Blind source separation-based IVA-Xception model for bird sound recognition in complex acoustic environments

Yusheng Dai,[1] Jin Yang,[1,✉] Yiwei Dong,[2]
Haipeng Zou,[3] Mingzhi Hu,[1] and Bin Wang[4]

[1] *School of Cyber Science and Engineering, Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu 610065, People's Republic of China*

[2] *College of Mathematics, Sichuan University, Chengdu, People's Republic of China*

[3] *College of Software Engineering, Sichuan University, Chengdu, People's Republic of China*

[4] *School of Information Science and Technology, Southwest Jiaotong University, People's Republic of China*

✉Email: jinyangscu@163.com

Identification of bird species from audio recordings has been a major area of interest within the field of ecological surveillance and biodiversity conservation. Previous studies have successfully identified bird species from given recordings. However, most of these studies are only adaptive to low-noise acoustic environments and the cases where each recording contains only one bird's sound simultaneously. In reality, bird audios recorded in the wild often contain overlapping signals, such as bird dawn chorus, which makes audio feature extraction and accurate classification extremely difficult. This study is the first to focus on applying a blind source separation method to identify all foreground bird species contained in overlapping vocalization recordings. The proposed IVA-Xception model is based on independent vector analysis and convolutional neural network. Experiments on 2020 Bird Sound Recognition in Complex Acoustic Environments competition (BirdCLEF2020) dataset show that this model could achieve a higher macro F1-score and average accuracy compared with state-of-the-art methods.

Introduction: Recent developments in the field of ecoacoustics have led to an increasing interest in audio-based species identification. Many ecoacoustics scientists have collected abundant acoustic information in the wild through remote sensing techniques and performed further soundspace analysis to explore potential ecological information [1, 2]. For example, Sumitani et al. used robot audition techniques to monitor spatio-spectro-temporal dynamics of bird vocalizations [3]. The audio based strategy has also been extensively adopted in automatic bird species recognition [4].

However, it is challenging to identify bird species accurately in a complex soundscape that contains various background noises. In recent years, machine learning [5, 6] and neural network [7–9] based methods have shown impressive performance in bird sound classification, whereas most of the previous work has been limited to scenarios where the recordings are carefully selected and of low noise [4]. In the meantime, studies aimed at complex acoustic environments, such as data augmentation methods [10, 11], spectral subtraction methods [9], and wavelet denoising methods [12], have paid too little attention to bird sound recognition in the case of overlapping vocalizations. Overlapping sound signals commonly exist in unattended field recordings. Allied bird species have similar acoustic characteristics and the overlapped signals interfere with each other in both time and frequency domain, which greatly affects ecological interpretations [1]. In this case, it is extremely difficult to eliminate these interfering sound sources without losing syllable information based on traditional noise reduction methods [4]. To date, it remains to be investigated how to identify bird species from audio recordings that contain multiple kinds of bird sounds at the same time [4, 8, 13].

Blind source separation is a powerful technique to recover the original source signals when only signal mixtures are accessible. In the past few years, this technology has been gradually used in the fields of wildlife monitoring and biodiversity assessment. Compared with model-based source separation methods, it requires a less amount of data and can perform well without the aid of prior information [1]. This letter is novel in that, it fills a gap in the research on overlapping bird sounds recognition from a source signal separation perspective. We take advantage of today's multi-channel recording devices, and establish the IVA-Xception model which can achieve high performance in identifying all foreground birds from overlapping bird sound recordings based on IVA and CNN.

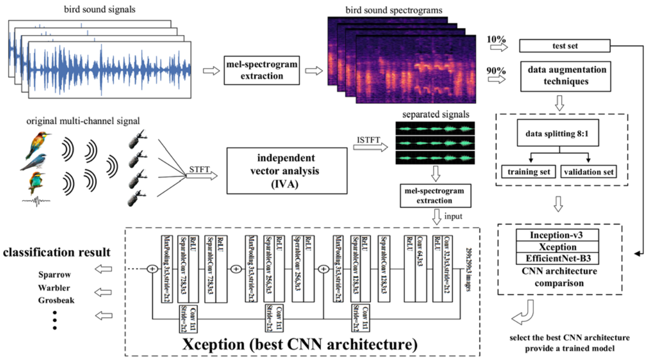


Fig. 1 *Pipeline of IVA-Xception model*

Pipeline: In this letter, we propose a robust audio-based bird identification model called IVA-Xception, which can identify all foreground bird species from given audios and achieve high accuracy. As Figure 1 shows, in a complete bird recognition process, the model first uses IVA [14] in the frequency domain to separate source signals from the original multi-channel signal. Then we utilize the CNN that has been trained to extract features from the converted spectrograms. Finally, the Softmax classifier is adopted to obtain the identification result. For CNN architecture selection, we utilize Xception [15] as the adaptive CNN architecture for our system after comparing neural networks' performance on spectrogram feature extraction. We also apply data augmentation techniques to converted spectrograms to improve the robustness of the system and solve the data imbalance problem.

Source signal separation: The signal recorded by a microphone array unit can be described as a convolution operation process of discrete signals from different sources, which can be expressed as

$$\hat{x}_m[t] = \sum_{k=1}^{K} (\hat{a}_{mk} \star \hat{s}_k)[t], \tag{1}$$

where $\hat{x}_m[t]$ is the m-th microphone unit signal, $\hat{s}_k[t]$ is the k-th source signal, and $\hat{a}_{mk}[t]$ is the impulse response between the two. The operator $\star$ denotes convolution. In our model, we applied IVA to separate K bird sound sources from M-channel recordings in the frequency domain. In contrast to the conventional blind source separation method, independent component analysis (ICA), IVA does not suffer the frequency permutation problem [14]. Meanwhile, it can yield a rather satisfying separation result. It should also be pointed out that the number of channels of today's recording devices is typically large; for example, SWIFT developed by Cornell Lab of Ornithology [8] has 64 recording units, which generally exceeds the number of bird sound sources. In addition, adding extra microphones can improve the performance of source separation. Thus, we assume $M \geqslant K$ and model the short-time Fourier transforms (STFT) of the recorded multi-channel signal as Equation (2).

$$\boldsymbol{x}(f,t) = A_s(f)\boldsymbol{s}(f,t) + A_z(f)\boldsymbol{z}(f,t) \in \mathbb{C}^M, \tag{2}$$

where $\boldsymbol{x}(f,t) = [x_1(f,t), \ldots, x_M(f,t)]^\top \in \mathbb{C}^M$ denotes the recorded signal; $\boldsymbol{s}(f,t) = [s_1(f,t), \ldots, s_K(f,t)]^\top \in \mathbb{C}^K$ and $\boldsymbol{z}(f,t) \in \mathbb{C}^{M-K}$ are the source and noise signals, respectively; $f \in \{1, \ldots, F\}$ denotes the frequency bin; and $t \in \{1, \ldots, T\}$ denotes the time-frame index. For mixing matrices, $A_s(f) \in \mathbb{C}^{M \times K}$ is the mixing matrix for bird sound sources, and $A_z(f) \in \mathbb{C}^{M \times (M-K)}$ is that for background noise. The objective of IVA is to estimate the demixing matrix $\boldsymbol{W}(f) = [\boldsymbol{w}_1(f), \ldots, \boldsymbol{w}_M(f)] \in \mathbb{C}^{M \times M}$ satisfying

$\boldsymbol{W}(f)^{\mathrm{H}}[A_s(f), A_z(f)] = I_M$. Eventually, the source vector $\boldsymbol{s}(f, t)$ can be recovered by multiplying $\boldsymbol{W}(f)$ by the observation signal $\boldsymbol{x}(f, t)$

$$s_k(f, t) = \boldsymbol{w}_k(f)^{\mathrm{H}}\boldsymbol{x}(f, t) \in \mathbb{C}, \quad k \in \{1, \dots, K\}$$

$$\boldsymbol{z}(f, t) = W_z(f)^{\mathrm{H}}\boldsymbol{x}(f, t) \in \mathbb{C}^{M-K}$$

$$W_z(f) = [\boldsymbol{w}_{K+1}(f), \dots, \boldsymbol{w}_M(f)] \in \mathbb{C}^{M\times(M-K)}$$

where I_M is the identity matrix and ^{H} is the Hermitian transpose. It is also noteworthy that there is often no need to separate the noise components. Some assumptions must be met to make the estimation possible. The most crucial one is that subcomponents $s_1(f, t), \dots, s_K(f, t)$ must be statistically independent and non-stationary. Armed with these preliminaries, we applied the OverIVA algorithm [16] to learn the demixing matrix $\boldsymbol{W}(f)$. OverIVA algorithm, whose object is minimizing the negative log-likelihood of the observed data, takes advantage of the iterative projection technique and reduces the computational cost theoretically. It is superior to the most commonly used IVA algorithms in terms of convergence speed and effectiveness.

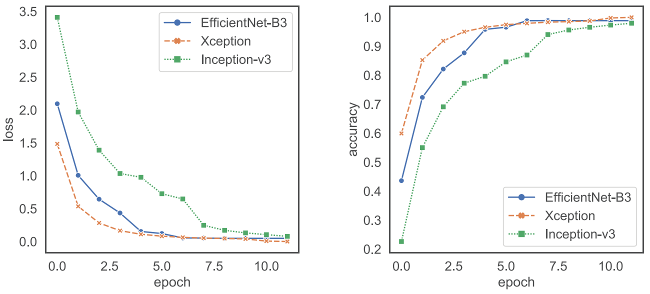
Spectrogram conversion and selection: In this part, each separated single-channel audio was converted to a log-amplitude Mel-spectrogram, and then spectrograms which are from bird sounds rather than noises were selected. Specifically, we first chopped the audio signals into 1-s chunks and transformed these chunks into log-amplitude Mel-spectrograms. We set the pre-emphasis factor to 0.95 and a frequency range of approximately 900 to 15,100 Hz. Each spectrogram size was set to be 299×299 px. However, we found that the obtained spectrograms contained many noise spectrograms, where there were little or almost negligible bird sound syllables. These noise spectrograms could be misleading when training the CNN, as they are labelled as bird sounds. Therefore, we conducted spectrogram selection work referred to prior work [6] by estimating the signal-to-noise ratio (SNR) of the spectrogram because spectrograms with a higher SNR are more likely to contain bird sounds.

CNN network training and species identification: This study utilized a CNN to extract features from spectrograms. During data preprocessing, we noticed the data imbalance problem after the spectrogram conversion process. The maximum number of spectrograms of certain bird species reaches up to 6924, yet the minimum is only 146. Referred to work [10, 11], the following data augmentation methods were used to solve this problem: horizontal shift, adjusting brightness, adjusting contrast, adding Gaussian noise, adding background noise, random pixel dropout, color space augmentations, volume shift, and pitch shift. As a result, We made the number of spectrograms of each species to be 1500. For the selection of CNN architecture, we conducted experiments to pick the most suitable one from three CNNs that achieved excellent performances in past image classification works: Inception-v3 (ranked first at BirdCLEF2019) [17], Xception (ranked second at BirdCLEF2020), and EfficientNet-B3 [18]. The experiment results suggested that Xception performs best among the three architectures, which is elaborated in the experiment validation section. It is worth mentioning that as we previously chopped the audios into 1-s chunks, each separated single-source audio generated multiple spectrograms, and each spectrogram had its own classification result. We defined a principle of how to interpret classification results for audio chunks based on classification results of its converted spectrograms: we arranged the predicted probability values for each spectrogram in descending order, like $p_1 > p_2 > \dots > p_i$, and set a dynamic threshold λ to describe the reliability of prediction result. Once $p_1 - p_2 > \lambda$, we put the prediction result of the corresponding spectrogram into a "confidence list," otherwise it was discarded. Then the final prediction probability distribution was produced by averaging prediction results of the spectrograms in the confidence list.

Experimental validation: The audio data used in this experiment is the official dataset provided by the BirdCLEF2020 competition [19]. The original data set comes from the Xeno-Canto community and contains nearly 80,000 recordings, covering more than 1500 birds. In particular, the audio contains both stationary and non-stationary background noises.

Table 1. Performance of the three CNNs on the test set

CNN architecture	F1-score (%)	Accuracy(%)
EfficientNet-B3	87.49	87.83
Xception	90.85	90.62
Inception-v3	81.37	80.97



(a) Change of loss (b) Change of accuracy

Fig. 2 *Change of loss and accuracy with epoch during training for three CNNs*

Two experiments were conducted in this letter. The first experiment was aimed to select an adaptive CNN architecture. For each model, we processed the spectrogram data with the same preprocessing methods and evaluated the network's performance on bird sound spectrogram classification according to macro F1-score and average accuracy. We selected 50 species of birds in this experiment and split the audio data of these birds into training, validation, and test sets at an 8:1:1 ratio. We converted the audio files into spectrograms according to the methods described in previous sections. After applying data augmentation methods to the training and validation set, the number of spectrograms for each bird in the training set was 1500, and the number of spectrograms for both the validation and test sets was 190. After completing CNN training, we evaluated the model's performance with the test set. The experimental results are given in Table 1.

As Table 1 shows, Xception obtained the best spectrogram classification result, and the F1-score score on the test set was 3.36% higher than EfficientNet-B3 and 9.48% higher than Inception-v3. It also obtained the best accuracy score. Moreover, Xception showed a faster convergence speed during the training process, as is shown in Figure 2.

In the second experiment, we explored the effect of IVA-based signal separation on the identification performance, and tested the model robustness when interference sources were added. From the primary 50 species, we randomly selected five bird species:Great Reed Warbler, Eurasian Reed Warbler, Long-tailed Tit, Black-throated Sparrow, and Slate-colored Grosbeak. Then, the single-channel audios of these species provided by BirdCLEF2020 were first unified to 8 s, 44,100 Hz, 32 bit with the same power. To simulate real recording environments, we set up a complex soundscape with overlapping vocalizations using the image source method implemented in the pyroomacoustics Python package [20]. Concretely, we simulated a 20 m×20 m×10 m 3-dimensional space with a reverberation time of 0.2 s, set the relative humidity of the air to be 50%, and set the temperature to be 20°C. The absorption of sound energy by the air was also simulated. For spatial relationship, target sources were placed in different directions at a distance of 6 to 7 m from the microphone array and 6 m from the ground. The microphone array units were placed on a fan-shaped area of radius 3 cm centred at [13, 10, 3.5] in the same height. To enhance the robustness of the proposed model, we also added sounds of other species as interference sources, which distributed randomly in a cuboid with x-axis spanning from 16 to 20, y-axis spanning from 2 to 20, and z-axis spanning from 6 to 9. As for the relationship of power between the target sources and interference sources, we referred to work [16]. The overall setup is illustrated in Figure 3.

During the experiment, we simulated two target sources and three target sources cases, denoted as two-birds and three-birds, respectively.

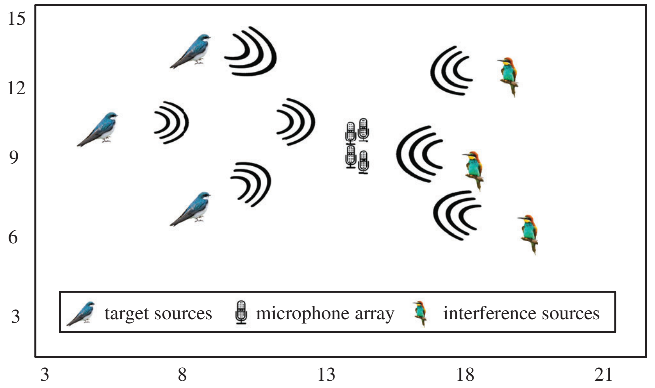


Fig. 3 *Vertical view of the simulated soundscape setup*

*Table 2. Performance comparison of IVA-Xception and Xception. The symbol * indicates sounds of other bird species are added as interference. For the 2-birds case, two recording units were used, and for 3-birds case, four recording units were used*

Num. of species	Method	F1-score (%)	Accuracy (%)
2-birds	Xception	75.38	64.80
	IVA-Xception	86.58	80.20
	IVA-Xception*	86.96	80.40
3-birds	Xception	65.21	53.07
	IVA-Xception	80.46	69.00
	IVA-Xception*	80.00	68.20

Specifically, for the two-birds case, the experiment traversed $C_5^2 = 10$ combinations. For each combination, we generated 50 audio files randomly, obtaining a total of 500 test audios. Similarly, in three-birds case, the experiment traversed $C_5^3 = 10$ combinations, and we also obtained 500 test audios. After completing the simulation process, we collected signals recorded by the microphone array and then input them into the models to identify all of the foreground species. To explore the effect of IVA-based signal separation on the identification performance, we compared the performance of our proposed IVA-Xception model with the Xception model without the source separation process. The experiment results are shown in Table 2.

As Table 2 shows, the F1-score and accuracy of the IVA-Xception were significantly higher than Xception in both the 2-birds case and the 3-birds case. In other words, our proposed model shows a generally improved classification performance of 10% to 16% over the state-of-the-art Xception model. Therefore, it is demonstrated that blind source separation indeed contributes to identification performance. Moreover, IVA-Xception yields stable results even in scenarios with interference, reflecting strong robustness.

Conclusion: This is the first study that set out to identify all foreground bird species from overlapping vocalizations audio recordings using a blind source separation method. Before this study, it was difficult to make accurate classifications in scenarios like dawn chorus. The established IVA-Xception identification model is robust and shows excellent performance. It is validated by experiments that Xception has the best ability of spectrogram feature extraction, the IVA-Xception model is superior to state-of-the-art models in the bird sound recognition area, and our proposed model is robust against all possible realizations of the modelled uncertainty. Further research could also be conducted to examine how to identify bird species in under-determined states and single-channel overlapping vocalization audio recordings. Finally, all source codes used in this letter could be accessed on GitHub [21].

Acknowledgements: This work was supported by the National Natural Science Foundation of China (61872254), the Key Lab of Information Network Security of Ministry of Public Security (C20606), and the Sichuan Science and Technology Program (21CXRC0103). We want to convey our grateful appreciation to the corresponding author of this paper, Jin Yang. He has offered advice with huge values in all stages when writing this essay to us.

Received: *27 January 2021* Accepted: *16 March 2021*
doi: 10.1049/ell2.12160

References

1 Lin, T.H., Tsao, Y.: Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval. *Remote Sensing in Ecology and Conservation*, **6**(3), 236–47 (2020)

2 Farina, A., et al.: *Ecoacoustics: The Ecological Role of Sounds*. John Wiley & Sons, New York (2017)

3 Sumitani, S., et al.: Fine-scale observations of spatio-spectro-temporal dynamics of bird vocalizations using robot audition techniques. *Remote Sensing in Ecology and Conservation* **7**, 18–35 (2020)

4 Priyadarshani, N., et al.: Automated birdsong recognition in complex acoustic environments: A review. *Journal of Avian Biology*, **49**(5), jav-01447 (2018)

5 Lasseck, M.: Bird song classification in field recordings: winning solution for NIPS4B 2013 competition. In: Proceedings of the International Symposium on Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, pp. 176–181 (2013)

6 Zhao, Z., et al.: Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, **39**, 99–108 (2017)

7 Himawan, I., et al.: 3D convolution recurrent neural networks for bird sound detection. In: Proceedings of the 3rd Workshop on Detection and Classification of Acoustic Scenes and Events 2018, pp. 1–4

8 Kahl, S., et al.: Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. *CLEF* (2020)

9 Bai, J., et al.: Xception based method for bird sound recognition of BirdCLEF 2020. *CLEF Working Notes* (2020)

10 Sprengel, E., et al.: Audio based bird species identification using deep learning techniques. CLEF Working Notes, pp. 547–559.Springer, Cham, Switzerland (2016)

11 Fazeka, B., et al.: A multi-modal deep neural network approach to birdsong identification [Internet]. arXiv [cs.SD]. (2018)

12 Priyadarshani, N., et al.: Birdsong denoising using wavelets. *PloS One* **11**(1), e0146790 (2016)

13 Goëau, H., et al.: Overview of BirdCLEF 2018: Monospecies vs. Soundscape bird identification. In: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum, Sep 10, 2018, Avignon, France (No. 2125)

14 Kim, T., et al.: Independent vector analysis: An extension of ICA to multivariate components. In: International Conference on Independent Component Analysis and Signal Separation, March 2006, pp. 165–172.Springer, Berlin, Heidelberg (2006)

15 Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258 (2017)

16 Scheibler, R., Ono, N.: Independent vector analysis with more microphones than sources. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 185–189.IEEE (2019)

17 Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)

18 Tan, M., et al.: EfficientNet: Rethinking model scaling for convolutional Neural Networks [Internet]. arXiv [cs.LG]. (2019)

19 [dataset] The organization of the BirdCLEF task; 2020; LIFECLEF 2020 BIRD DATASET; https://www.aicrowd.com/clef_tasks/22/task_dataset_files?challenge_id=211

20 Scheibler, R., et al.: Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 351–355

21 [dataset] Yusheng Dai and Haipeng Zhou; 2021; IVA-Xception; v1.0; https://github.com/dalision/IVA-Xception/