Honey Bee Queen Presence Detection from Audio Field Recordings using Summarized Spectrogram and Convolutional Neural Networks

Agnieszka Orlowska, Dominique Fourer, Jean-Paul Gavini, Dominique Cassou-Ribehart

Honey Bee Queen Presence Detection from Audio Field Recordings using Summarized Spectrogram and Convolutional Neural Networks

Agnieszka Orlowska, Dominique Fourer, Jean-Paul Gavini and Dominique Cassou-Ribehart

abstract

The present work proposes a simple supervised method based on a downsampled time-frequency representation of the input audio signal for detecting the presence of the queen in a beehive from noisy field recordings. Our proposed technique computes a "summarized-spectrogram" of the signal that is used as the input of a deep convolutional neural network. This approach has the advantage of reducing the dimension of the input layer and the computational cost while obtaining better classification results with the same deep neural architecture. Our comparative evaluation based on a cross-validation beehive-independent methodology shows a maximal accuracy of 96% using the proposed approach applied on the evaluation dataset. This corresponds to a significant improvement of the prediction accuracy in comparison to several state-of-the-art approaches reported by the literature. Baseline methods such as MFCC, constant-Q transform and classical STFT combined with a CNN fail to generalize the prediction of the queen presence in an unknown beehive and obtain a maximal accuracy of 55% in our experiments.

Key words: Honey Bee Queen Detection, Audio Classification, Time-Frequency Analysis, Convolutional Neural Networks

A. Orlowska and D. Fourer
IBISC (EA 4526), Univ. Évry/Paris-Saclay, Evry-Courcouronnes, France
e-mail: dominique.fourer@univ-evry.fr

J-P. Gavini and D. Cassou-Ribehart
Starling Partners Company, Paris, France
e-mail: jpg|dcr@starling.partners

1 Introduction

Smart beekeeping is an emerging and promising research field which aims at providing computational solutions for aiding the monitoring of bee colonies. It is known [13, 12] that bees produce specific sounds when exposed to stressors such as failing queens, predators, airborne toxicants. However experienced beekeepers are not always able to explain the exact causes of the sound changes without a hive inspection. Nonetheless, hive inspections disrupt the life cycle of bee colonies and can involve additional stress factors for the bees [3, 2]. With this in mind, several recent studies propose to analyze the audio signature of a beehive through a machine learning approach [10, 3] in order to develop systems for automatically discriminating the different health states of a beehive. For example, in [8, 7], the authors propose a method which combines the Short-Time Fourier Transform (STFT) of the analyzed audio recording with convolutional neural networks (CNN) to discriminate bee sounds from the chirping of crickets and ambient noise. This approach outperforms classical machine learning methods such as k-nearest neighbors, support vector machines or random forests for classifying audio samples recorded by microphones deployed above landing pads of Langstroth beehives[1]. The detection of the queen presence appears to be one of the most important tasks for smart beekeeping and is addressed in [4] with a complete beehive machine-learning-based audio monitoring system. More recently in [2, 11], the authors investigate Music Information Retrieval (MIR)-inspired approaches based on mel-frequency cepstral coefficients (MFCC), and on the spectral parameters of sinusoidal signal components as input features of a supervised classification method based on a CNN to predict for the presence of the queen in a beehive from recorded audio signals.

In spite of promising results reported in the literature, a further evaluation of the state-of-the-art approaches reveals overfitting problems using the trained models for detecting the queen presence, when applied to distinct beehives as presented for example in [10] through a beehive-independent classification experiment. This lack of generality for the trained model can be critical in real-world application scenarios because the trained models cannot efficiently be applied to another arbitrary chosen beehive without a new beehive-specific training of the model using annotated examples. Thus, the present work introduces a very simple but efficient transformation technique which improves the results of a CNN-based audio classification method in beehive-independent configurations. We compute a “summarized” time-frequency representation through a specific downsampling technique which experimentally reveals a better generalization of the trained model based on a convolutional neural network architecture. Our technique can arbitrary reduce the dimension of the input features provided to the CNN to obtain the best trade-off between the model accuracy and the computational cost.

This paper is organized as follows. In Section 2, we present the framework of the problem addressed in this study with a description of the experimental materials. In Section 3 we present the proposed approach and we introduce our supervised technique based on the summarized spectrogram for automatically predicting the queen presence in a beehive from audio recordings. In Section 4, we comparatively

assess our new proposed method with several state-of-the-art approaches, in terms of prediction accuracy with a consideration for the dimension of the computed audio features. Finally, this paper is concluded by a discussion including future work directions in Section 5.

2 Framework

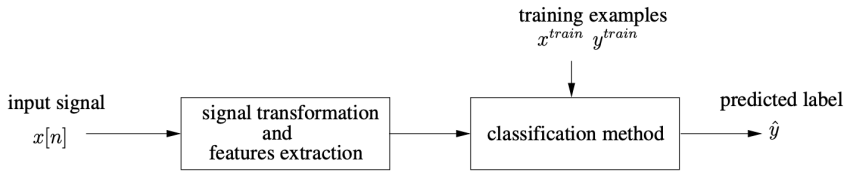

Fig. 1 Illustration of the overall proposed approach.

2.1 Problem formulation and notations

We address the problem of prediction the state of a beehive using an audio signal x resulting from a field recording of a monitored beehive. The overall approach is based on a supervised machine learning approach depicted in Fig. 1 in which relevant audio features are first computed from x before being processed by a classification method. At the training step, training examples x^{train} and labels y^{train} are used to fit the model parameters of the classification method. At the testing step, the trained model is used to predict from x the associated state of the beehive associated to a label $\hat{y}$ (y being the unknown ground truth). This work aim at proposing the best processing pipeline allowing an accurate prediction of the beehive health state, and focuses on the signal transformation and feature extraction step.

2.2 Materials

We use the publicly available dataset introduced by Nolasco and Benetos in [10] during the Open Source Beehive (OSBH) project and the NU-Hive project[1]. The dataset contains annotated audio samples acquired from six distinct beehives. The present work focuses on the audio signals which were annotated as “bee”, corresponding to

[1] `https://zenodo.org/record/1321278`.

sounds emitted by the beehive. Hence, the "no bee" annotated signals correspond to external noises and are simply not investigated in our study. At a pre-processing step, each audio recording is resampled at rate of $F_s = 22.05$ kHz as in [10] and is transformed to single-channel signals by averaging samples from the available channels. Each recording is then split in one-second-long homogeneous time series (associated to the same annotation label). As a result, we obtain a dataset of 17,295 distinct individuals where 8,444 ones are labeled as "*queen*" ($y = 1$) and 8,851 ones are labeled as "*no queen*" ($y = 0$). An overview of the investigated dataset with the considered labels for each beehive is presented in Table 1.

Table 1 Description of the dataset content investigated in the present study. Each individual corresponds to a one-second-long audio signal sampled at $F_s = 22.05$ kHz.

Beehive name	*queen*	*no queen*	Total
CF001	0	16	16
CF003	3,700	0	3,700
CJ001	0	802	802
GH001	1,401	0	1,401
Hive1	2,687	1,476	4,163
Hive3	656	6,557	7,213
Total	8,444	8,851	17,295

3 Proposed Method

3.1 Time-frequency representation computation

The Short-Time Fourier Transform (STFT) is a popular technique designed for computing time-frequency representations of real-world signals. STFT appears in a large number and variety of signal processing methods which involve non-stationary multicomponent signals that can efficiently be disentangled using a Fourier transform combined with a sliding analysis window [6]. Given a discrete-time finite-length signal $x[n]$, with time index $n \in \{0, 1, ..., N-1\}$, and an analysis window h, the discrete STFT of x can be computed as:

$$F_x^h[n, m] = \sum_{k=-\infty}^{+\infty} x[k]h[n-k]^* e^{-j\frac{2\pi mk}{M}} \tag{1}$$

with z^* the complex conjugate of z and $j^2 = -1$. Here, $m \in \{0, 1, ..., M-1\}$ corresponds to an arbitrary frequency bin associated to the frequency $f = \frac{m}{M}F_s$ expressed in Hz, for $m \leq \frac{M}{2}$.
The spectrogram is defined as the squared-modulus of the STFT $|F_x^h[n, m]|^2$ [5]. In practice, it is fractioned along the time axis by considering an integer hop size

$\Delta n > 1$ with a possible overlap between adjacent frames. As a result, we obtain a $M \times L$ matrix with M the arbitrary number of computed frequency bins, and $L = \lfloor \frac{N}{\Delta n} \rfloor$ ($\lfloor . \rfloor$ being the floor function) the resulting number of time indices such as $L \leq N$ when $\Delta n > 1$.

3.2 Summarizing process

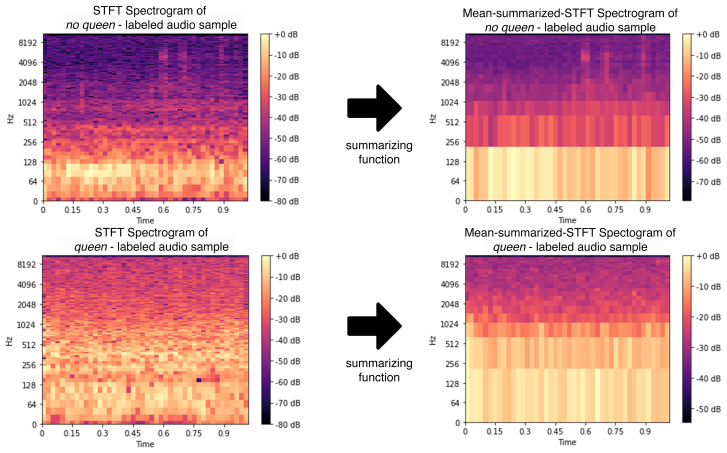

Fig. 2 Classical- and summarized-spectrogram time-frequency representation comparison of a one-second-long beehive audio recording.

The main problem occurring when a STFT is used as the input of a neural network is the high number of input coefficients which can lead to a high memory consumption and a heavy computation cost during the training step. Hence, we propose a simple dimension reduction method of the spectrogram which aims at preserving the relevant information present in the time-frequency plane without modifying the original time-frequency resolution related to the analysis window. To this end, we use a summary process on the computed spectrogram $|F_x^h[n,m]|^2$ which consists in two steps. First, the positive frequency axis ($m \in [0, \lfloor \frac{M}{2} \rfloor]$) is partitioned into a finite number of equally spaced frequency bands such as $B < \frac{M}{2}$. Second, at each time index, the information of each frequency band is summarized into a unique coefficient by applying a summary aggregating function denoted $g()$ along the frequency axis (the best choice for g is discussed later). The summarized-spectrogram SF_x^h with a reduced dimension of $B \times L$ is computed as:

$$\text{SF}_x^h[n,b] = g\left(|F_x^h[n,m_b]|^2\right)_{\forall m_b \in \left[b\lfloor \frac{M}{2B} \rfloor, (b+1)\lfloor \frac{M}{2B} \rfloor - 1\right]} \tag{2}$$

with $b \in [0, B-1]$ the new frequency bin. We illustrate in Fig. 2 the result obtained by computing the summarized-spectrogram of two audio signals corresponding to beehive recordings respectively labeled as "*queen*" and "*no queen*" using the arithmetic mean as g function.

3.3 2D Convolutional Neural Network

CNN is a natural choice for analyzing a time-frequency representation that can also be considered as an image. To predict the label corresponding to the state of a beehive from an audio signal, the resulting summarized-spectrogram is processed by the deep neural network architecture inspired from [2] using 2 additional convolutional layers. It consists of 6 convolutional blocks including with a 3×3 kernel size, followed by a batch normalization, a 2×2 max-pooling and a 25% dropout layers. The output is connected to a 3 fully connected layers (FC) including 2 droupout layers of respectively 25% and 50% followed by a softmax activation function to compute the predicted label $\hat{y}$ (rounding to the closest integer 0 or 1). Convolutional and FC layers both use a LeakyReLU activation function defined as $LeakyRELU(x) = \max(\alpha x, x)$, with $\alpha = 0.1$.

4 Numerical Experiments

4.1 Experimental protocol

We propose here two distinct experiments for comparatively assessing our new proposed method described in Section 3 with several other state-of-the-art approaches for predicting the queen presence.
Experiment 1: We merge the 6 available beehives and then we apply a random split to obtain 70% of the individuals for training and 30% for testing.
Experiment 2: We use a 4-fold cross-validation methodology where the beehives are independent. To this end, the folds have been manually created to assign each beehive to a unique fold. An exception is made for the testing folds 1 and 2 which contain two beehives since CF001, CF003, CJ001 and GH001 only contain individuals from the same annotation label. The proposed partitioning of the whole dataset in Experiment 2 is detailed in Table 2.

4.2 Implementation details

The investigated methods have been implemented in python using when needed the following libraries: Librosa is used for audio processing and features extrac-

Table 2 Description of the partitioned dataset investigated in Experiment 2.

Fold	Training set	Testing Set
Fold 1	CJ001 + GH001 + Hive3 + Hive 1	CF001 + CF003
Fold 2	CF001 + CF003 + Hive3 + Hive 1	CJ001 + GH001
Fold 3	CJ001 + GH001 + Hive3 + CF001 + CF003	Hive1
Fold 4	CJ001 + GH001 + Hive1 + CF001 + CF003	Hive3

	Fold 1	Fold 2	Fold 3	Fold 4
queen	3700	1401	2687	656
no queen	16	802	1476	6557
Total	3716	2203	4163	7213

tion, Keras with Tensorflow are used for the implementation and the use of the proposed CNN architecture, and scikit-learn [9] is used for computing the evaluation metrics. The training of our CNN was configured for a constant number of 50 epochs with a batch size of 145. The numerical computation was performed using an Intel(R) Xeon(R) W-2133 CPU @ 3.60GHz CPU with 32GB of RAM and a NVIDIA GTX 1080 TI GPU. The Python code used in this paper is freely available at `https://github.com/agniorlowska/beequeen_prediction` for the sake of reproducible research.

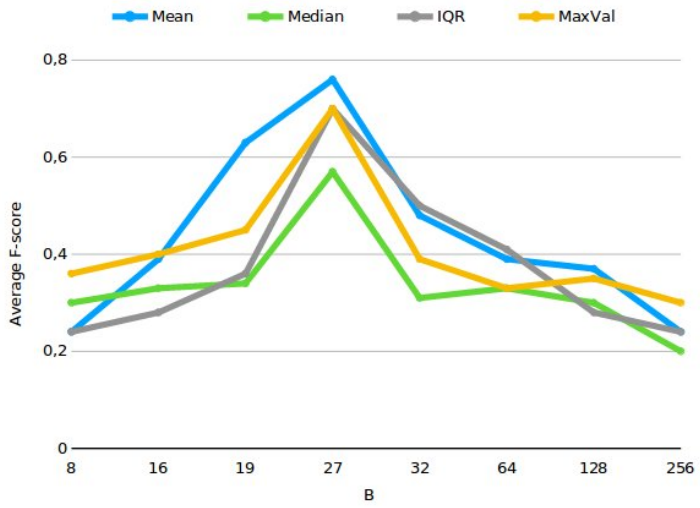

Fig. 3 Average F-measure for different summary function g and B value configurations.

4.3 Hyperparameters tuning and data augmentation

To define the best value of B with the best summary function $g()$, we evaluated several configurations by considering the beehive-independent Experiment 2 protocol. According to the results presented in Fig. 3, we chose $B = 27$ and the mean function which obtained the best results. We also tried to apply the summarizing process separately on the real and the imaginary part of the STFT before computing the spectrogram however this provides very poor results in each configuration. To improve the performance of the trained model, we used a data augmentation (DA) technique which artificially increases of 50% the number of training individuals by the addition of a white Gaussian noise to existing ones. The variance of the noise signal has been define to obtain a resulting signal-to-noise-ratio (SNR) equal to 30 dB. Due to the increase of computation time, we only applied DA on the best resulting method presented in Tables 3 and 4. Our simulations show that DA does not significantly improve the results obtained with MFCC and CQT-based methods which are poorer than with the STFT.

4.4 Comparative results

Our proposed method is compared to several existing approaches introduced in [10, 11]. The Mel Frequency Cepstral Coefficients (MFCCs) + CNN method is a popular approach proposed in [11] where the number of computed MFCC is set to 20. The constant-Q transform (CQT) + CNN, was also investigated as a baseline method since CQT which can be viewed as a modified version of the discrete STFT with an varying frequency resolution.The so-called Q-factor corresponds to $Q = \frac{f}{\Delta f}$ where Δf is the varying frequency resolution (difference between two frequency bins). All the investigated signal representations use exactly the same CNN architecture for which the dimension of the input layer is adapted. The classification results are obtained with our proposed method based on the mean summary function for a number of frequency bands $B = 27$ computed from a STFT or CQT with $M = 1025$ and with an overlap of 50% ($\Delta n = 512$) between adjacent frames to obtain an input features matrix of dimension 27×44. The results in the two experiments respectively expressed in terms of Precision, Recall, F-score and Accuracy metrics are reported in Tables 3 and 4. According to Table 3, all the compared methods are almost equivalent since they obtain excellent classification results with an almost perfect accuracy of 1. These results are comparable with those reported in the literature and can be explained by the fact that all the available beehives are merged in the same training set. The beehive-independent results are presented in Table 4 and are very different. Now, the best results in Experiment 2 are only obtained with our proposed method (denoted mean-STFT) which uses the summarized-spectrogram combined with a CNN and obtains an average F-score of 0,75. The use of the data augmentation improves the results and leads to a maximum accuracy of 0,96.

Table 3 Comparison of the classification results in Experiment 1 (random split).

Method	Features	Label	Precision	Recall	F-score	Accuracy
MFCCS+CNN [11]	20×44	Queen	1.00	0.99	0.99	0.99
		No queen	0.99	1.00	0.99	
STFT+CNN	513×44	Queen	1.00	0.93	0.97	0.97
		No queen	0.94	1.00	0.97	
CQT+CNN [11]	513×44	Queen	0.96	0.93	0.95	0.95
		No queen	0.92	1.00	0.95	
mean-CQT+CNN	27×44	Queen	0.98	1.00	0.99	0.99
		No queen	0.99	0.98	0.98	
mean-STFT+CNN	27×44	Queen	**0.99**	**1.00**	**1.00**	**1.00**
		No queen	**1.00**	**0.99**	**1.00**	
mean-STFT+CNN+DA	27×44	Queen	**0.99**	**1.00**	**1.00**	**1.00**
		No queen	**1.00**	**0.99**	**1.00**	

Table 4 Comparison of the classification results in Experiment 2 (4-fold hive-independent cross-validation).

Method	Features	Label	Precision	Recall	F - score	Accuracy
MFCCs+CNN [11]	20x44	Queen	0.36	0.44	0.40	0.31
		No queen	0.22	0.16	0.19	
STFT+CNN	513×44	Queen	0.77	0.76	0.66	0.55
		No queen	0.33	0.20	0.33	
CQT+CNN	513×44	Queen	0.10	0.07	0.08	0.25
		No queen	0.32	0.41	0.36	
mean-CQT+CNN	27x44	Queen	0.25	0.11	0.16	0.38
		No queen	0.41	0.65	0.50	
mean-STFT+CNN	27x44	Queen	0.71	0.86	0.78	0.75
		No queen	0.81	0.64	0.71	
mean-STFT+CNN+DA	27×44	Queen	**0.96**	**0.99**	**0.96**	**0.96**
		No queen	**0.99**	**0.94**	**0.96**	

5 Conclusion

We have introduced and evaluated a new downsampling method for improving the prediction of the presence of a queen bee from audio recordings using a deep CNN. Despite its simplicity, the summarized-spectrogram has a better efficiency in comparison to other perception-motivated representations such as MFCC or CQT, when they are used as input features for the queen presence detection problem. Hence, we have obtained a maximal resulting accuracy of 96% in a beehive-independent split configuration which is very promising. This result paves the way of future real-world applications of smart beehive monitoring techniques based on embedded systems. Future work consists in a further investigation including more data provided by monitored beehives. Moreover, we expect a further investigation of the relevant

information conveyed by the summarized-spectrogram when used for providing audio features, in order to design new audio classification methods.

References

1. Abu, T., Sahile, G., et al.: On-farm evaluation of bee space of langstroth beehive. Livestock Research for Rural Development **23**(10) (2011)
2. Cecchi, S., Terenzi, A., Orcioni, S., Piazza, F.: Analysis of the sound emitted by honey bees in a beehive. In: Audio Engineering Society Convention 147 (2019)
3. Cecchi, S., Terenzi, A., Orcioni, S., Riolo, P., Ruschioni, S., Isidoro, N.: A preliminary study of sounds emitted by honey bees in a beehive. In: Audio Engineering Society Convention 144. Milan, Italy (2018)
4. Cejrowski, T., SzymaÅĎski, J., Mora, H., Gil, D.: Detection of the Bee Queen Presence Using Sound Analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10752 LNAI, pp. 297–306 (2018). DOI 10.1007/978-3-319-75420-8_28
5. Flandrin, P.: Time-frequency/time-scale analysis. Academic press (1998)
6. Fourer, D., Harmouche, J., Schmitt, J., Oberlin, T., Meignen, S., Auger, F., Flandrin, P.: The ASTRES toolbox for mode extraction of non-stationary multicomponent signals. In: Proc. EUSIPCO 2017, pp. 1170–1174 (2017)
7. Kulyukin, V., Mukherjee, S., Amlathe, P.: Toward Audio Beehive Monitoring: Deep Learning vs. Standard Machine Learning in Classifying Beehive Audio Samples. Applied Sciences **8**(9), 1573 (2018). DOI 10.3390/app8091573
8. Kulyukin, V.A., Mukherjee, S., Burkatovskaya, Y.B., et al.: Classification of audio samples by convolutional networks in audiobeehive monitoring. Tomsk State University Journal of Control and Computer Science (45), 68–75 (2018)
9. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, vol. 8, pp. 18–25 (2015)
10. Nolasco, I., Benetos, E.: To bee or not to bee: Investigating machine learning approaches for beehive sound recognition. In: Proc. DCASE (2018)
11. Nolasco, I., Terenzi, A., Cecchi, S., Orcioni, S., Bear, H.L., Benetos, E.: Audio-based identification of beehive states. In: Proc. IEEE ICASSP, pp. 8256–8260 (2019)
12. Papachristoforou, A., Sueur, J., Rortais, A., Angelopoulos, S., Thrasyvoulou, A., Arnold, G.: High frequency sounds produced by Cyprian honeybees *Apis mellifera cypria* when confronting their predator, the Oriental hornet *Vespa orientalis*. Apidologie **39**(4), 468–474 (2008). DOI 10.1051/apido:2008027
13. Wenner, A.M.: Sound production during the waggle dance of the honey bee. Animal Behaviour **10**(1-2), 79–95 (1962). DOI 10.1016/0003-3472(62)90135-5