before. Thanks to the new technique, we are able to address new biological questions that need large amount of acoustic data.

Keywords: bird song, deep-learning, object detection, Collared Flycatcher, automatic segmentation

Összefoglalás Az állati bioakusztikai kutatások jelentős mennyiségű digitalizált hangfelvételt produkálnak, így hatékony automatikus feldolgozási módszerekre van szükség a felvételek információtartalmának kinyerésére. Kutatásunk középpontjában az örvös légykapó *(Ficedula albicollis)* énekének viselkedésökológiai szempontból történő vizsgálata áll. Az elmúlt 20 évben több száz órányi hangfelvételt készítettünk a faj természetes élőhelyén, és ezek feldolgozására automatikus módszereket kerestünk. Tanulmányunkban egy nyílt forráskódú, mélytanulású (deep learning) képdetektálási módszert használtunk az örvös légykapó (1) énekének hangfelvételen belüli megtalálására, és (2) az éneket felépítő egységek, a szillabusok megkeresésére. Mindkét esetben az énekeket spektrogrammá alakítottuk, és két külön modellt tanítottunk be a detektálási feladatokra. Mindkét feladat esetében a módszer ígéretesnek tűnik, jelentősen csökkentve a feldolgozáshoz szükséges emberi időt, ami lehetővé teszi minőségileg új, bioakusztikával kapcsolatos kérdések vizsgálatát.

Kulcsszavak: mélytanulás, örvös légykapó, automatikus szegmentálás, madárének

[1] *Behavioural Ecology Group, Department of Systematic Zoology and Ecology, Eötvös Loránd University, 1117 Budapest, Pázmány Péter sétány 1/C, Hungary*
[2] *Wigner Research Centre for Physics, 1121, Budapest, Konkoly-Thege Miklós út 29-33. Hungary*
[3] *The Barn Owl Foundation, 8744 Orosztony, Temesvári út 8., Hungary*

[4] MTA-ELTE, Theoretical Biology and Evolutionary Ecology Research Group, Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University, 1117 Budapest, Pázmány Péter sétány 1/C, Hungary
[5] Evolutionary Ecology Group, Centre for Ecological Research, Institute of Ecology and Botany, 2163 Vácrátót, Alkotmány utca 2-4. Hungary
* corresponding author: zsebok.s@gmail.com

Introduction

Bird song is an important model for study the ontogeny and evolution of signals and sexual selection (Catchpole & Slater 2008, Vellema *et al.* 2019), therefore it attracts great interest from behavioural ecologists. Furthermore, many faunistic, applied and conservational researches are based on bird song (Laiolo 2010, Borker *et al.* 2015, Zachar *et al.* 2019). Many of these investigations need to collect large amount of acoustic data, where the processing of the recordings may be challenging. Usually, the main steps of the processing are the search of the vocalization of the focal species on the recording, the segmentation of the signals, the extraction of the acoustic features of interest, and the clustering or classification of the elements (Hopp *et al.* 1998). To find automatic processing for all these steps are at the centre of the interest of current research programs (Priyadarshani *et al.* 2018).

One of the most time-consuming steps is the search for the signals in long recordings. Several computer programs were developed to help the researchers to make the manual search easier (Bioacoustics Research Program 2014, Zsebők *et al.* 2018a). Also, several automatic solutions were published based on amplitude or combined amplitude and other acoustic variables like Sound Analysis Pro (Tchernichovski *et al.* 2000) developed for laboratory studies or Luscinia (e.g. Lachlan *et al.* 2018) used in many field studies. Other direction is to use one example of targeted sound and use spectrographic cross-correlation e.g. monitoR package in R (Hafner & Katz 2017). A more sophisticated solution is to build models based on many samples of the targeted vocalization. One of the most promising directions is the deep-learning method based on artificial neural networks, used successfully to detect bat sounds (Mac Aodha *et al.* 2018), identify individuals by their vocalization (Stowell *et al.* 2018), and many models were published in the framework of Bird song Detection Challenge (Stowell *et al.* 2019) to recognize the bird song independently of the species. However, several studies focus only on one species where all the signals have to be found, therefore researchers have to develop a one-species detection method. It can be especially challenging when the vocalization is largely variable like in bird species with large repertoire.

Here, we show how a deep-learning framework can be easily used and tailored by the researchers for one-species detection with complex signals. We chose a ready-made object detection program called 'You Only Look Once' (YOLO) that is developed for object detection in images and videos (Redmon *et al.* 2016, Redmon & Farhadi 2018). YOLO uses deep-convolutional network method, where the dimensions of the input layers can be tailored to the input image size and the characteristics of the network layers can be adjusted to the difficulty of the object detection problem. The idea behind the framework is that the acoustic recordings can be represented as spectrogram images and these images can be fed into the input of YOLO.

Our model species is the Collared Flycatcher *(Ficedula albicollis)* of which vocalization is intensively studied (Haavie *et al.* 2004, Garamszegi *et al.* 2007, 2008, 2012, 2018, Zsebők *et al.* 2017, 2018b). The song is diverse and variable, constitutes of small elements called syllables *(Figure 1)*. Males express 20–90 syllable types (based on 20 songs sampled), and there are large individual differences in the repertoire of syllables (Garamszegi *et al.* 2012). Finding both the songs in the recordings and segmenting the syllables within the songs are time-consuming processes that demand the search for automatic solutions for these steps.

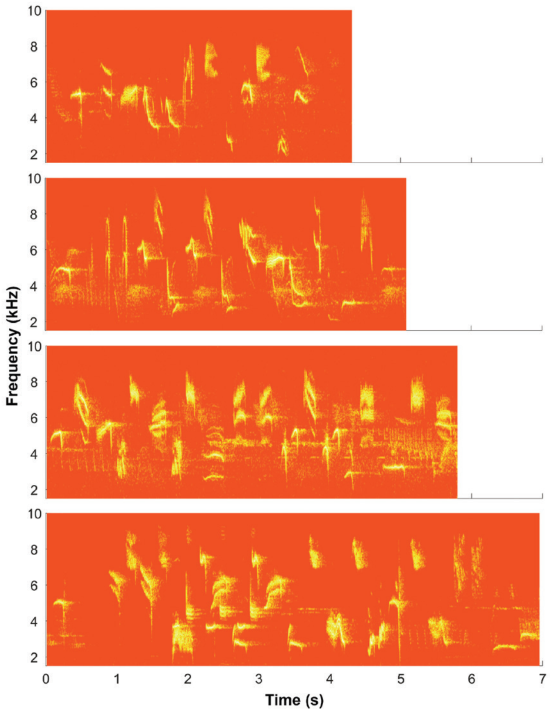

Figure 1. Spectrograms of Collared Flycatcher songs from four different individuals

1. ábra Az örvös légykapó négy különböző egyedétől származó énekének spektrogramja

In this study, our objectives were to build separate models with the YOLO object detection method to identify (1) the songs in the raw acoustic recordings and (2) the syllables within the songs, and evaluate the performance of the two models. We also provide the computer programs that ease the use of YOLO: scripts transforming the sound into images for teaching and testing, extracting the results from the output of the YOLO, and an interactive segmentation tool to verify and correct the mistakes.

Materials and methods

Recordings were obtained in the Pilis-Visegrádi Mountains, Hungary (47°43' N 19°01' E), during the mating period (April–May) between 1999 and 2015. For detailed protocol, see Garamszegi *et al.* (2012). We manually cut the songs out from the recordings and segmented the syllables with the Ficedula Toolbox (Zsebők *et al.* 2018a). We stored the start and end time positions of the songs in the recording file, and the start and end position, minimum frequency and maximum frequency of the syllables within the song file. Altogether we used 3275 songs from 146 recordings for the song database, and 9200 syllables from 9342 songs for the syllable database from our Collared Flycatcher sound library.

We also included sound recordings to our song and syllable database from different bird species frequently appeared during sampling: *Poecile palustris, Cyanistes caeruleus, Parus* *major, Phylloscopus collybita, Sylvia atricapilla, Certhia familiaris, Certhia brachydactyla, Turdus philomelos, Turdus merula, Erithacus rubecula, Fringilla coelebs, Emberiza citrinella.* These recordings were originated from the online sound library of Xeno-canto (xeno-canto.org). The list of sound files and their recorders are provided in the Supplementary *(Table S1)*. Altogether, 390 recordings were used 30 recordings by species.

We built two image libraries, one for the song and another one for the syllable segmentation. For both image libraries, we calculated the spectrograms with 512 FFT (Fast Fourier Transformation) window and 50% overlap. The images contained the spectrograms between 1.5 kHz and 10 kHz frequency, and the resulted images were 300 pixels wide and 150 pixels high.

For the song image library, the images contained 5 s long parts from the recordings. The flycatcher recordings were sampled in a way that the images contained at least 0.1 sec long part of song. The xeno-canto recordings were sampled continuously from the beginning of the recordings by 5 seconds (maximum 10 samples per recordings) without knowing the time information when the given species was vocalized but serving as negative samples without Collared Flycatcher songs. The song image library contained 6831 images, 56% of them contained Collared Flycatcher songs.

The syllable image library contained 1 s parts of the recordings. The images of the flycatcher songs contained at least 1 syllable. The xeno-canto recordings were sampled continuously from the recordings by 1 sec (maximum 50 samples per recording). The syllable image library contained 41229 images, 56% of them contained Collared Flycatcher syllables.

The time information of syllables and songs were provided only for the Collared Flycatcher images. 90% of the images were used as training and 10% as test samples. For the song detection, YOLO model contained 15 layers, and for syllable detection, 31 layers. The learning rate was 0.001 for both models. For the detailed description of the models, see the supplementary files. The models were trained on a GPU (Nvidia GeForce GTX 1080 Ti) in a cluster housed in Wigner GPU Laboratory in the Wigner Research Centre for Physics, Hungarian Academy of Sciences.

The performance of the models was evaluated based on the cross-validation output of the YOLO program, well-known measures in machine learning: recall, average loss and the intersection over union (IOU) (Redmon *et al.* 2016). The calculation of the final mean and standard deviation (SD) of these measures were based on the last 1000 epochs.

All the programs for generating the image library and evaluating the models were written in R environment (R Core Team 2018) with the help of the Seewave package (Sueur et al. 2008). The source codes are freely available on GitHub (https://github.com/zsebok/YOLO).

Results

Both the song and the syllable detection models showed fast learning based on the curves of loss function and IOU *(Figure 2)*. In song detection, in 40,000 epochs, the average loss reached 0.050 ± 0.005 (mean ± *SD*), the recall was 0.978 ± 0.024, and the IOU was 0.809 ± 0.020. In syllable detection, after 80,000 epochs, the average loss was 0.287 ± 0.019, the

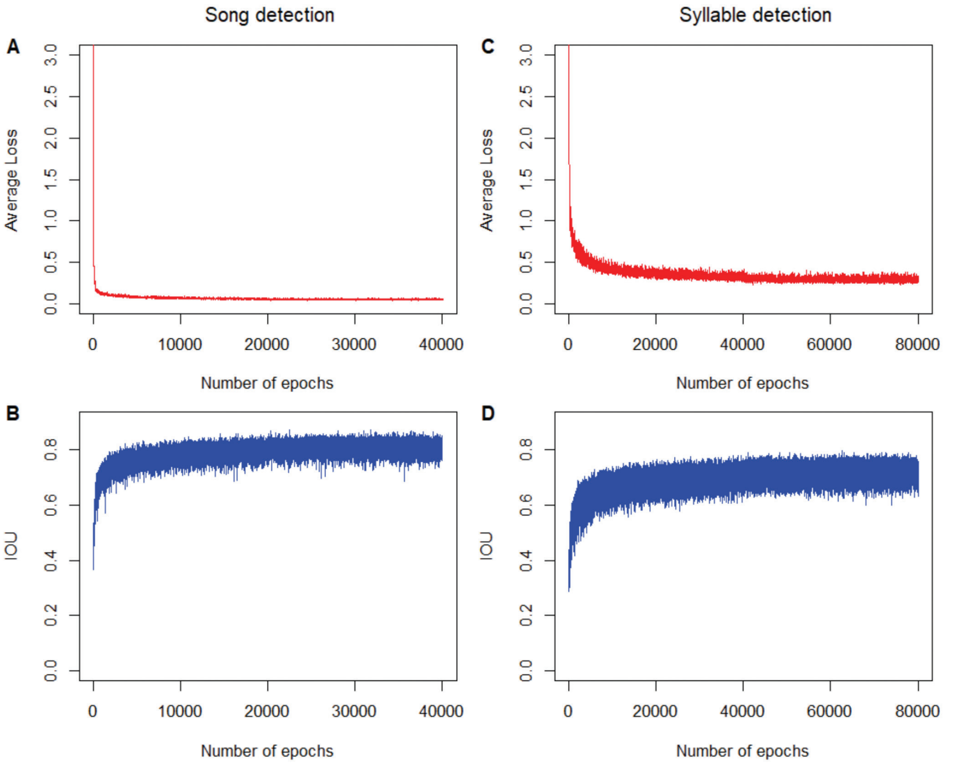

Figure 2. Evaluation of the song (A, B) and the syllable (C, D) detection models through the learning process

2. ábra Az ének (A, B) és szillabus (C, D) detektáló modellek tanulási görbéi

recall was 0.906 ± 0.035, and the IOU was 0.722 ± 0.025. After the training, the visual inspection of the segmentation rectangular seemed acceptable at both the detected songs and syllables on test sounds *(Figure 3, 4)*.

Discussion

According to the visual inspection of the object detection results, the song and syllable segmentation looks promising showing no large error. However, in the syllable segmentation, the mean IOU that is lower than the human inter-observer IOU (0.84 ± 0.17, unpublished results) seems sufficient to identify the syllables and perform automatic measurements on them. In general, IOU over 70% is taken as good performance in object detection (Rahman & Wang 2016, Redmon *et al.* 2016), and both song and syllable detection reached that limit.

In line with the previous publications (Mac Aodha *et al.* 2018, Stowell *et al.* 2019), we also found that the deep-learning technique with convolutional layers can cope with the highly variable acoustic signals and indicate a promising method for segmenting the acoustic recordings to significantly decrease the processing time by the human observers. The disadvantage of

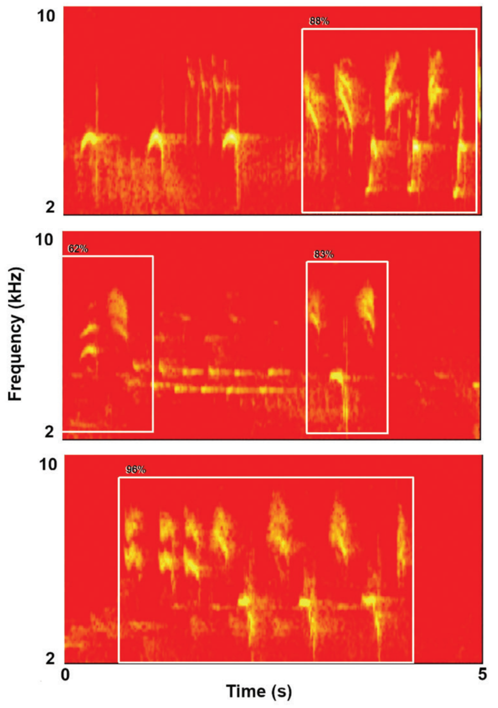

Figure 3. Example images of the song detection. The numbers above the detected songs are representing the Intersection Over Union values

3. ábra Spektrogramon ábrázolt példák az énekdetektálás bemutatására. A detektált énekek fölötti számok az IOU értéket mutatják

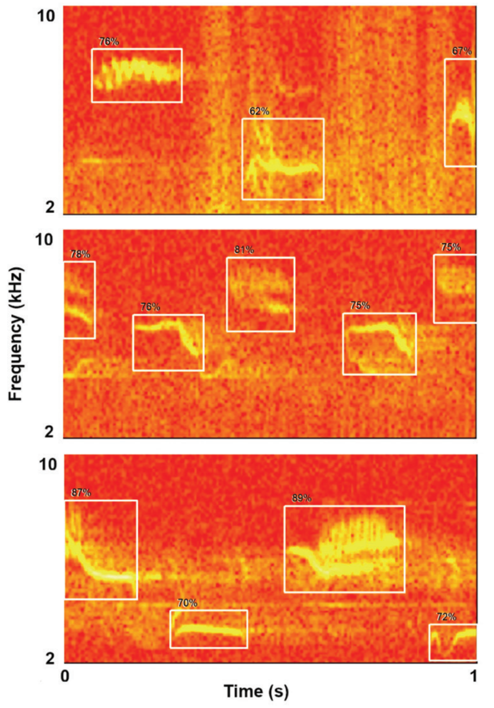

Figure 4. Example images of the syllable detection. The numbers above the detected syllables are representing the Intersection Over Union values

4. ábra Spektrogramon ábrázolt példák a szillabus detektálás bemutatására. A detektált szillabusok fölötti számok az IOU értéket mutatják

using deep-learning method is the need of huge amount of previously segmented recordings to train and test the models. It can be feasible, like in our research program, where long-term or broad-scale investigations can already provide such data. However, large acoustic libraries (like xeno-canto.org) can be a great help in the start of building such datasets.

Here, we showed that for a one-species recognition, a free and open-source object detection program like YOLO developed for image and video processing can be used effectively. With the scripts written in R and provided as a supplementary to this paper, a user without much knowledge is able to build acoustic bird detector for specific species. It is worth to mention that YOLO is able to detect objects belongs to multiple classes (Redmon *et al.* 2016), thus our suggested framework can be broaden to multiple species in birds and other animals.

To further increase the general usefulness of our models for the detection of Collared Flycatcher and other bird species' songs and syllables, it is worth to apply data augmentation technique by using artificially prepared recordings with different background noises (Stowell *et al.* 2018). Also, to use such models on more broad geographical range, further recordings from other populations are needed in the training phase of the process.

Acknowledgement

We are grateful to the members of the Behavioural Ecology Group (Department of Systematic Zoology and Ecology, Eötvös Loránd University, Budapest). The project was supported by the National Research, Development and Innovation Office, Hungary (NKFIH K-115970, K-129215, PD-115730) and the Pilisi Parkerdő Zrt.

Supplementary material

Table about the used sound recordings downloaded from xeno-canto.org (available on the homepage of Ornis Hungarica)
Táblázat a xeno-canto.org oldalról letöltött hangfelvételekről (elérhető az Ornis Hungarica honlapjáról)

References

Bioacoustics Research Program 2014. Raven Pro: Interactive Sound Analysis Software (Version 1.5) [Computer software]. – Ithaca, NY: The Cornell Lab of Ornithology Available from http://www.birds.cornell.edu/raven.

Borker, A. L., Halbert, P., McKown, M. W., Tershy, B. R. & Croll, D. A. 2015. A comparison of automated and traditional monitoring techniques for marbled murrelets using passive acoustic sensors. – Wildlife Society Bulletin 39: 813–818. DOI: 10.1002/wsb.608

Catchpole, C. K., Slater, P. J. B. 2008. Bird song: biological themes and variations, 2nd ed. – Cambridge University Press, Cambridge

Garamszegi, L. Zs., Eens, M. & Török, J. 2008. Birds Reveal their Personality when Singing. – PLoS One 3(7). DOI: 10.1371/journal.pone.0002647

Garamszegi, L. Zs., Török, J., Hegyi, G., Szöllõsi, E., Rosivall, B. & Eens, M. 2007. Age-dependent expression of song in the Collared Flycatcher, *Ficedula albicollis*. – Ethology 113: 246–256. DOI: 10.1111/j.1439-0310.2007.01337.x

Garamszegi, L. Zs., Zagalska-Neubauer, M., Canal, D., Blazi, Gy., Laczi, M., Nagy, G., Szőllősi, E., Vaskuti, É. Török, J. & Zsebők, S. 2018. MHC-mediated sexual selection on birdsong: Generic polymorphism, particular alleles and acoustic signals. – Molecular Ecology 27: 2620–2633. DOI: 10.1111/mec.14703

Garamszegi, L. Zs., Zsebők, S. & Török, J. 2012. The relationship between syllable repertoire similarity and pairing success in a passerine bird species with complex song. – Journal of Theoretical Biology 295: 68–76. DOI: 10.1016/j.jtbi.2011.11.011

Haavie, J., Borge, T., Bures, S., Garamszegi, L. Zs., Lampe, H. M., Moreno, J., Qvarnström, A., Török, J. & Sætre, G. P. 2004. Flycatcher song in allopatry and sympatry – Convergence, divergence and reinforcement. – Journal of Evolutionary Biology 17: 227–237. DOI: 10.1111/j.1420-9101.2003.00682.x

Hafner, S. D. & Katz, J. 2017. {monitoR}: Acoustic template detection in R. Retrieved from http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm

Hopp, S. L., Owren, M. J. & Evans, C. S. 1998. Animal acoustic communication: sound analysis and research methods. – Springer-Verlag Berlin Heidelberg

Lachlan, R. F., Ratmann, O. & Nowicki, S. 2018. Cultural conformity generates extremely stable traditions in bird song. – Nature Communications 9: 2417. DOI: 10.1038/s41467-018-04728-1

Laiolo, P. 2010. The emerging significance of bioacoustics in animal species conservation. – Biological Conservation 143: 1635–1645. DOI: 10.1016/j.biocon.2010.03.025

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G. & Jones, K. E. 2018. Bat detective-Deep learning tools for bat acoustic signal detection. – PLoS Computational Biology 14: 1–19. DOI: 10.1371/journal.pcbi.1005995

Priyadarshani, N., Marsland, S. & Castro, I. 2018. Automated birdsong recognition in complex acoustic environments: a review. – Journal of Avian Biology 49(5): 1–27. DOI: 10.1111/jav.01447

R Core Team 2018. *R*: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria – Available online at https://www.R-project.org/

Rahman, M. A. & Wang, Y. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. – Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10072 LNCS: 234–244. DOI: 10.1007/978-3-319-50835-1_22

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. Retrieved from https://arxiv.org/abs/1506.02640v5

Redmon, J. & Farhadi, A. 2018. YOLOv3: An Incremental Improvement. – Retrieved from http://arxiv.org/abs/1804.02767

Stowell, D., Petrusková, T., Šálek, M. & Linhart, P. 2018. Automatic acoustic identification of individual animals: Improving generalisation across species and recording conditions. – Retrieved from http://arxiv.org/abs/1810.09273

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y. & Glotin, H. 2019. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. – Methods in Ecology and Evolution 10: 368–380. DOI: 10.1111/2041-210X.13103

Sueur, J., Aubin, T. & Simonis. C. 2008. Seewave, a Free Modular Tool for Sound Analysis and Synthesis. Bioacoustics The International Journal of Animal Sound and its Recording 18:213–226. DOI: 10.1080/09524622.2008.9753600

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. 2000. A procedure for an automated measurement of song similarity. – Animal Behaviour 59: 1167–1176. DOI: 10.1006/anbe.1999.1416

Vellema, M., Diales Rocha, M., Bascones, S., Zsebők, S., Dreier, J., Leitner, S., Van der Linden, A., Brewer, J. & Gahr, M. 2019. Accelerated redevelopment of vocal skills is preceded by lasting reorganization of the song motor circuitry. – Elife 8: 1–46. DOI: 10.7554/elife.43194

Zachar, G., Tóth, A. S., Gerecsei, L. I., Zsebők, S., Ádám, Á. & Csillag, A. 2019. Valproate exposure in ovo attenuates the acquisition of social preferences of young post-hatch Domestic Chicks. – Frontiers in Physiology 10: 881. DOI: 10.3389/fphys.2019.00881

Zsebők, S., Blázi, G., Laczi, M., Nagy, G., Vaskuti, É. & Garamszegi, L. Zs. 2018a “Ficedula”: an open-source MATLAB toolbox for cutting, segmenting and computer-aided clustering of bird song. – Journal of Ornithology 159: 1105–1111. DOI: 10.1007/s10336-018-1581-9

Zsebők, S., Herczeg, G., Blázi, G., Laczi, M., Nagy, G., Török, J. & Garamszegi, L. Zs. 2018b Minimum spanning tree as a new, robust repertoire size comparison method: simulation and test on birdsong. – Behavioral Ecology and Sociobiology 72: 48. DOI: 10.1007/s00265-018-2467-9

Zsebők, S., Herczeg, G., Blázi, G., Laczi, M., Nagy, G., Szász, E., Markó, G., Török, J. & Garamszegi, L. Zs. 2017. Short- and long-term repeatability and pseudo-repeatability of bird song: sensitivity of signals to varying environments. – Behavioral Ecology and Sociobiology 71: 154. DOI: 10.1007/s00265-017-2379-0