Developing Birds Sound Recognition System Using an Ontological Approach

Yauheniya Zianouka and Dzianis Bialiauski
Liesia Kajharodava and Aliaksandr Trafimau
Vitalij Chachlou and Juras Hetsevich
United Institute of Informatics Problems
National Academy of Sciences of Belarus
Minsk, Belarus
{evgeniakacan, dzianis.bialiauski,
lesia.cordell, cncntrt,
vitalikhokhlov, yuras.hetsevich}@gmail.com

Vadim Zahariev and Kuanysh Zhaksylyk
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
zahariev@bsuir.by, kuanysh.zhk@gmail.com

***Abstract*—The article presents an intelligent model of automated voice recognition systems (on the example of birds). To develop it, a dataset of birds' voices was annotated and processed using Mel-Frequency Cepstral Coefficient as an effective tool for modelling the subjective pitch and frequency content of audio signals. For composing and training the model, Convolutional Neural Network is used to implement high level results. The possibilities of using ontological approaches and OSTIS technology for further improvement of the quality of ML models are shown.**

***Keywords*—recognition system, machine learning, dataset, automatic processing, Mel-frequency cepstral coefficients (MFCCs), Convolutional Neural Network, EfficientNet, ontological approach**

I. Introduction

There are many voice signals in nature, and each type of sound signal has its own function. In general, sound signals can be divided into singing and other voices (for example, streams, noise). Singing is a more melodic type of bird voice. It is usually longer and more complex than the stream. Due to many variations of the sound signals of different bird species, there is a problem of their recognition. The relevance of creating such a system is due to the fact that all existing developments on the recognition of animal species are not suitable for Belarusian birds. Only some of them are able to recognize the sound signals of European species, which also affect our project. However, the development and use of an automated system facilitates the recognition process, the system itself has recognition accuracy problems that need to be mentioned. Software for determining the species of animals (for example, birds) by voice signals is based on mathematical calculation models (most often twisted neural networks), which, with insufficient training, can make a computational error that leads to incorrect determination of the biological species we need [1]. Thus, the tasks were set:

- to develop methodological foundations for collecting, annotating and recognizing animal voice signals on the territory of Belarus (in terms of technical implementation).
- to compose a structural scheme for automated recognition of animal voice signals for autonomous continuous monitoring of rare, threatened species and indicator species.
- to increase the accuracy of recognition of animal species (for example, birds).

II. Dataset for training the recognition model

A dataset of electronic voice signals corpora is a substantial component for training the recognition model. The primary source of publicly available arrays of animal vocalisations used in the dataset is the Xeno-Canto (http: https://www.xeno–canto.org). Its resources are available for listening, downloading, and studying the characteristics of sound recordings. It is one of the largest sources of audio data of bird vocalisations collected from around the world. The site has API endpoints that can be used to automatically search, download data by scientific or common name of a species or family, region tags, sound types, country, etc.

Machine learning is heavily dependent on data. That's why one of the main tasks that need to be performed during its preparation is the annotation of audio recordings. When developing the method of data annotation for bird voice recognition systems, it is taken into account that the composition of the Belarusian fauna includes rare species of animals and birds, for which there is no or limited scope of annotated sounds and graphic data [2]. It is also considered that the selected data may contain different sounds of the same species, sounds of many other species that are heard together with this species.

In order to improve the results of processing and recognizing birds' voices, each individual audio file is

labelled with the name of its own species. The data downloaded during the collection phase from open sources may include some part of the annotated data, but needs to review and fix the annotation according to the audio event classes. Each audio part of the signal is detected as a silent audio segment or an audio event. Then it is assigned to its appropriate audio subclass based on the nature of its content. Next, all the annotated information with the corresponding timestamps is written to a text file markup. Since the data is mostly recorded in the birds' natural habitats, each recording may also contain some background information, including various noises from other animals or people. In order to adjust and improve the performance of the recognition algorithm, it is recommended to add and annotate recordings in the base for training the algorithm, where there are no bird sounds, but there is a background sound of the environment in which certain species of birds usually exist.

In our work, the system for recognizing the voices of birds was built on fourteen species: Parus Major (Sinica vialikaja), Fringilla Coelebs (Bierascianka), Turdus Philomelos (Drozd-spiavun), Emberiza Citrinella (Strynatka zvyčajnaja), Phylloscopus Collybita (Piačuraŭka-cieńkaŭka), Turdus Merula (Drozd čorny), Sylvia Atricapilla (Lieska-čornahaloŭka), Luscinia Luscinia (Salaviej uschodni), Acrocephalus Dumetorum (Čarotaŭka sadovaja), Erithacus Rubecula (Malinaŭka), Loxia Curvirostra (Kryžadziub-jalovik), Hippolais Icterina (Pierasmieška), Periparus Ater (Sinica-maskoŭka), Sylvia Communis (Lieska šeraja). Using the API (Application Programming Interface) a dataset was collected (audio recordings) according to the above mentioned bird species. The number of entries was about two hundred for each species. The criteria by which records were selected for training were as follows:

- The duration of audio recordings is more than three seconds and less than 10 minutes.
- Proximity by distance to Belarus (Minsk).

The following information is available for each uploaded audio file: Bird species (Specific epithet); Bird subspecies of the (subspecific epithet); The group to which the species belongs (bird, grasshopper); The name of the species in English; The name of the person who made the audio recording; The country where the sound was recorded; The name of the area where the recording was conducted; Geographical latitude of the place where the recording was conducted; Geographical longitude of the place where the recording was made; The type of sound that a bird makes (singing, streaming, etc.); Gender (female or male), etc.

Currently, a total dataset contains 21737 audio recordings with a planned test sample of 4348 audio recordings. On the basis of this corpus, work on improving recognition and optimization algorithms will continue.

III. Data preprocessing

Before starting to create a machine learning model and predicting or classifying, it is necessary to carry out preliminary data processing (Preprocessing) [3]. It is the first and integral step of machine learning, as the quality of the data and the useful information that can be extracted from it, directly affects the learning ability of the model. The main tasks at the data preprocessing stage are:

- Processing of zero values;
- Data normalisation (its transformation to some dimensionless units);
- Control of outliers. These are not errors, but values that abnormally stand out and can distort the model's operation;
- Processing of categorical features;
- The problem of multicollinearity (the presence of a high mutual correlation between two or more independent variables in the model).

The technology stack used to develop the recognition model includes Python programming language. Open Source Python Libraries are *Librosa; Tensorflow; NumPy; Keras; Pandas*. Development environments are *Jupyter and PyCharm*. Optimization of algorithms and metrics for building a recognition model allows reducing training time by orders of magnitude. Therefore, it usually makes sense to start building the model using reduced datasets, successively improving the process parameters (see figure 1).

IV. Models for building voice recognition systems

To date, there are several approaches to the construction of voice recognition systems.

- Recognition models based on spectrograms. The audio signal is converted into a spectrogram - a visual representation of the signal's frequencies over time.
- Models based on the amplitude component of the signal without analysing the frequency characteristics. Very often, recurrent networks RNN, LSTM and GRUs are used as models in this case.
- Models, based on the synthesis of specific useful characteristics of the signal. Among such characteristics, one can distinguish mel-cepstral coefficients (MFCCs), coding with a linear predictor (LPC), gammatone filters (Gammatone filter banks).
- Hybrid models. Usually include several types of models for the synthesis of the recognition model to get the best recognition quality.
- Transfer learning. Training is based on an already pre-trained model on other data where weights are already present. At the same time, the model is further studied on the available data of audio recordings.

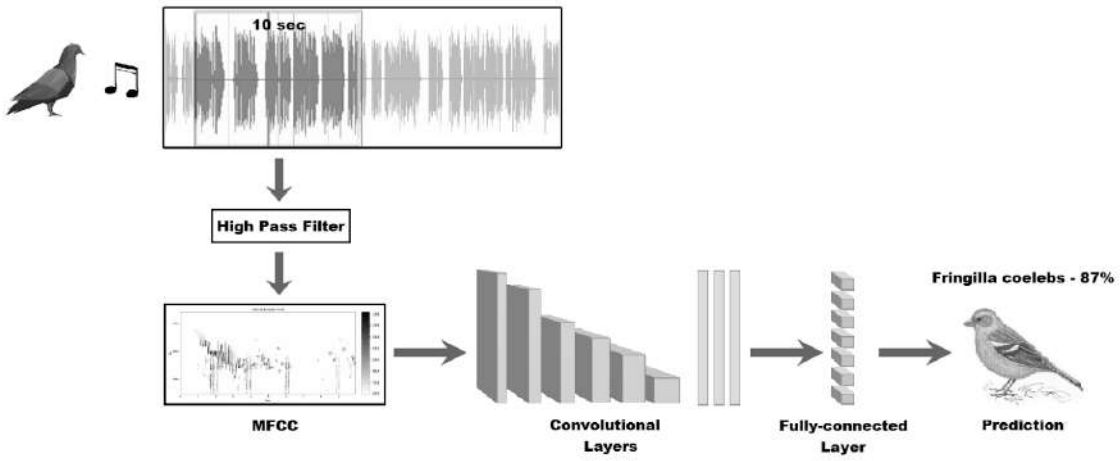

Figure 1. An Audio data preprocessing and neural network model

In our work, we use the first approach, since it certainly results in the best quality of recognized audio signals. For this, we need annotated audio recordings that allow training the model more accurately and recognize bird species.

V. Recognition model based on spectrograms

Spectrogram-based models are a powerful tool for sound prediction tasks due to their ability to capture both temporal and spectral features of audio signals. This approach has proven effective in a wide range of applications [4].

Every sound we hear consists of sound frequencies at the same time interval. The essence of a spectrogram is the visualisation of this set of frequencies on a single graph, as opposed to a sonogram, where only the amplitude of the signal is displayed. A spectrogram is a graphical representation of the spectrum of an audio signal as a function of time. It shows which sound frequencies are present in the audio recording at any given time. A spectrogram is built by applying a Fourier transform to short sections of an audio signal called windows. The resulting spectrum is then displayed as a colour map with time on the horizontal axis and frequency on the vertical axis. The colour of each pixel of the spectrogram corresponds to the amplitude of the corresponding frequency.

Mel-spectrogram is a graphical representation of an audio signal in which frequencies are represented on a Mel scale instead of the linear frequency scale used in a conventional spectrogram. The Mel scale is a reproducible scale based on human perception of sound. It is based on the fact that at low frequencies the audio signal can be distinguished with greater resolution, while at higher frequencies the human ear is less sensitive to changes. Thus, the Mel scale reduces resolution at high frequencies and increases it at low frequencies to better match the human perception of sound. If we combine these two ideas into one, we get a modified spectrogram (MFCC, mel frequency cepstral coefficients), which filters out the frequencies of sounds that a person does not hear, and leaves the most characteristic ones.

Mel scale is calculated as (1):

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}), \quad (1)$$

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi/k], \quad n = 1, 2, \cdots, N \quad (2)$$

where Mel(f) is the logarithmic scale of the normal frequency scale f. Mel scale has a constant mel-frequency interval, and covers the frequency range of 0 Hz - 20050 Hz. The Mel-Frequency Cepstral Coefficients (MFCCs) are computed from the FFT power coefficients which are filtered by a triangular band pass filter bank. The filter bank consists of 12 triangular filters. The MFCCs are calculated as (2)

where Sk(k = 1, 2, ··· K) is the output of the filter banks and N is the total number of samples in a 20 ms audio unit.

Since birds sing at high frequencies, a high-pass filter is used to remove unnecessary noise (leave frequencies at a minimum value of 1400 Hz). These coefficients will be sent to the input of the recognition model. The Python library librosa was used to generate the Mel spectrogram:

signal, sr = librosa.load(fp, sr=self.sr, duration=self.duration, mono=self.mono)

S_ms = librosa.feature.melspectrogram(y=signal, sr=sr, n_fft=self.n_fft, hop_length=self.hop_length, n_mels=self.n_mels, fmin=self.fmin, htk=self.is_htk,)

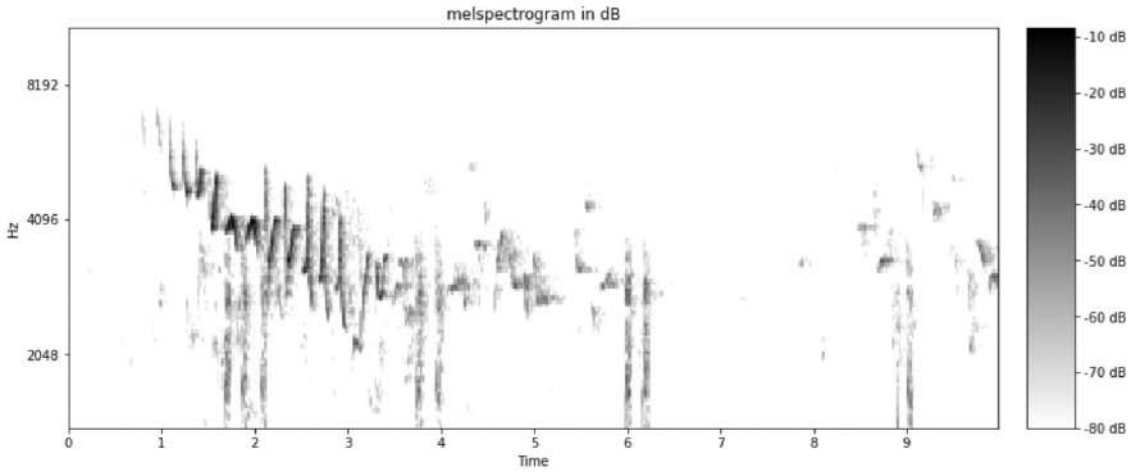

Figure 2. A spectrogram of preprocessing the voice of a Sylvia Atricapilla bird (Lieska-cornahalouka)

S_ms_db = librosa.amplitude_to_db(np.abs(S_ms), ref=np.max)

Where parameters are

sr=22050, n_fft=1024, hop_length=512, n_mels=128.

An example of preprocessing the voice of a Sylvia Atricapilla bird (Lieska-cornahalouka) is a high-pass filter, calculation of mel frequency cepstral coefficients and display as a spectrogram on a graph is shown on figure 2.

VI. Model architecture

Having a Mel spectrogram of the bird's sounds, we will recognize its species among a pre-prepared list of species (classes). The model extracts specific characteristics of the processed data (images), sends them to the input of a deep neural network, and outputs a set of probabilities that correspond to the likelihood that the image belongs to each of these classes. We assume that the class with the highest of these probabilities is the output of the model. This is a deep neural network of the CNN type (Convolutional Neural Network - a convolutional neural network) for recognizing the image class of birds' voices. It is a type of deep learning algorithm that is commonly used in image and video processing applications. Convolutional neural networks are specifically designed to process data that has a grid-like topology, such as images, where the goal is to classify an image into one of several categories. They work by applying a series of filters or convolutions to the input data, which extract features from the input data. These features are then used to make predictions about the input data. They are also used in object detection to locate objects within an image, and in image segmentation for partitioning an image into regions or segments. Overall, CNNs have revolutionised the field of computer vision and have led to significant improvements in image and video processing applications.

In the architecture, we used the EfficientNetB3 network as well as three more layers:

(Flatten, Dropout, Dense with the softmax function as an output) to build a CNN network. The following network structure is *EfficientNetB3 Flatten Dropout Dense(with softmax function as output).*

EfficientNet (https://arxiv.org/abs/1905.11946, https://keras.io/api/applications/efficientnet/efficientnetb3-function) is a modern development of a convolutional neural network from Google Brain (see figure 3). The main objective of EfficientNet was to thoroughly test how to scale the size of convolutional neural networks. For example, one can scale ConvNet based on layer width, layer depth, input image size, or a combination of all these parameters. Thus, the final model was built on the basis of EfficientNetB3 and 14 different classes (bird species) with Adam optimizer, categorical cross-entropy loss function and balanced class weights.

VII. Quality of model training

The quality of bird voice recognition by the represented model is 75.6 percent of the total accuracy on the test data, where

- 7 classes have an F1-score of more than 80 percent
- 3 classes have an F1-score between 70 percent and 80 percent
- 2 classes have an F1-score between 60 percent and 70 percent
- 2 classes have F1-score less than 60 percent.

Moreover, a prototype of the model for automated bird voice recognition *"Bird Sound Recognizer"* was created for the implementation of autonomous continuous monitoring of rare threatened species, indicator species and the state of biodiversity in forest ecosystems [4]. It is located on the online platform corpus.by (http: https://corpus.by/BirdSoundsRecognizer/?lang=en) and is open and free for using [5].

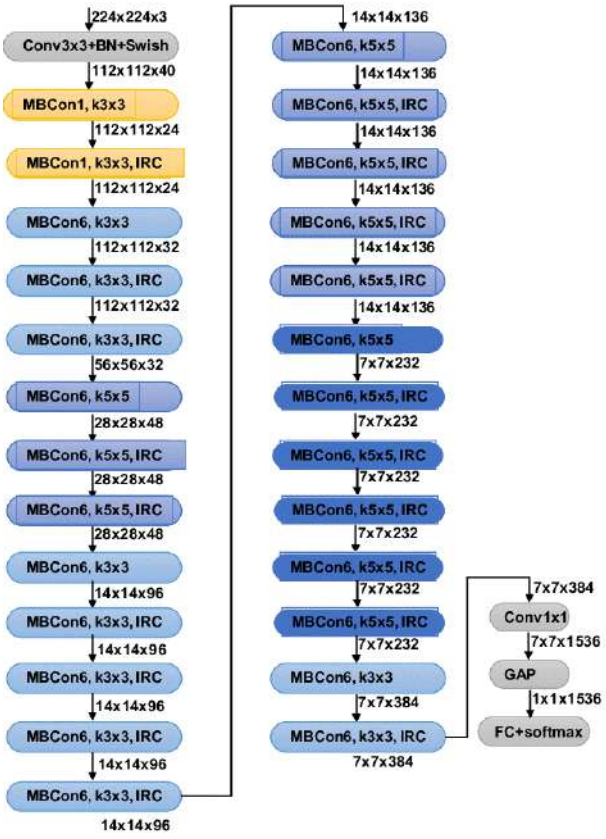

Figure 3. EfficientNet Network architecture

VIII. Model improvements based on ontological approaches

One of the ways to further improve the quality of the model is the ability to take into account, in addition to the sound signal itself, various meta-information: place, time, habitat of a particular species of birds. For a comprehensive solution of this problem, allowing to take into account the various features of the subject area, it is proposed to use the ontological approach.

When it comes to recognizing birdsong from an audio signal, ontological approaches are essential for several reasons. First and foremost, ontologies provide a way to organize and structure the vast amounts of data that are available for each uploaded audio file. In this case, the meta-information available for each audio file can be used to create a taxonomy of bird species, subspecies, and groups. This makes it easier to organize and analyze the data, and can help to improve the accuracy of the recognition process.

In addition, ontological approaches can help to address issues related to data inconsistency and incompleteness. By using a well-defined and structured ontology, it becomes easier to identify and correct inconsistencies in the data, such as misspellings or variations in naming conventions.

Furthermore, ontologies can help to facilitate the integration of different data sources, such as data from different recording locations or from different researchers. By using a common ontology, it becomes easier to merge and compare data from different sources, which can help to improve the accuracy of the recognition process.

Finally, ontological approaches can help to support the development of intelligent systems that are capable of learning and adapting to new data. By using a well-defined ontology as a basis for machine learning algorithms, it becomes possible to create systems that can recognize new bird species and adapt to new recording conditions.

The choice of technology for implementing the ontological approach is also important. Because it should provide a sufficient level of flexibility and scalability to integrate various types of systems and knowledge into them. As a technological basis, it is proposed to use the OSTIS technology, the main advantages and specifics of which in signal processing and integration with DNN are presented in papers [6-10]. An example of a fragment of a top-level ontology using OSTIS sc-code is presented below.

Sylvia Atricapilla
:= [Eurasian blackcap]
∈ *English language*
:= [Lieska-čornahaloŭka]
∈ *Belarusian language*
∈ *specie*
⊂ *sylvia*
∈ *genus*
⊂ *old world warbler*
∈ *family*
⊂ *perching birds*
∈ *order*
⊂ *vertebrates*
∈ *phylum*
⊂ *animal*
∈ *kingdom*
⇒ *habitat**:
{• *Eastern Europe*
⇒ *coordinates**:
{• [From: 54.1343° N, 28.5079° E]
• [To: 54.5028° N, 28.8794° E]
}
• *Western Asia*
• *Northwestern Africa*
}

By using a structured and well-defined ontology to organize and analyze the available data, it becomes possible to create intelligent systems that are capable of accurately recognizing a wide range of bird species, subspecies, and groups.

IX. Conclusion

The article describes a model for recognizing the voices of Belarusian birds. It is based on the analysis of a mel-frequency cepstrum (MFCC, mel-frequency cepstrum). The Mel spectrogram is a graphical representation of an audio signal in which frequencies are represented on Mel scale instead of the linear frequency scale used in a conventional spectrogram. For machine learning of the model, a deep neural network of the CNN (Convolutional Neural Network) was used to recognize the image class of birds' voices, since this type of network is more suitable for image recognition tasks. To build the CNN network, we chose the EfficientNetB3 network, as well as three more layers (Flatten, Dropout, Dense with the softmax function as output). Thus, the final model was built on the basis of EfficientNetB3 and 14 different classes (bird species) with Adam optimizer, categorical cross-entropy loss function and balanced class weights.

The overall recognition quality of the model was 75.6 percent. The conduction of the experiment was fulfilled on 14 species of birds with 200 records for each of the species using a CNN-based model with spectrograms as a characteristic of the input signal to the model. In the future, it is planned to expand the list of bird species for recognition to 116. The next step of the project is to monitor the continuous signal for the detection of the bird's voice in real time.

To solve this task, a recognition system will be designed using another data set, the audio files of which are annotated in detail by ornithologists, taking into account the time stamps for the bird's voice.

In paper also proposed approaches to further improve the quality of ML-models through the use of the ontological approach and OSTIS technology.

References

[1] F. Briggs, R. Raich, X. Z. Fern Audio Classification of Bird Species: A Statistical Manifold Approach. ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009, pp. 51–60.

[2] D. Stowell, M. D. Plumbley An open dataset for research on audio field recording archives: freefield1010. *ArXiv*, 2013, vol. abs/1309.5275.

[3] D. Stowell, M. D. Plumbley Automatic large–scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2014, vol. 2.

[4] S. A.Hajdurau, D. I. Latysevic, A. A. Bakunovic, L. I. Kajharodava, V. A. Chachlou, Ja. S. Zianouka, Ju. S. Hiecevic Madel baz danych dlia technalohii autamatyzavanaha raspaznavannia halasavych sihnalau zyviol [Database model for the technology of automated recognition of animal voice signals]. XXI International Scientific and Technical Conference "Development of Informatization and State System of Scientific and Technical Information RINTI-2022", UIIP NASB, Minsk, 2022, pp. 236–240.

[5] J. S. Hiecevič, J. S. Zianoŭka, A. S. Trafimaŭ, A. A. Bakunovič, D. I. Latyševič, A. JA. Drahun, M. M. Sliesarava, M. S. Tukaj Komplieks srodkaŭ realizacyi zadač štučnaha inteliektu dlia bielaruskaj movy [A complex of means to implement tasks of artificial intelligence for the Belarusian language]. *Piervaja vystavka-forum IT-akadiemhrada Iskusstviennyj intielliekt v Bielarusi*. UIIP NASB. Minsk. 2022. P. 64-73. (In Belarusian)

[6] V. V. Golenkov, N. A. Gulyakina, D. V. Shunkevich Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems, Bestprint, Minsk, 2021, P. 690

[7] V. Golovko and et el. Integration of artificial neural networks and knowledge bases *Open Semantic Technologies for Intelligent Systems (OSTIS-2018)*, BSUIR, Minsk, 2018, pp. 133–146.

[8] M. Kovalev, A. Kroshchanka, V. Golovko, Convergence and integration of artificial neural networks with knowledge bases in next-generation intelligent computer systems *Open Semantic Technologies for Intelligent Systems (OSTIS-2022)*, BSUIR, Minsk, 2022, pp. 173–186.

[9] V. Zahariev, E. Azarov, K. Rusetski.An approach to speech ambiguities eliminating using semantically-acoustical analysis *Open Semantic Technologies for Intelligent Systems (OSTIS-2018)*, BSUIR, Minsk, 2018, pp. 211–222.

[10] V. Zahariev, K. Zhaksylyk, D. Likhachov, N. Petrovsky, M. Vashkevich, E. Azarov. Audio interface of next-generation intelligent computer systems *Open Semantic Technologies for Intelligent Systems (OSTIS-2022)*, BSUIR, Minsk, 2022, pp. 239–250.

Разработка системы распознавания звуков птиц с использованием онтологического подхода

Зеновко Е., Белявский Д., Кайгородова Л.,
Трофимов А., Хохлов В., Гецевич Ю.,
Захарьев В., Жаксылык К.

В работе предложена модель распознавания голосов птиц Республики Беларусь, основанная на анализе мел-спектрограмм (MFCC, mel-frequency cepstrum). Мел-спектрограмма — это графическое представление звукового сигнала, в котором частоты представлены в мел-шкале вместо линейной шкалы частот, используемой в обычной спектрограмме. Шкала Mel — шкала высоты звуков, отсеивающая частоты звуков, которые человек не слышит, и оставляет самые характерные, находящихся на одинаковой дистанции для слушателя. Для машинного обучения модели была использована глубокая нейронная сеть типа CNN (Convolutional Neural Network) для распознавания класса изображения голоса птиц, так как именно этот вид сети больше подходит для задач распознавания изображений. Для построения сети CNN мы применили сеть EfficientNetB3, а также еще три слоя (Flatten, Dropout, Dense с функцией softmax в качестве выхода). Таким образом, окончательная модель была построена на основе EfficientNetB3 и 14 различных классов (видов птиц) с оптимизатором Адама (Adam optimizer), категориальной функцией потерь перекрестной энтропии (categorical cross-entropy loss function) и сбалансированными весами классов.

При проведении данного эксперимента на 14 видах птиц с 200 записями для каждого из видов и использованием модели на базе CNN со спектрогарамами в качестве характеристики сигнала для входа на модель, получено общее качество распознавания 75,6 процентов. В дальнейшем планируется расширение списка видов птиц для распознавания до 116.

Показаны возможности использования онтологических подходов и технологии OSTIS для дальнейшего повышения качества моделей машинного обучения.

Received 13.03.2023