AUDIO-BASED IDENTIFICATION OF BEEHIVE STATES

Inês Nolasco[2], Alessandro Terenzi[1], Stefania Cecchi[1], Simone Orcioni[1], Helen L. Bear[2], and Emmanouil Benetos[2,3]

[1]Department of Information Engineering, Universitá Politecnica delle Marche, Italy
[2] School of EECS, Queen Mary University of London, UK [3] The Alan Turing Institute, UK

ABSTRACT

The absence of the queen in a beehive is a very strong indicator of the need for beekeeper intervention. Manually searching for the queen is an arduous recurrent task for beekeepers that disrupts the normal life cycle of the beehive and can be a source of stress for bees. Sound is an indicator for signalling different states of the beehive, including the absence of the queen bee. In this work, we apply machine learning methods to automatically recognise different states in a beehive using audio as input. We investigate both support vector machines and convolutional neural networks for beehive state recognition, using audio data of beehives collected from the NU-Hive project. Results indicate the potential of machine learning methods as well as the challenges of generalizing the system to new hives.

Index Terms— Empirical mode decomposition, Hilbert-Huang transform, bioacoustics, computational bioacoustic scene analysis, beehive state recognition.

1. INTRODUCTION

Among insects, honey bees (*Apis mellifera L.*) are well known for their positive effects. Their importance is not limited to the production of honey, beeswax, royal jelly, and propolis but they are also at the basis of plant pollination, playing a key role in the proliferation of both spontaneous and cultivated flora. In recent years multiple stress factors have led to a decline of honey bee colonies [1] and this event has emphasized the significance of a continuous and extensive monitoring to investigate factors that may negatively affect the life cycle of bees.

In this context, the analysis of sound generated within the bee hives is an important approach for non-invasive monitoring [2]. Vibration and sound signals are used by honey bees to communicate within the colony [3, 4]. Honey bees produce their sounds by means of gross body movements, wing movements, high-frequency muscle contractions without wing movements, and pressing the thorax against the substrates or another bee [5, 6, 7].

In recent years, several studies have underlined that some behaviors of the honey bees are strictly related to variation in produced sound [5, 7, 8, 9]. In particular, these works have proved that there is a strict correlation between the amplitudes and frequencies of the bee hive sounds and some events like swarming [10, 11, 12, 13] and queen presence [8, 7, 9]. In [14], a relation between sound and changes in environmental conditions have been reported. Furthermore, it has been demonstrated that sound analysis could be a powerful instrument in pest monitoring e.g., in [13] it has been used for varroa-mite detection.

Recent works in beehive sound analysis are carried out through a computational bioacoustic scene analysis perspective [15]. In this context, relevant representations for these audio signals are combined with machine learning methods in order to develop systems that can automatically distinguish between different states of a hive. In [16] and [17], the authors explore the use of Mel spectra and Mel-frequency cepstral coefficients (MFCCs) together with different machine leaning methods to detect hives with and without the queen bee. In the context of computational sound scene analysis research, state-of-the-art methods for sound scene recognition as [18] show, are mainly based in Convolutional Neural Networks (CNNs). In [19], the authors explore the use of CNNs to the problem of beehive sound identification and highlight the long-term aspects of such sounds. They stress the need for long-term contextual representations for modeling such data. Also, in [20], the authors present a method to extract long-term features from spectrograms for the task of sound scene classification.

This work expands the preliminary work of [17] to investigate the potential of traditional and neural network-based machine learning methods exploiting MFCCs, Mel spectrograms, and the Hilbert Huang Transform (HHT) [21] as features to determine the presence of the queen bee in a hive.

The novelty of the proposed approach is related not only to the application of deep learning methods to this problem, but also to the use of HHT as spectral representations and the design of representations suitable for modelling long-term temporal context. The novelty is underlined in the experimental results where the proposed system has been tested on real audio data collected from the NU-Hive [22] project. By designing a *hive-independent* evaluation setup, inspired by speaker-independent evaluations in speech processing [23], we demonstrate the validity and potential of the developed system in a real-world scenario.

The paper is organized as follows. Section 2 describes the proposed approach, including the feature extraction and classification methodologies used. Section 3 presents the data acquired in honey bee hives, the experimental setup, and the obtained results. Finally, conclusions and future directions are reported in Section 4.

2. PROPOSED APPROACH

The proposed approach is based on two steps. Firstly, feature extraction is carried out using MFCCs, Mel spectrograms, and the HHT algorithm, with the aim of determining the frequency behaviour of the beehive when the queen is present or not. Secondly, classification of beehive states is achieved using both support vector machines (SVMs) and convolutional neural networks (CNNs), with different

This work was supported by a Universitá Politecnica delle Marche Research Grant, EPSRC Grant EP/R01891X/1, RAEng Research Fellowship RF/128, and by The Alan Turing Institute under EPSRC grant EP/N510129/1.

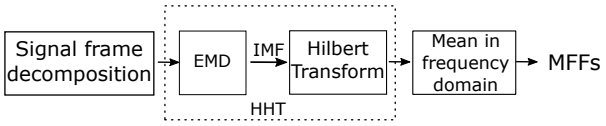

Fig. 1. Feature extraction procedure based on HHT.

combinations of features and parameters as appropriate.

2.1. Feature extraction

Three types of features are extracted to use within the classifiers. First, both Mel spectra and MFCCs, which are commonly used representations in the context of computational sound scene analysis [24], are used in this work. We compute MFCCs with 20 coefficients and Mel spectra with 120 frequency bands.

The extraction of a new feature based on the HHT [21] is considered in the proposed procedure. This choice derives from the fact that it is known from the literature that honey bee sounds are non-stationary signals composed by a superimposition of tones at various frequencies [13]. Figure 1 shows each step of the procedure. The signal is divided in frames of length $N = 32000$ samples using a frequency sampling $f_s = 32$ kHz, with analysis performed every 1 second. Each signal frame is decomposed with Empirical Mode Decomposition (EMD) [25] to obtain a set of basis functions which then are analyzed with the Hilbert transform. EMD uses adaptive basis functions for signal decomposition and this is useful when measured data are non-stationary and non-linear. The decomposition is based on the idea that the signals are composed of simple intrinsic modes of oscillations. Each intrinsic mode represents a simple oscillation with the same number of extrema and zero crossings. Each oscillation is represented by an Intrinsic Mode Function (IMF), which are the EMD basis functions and are defined by the following properties: (A) The number of extrema and the number of zero-crossings must be either equal or differ at most by one. (B) The mean value of the envelope defined by the local maxima and the local minima is zero.

Starting from this definition, any signal frame $x(n)$ with $n = 1, \ldots, N$, can be decomposed following these steps:

1. Identify all local extrema.
2. Connect all the local maxima by a cubic spline.
3. Repeat the procedure to produce the lower envelope.
4. Estimate their mean $m_1(n)$.
5. The first estimation of the IMF can now be written as $h_1(n) = x(n) - m_1(n)$.
6. Repeat the procedure up to k times until the function $h_{1k}(n) = h_{1(k-1)}(n) - m_{1k}(n)$ does not satisfy the IMF properties.
7. Now the first IMF component is equal to $c_1(n) = h_{1k}(n)$.
8. Remove from the original signal the component $c_1(n)$ obtaining the first residue $r_1(n) = x(n) - c_1(n)$.
9. Treat $r_1(n)$ as the new signal for the decomposition procedure.

The process is stopped when the residue $r_n(n)$ becomes a monotonic function, or when the amplitude is less then a predetermined value. The number of components $c_j(n)$ is M and it is not a predetermined value since it depends on the complexity of the signal. Therefore, the original signal $x(n)$ can be reconstructed as a superimposition of estimated IMF plus the residue, i.e.,

$$x(n) = \sum_{j=1}^{M} c_j(n) + r_n(n). \tag{1}$$

In the specific case of honey bee sounds, we have empirically found that it is possible to obtain the original signal setting $M = 10$.

After the EMD decomposition, the Hilbert Transform [26] is applied to each IMF and used for the estimation of the analytic signal $a_j(n)$ as follows:

$$a_j(n) = c_j(n) + j\mathcal{H}\{c_j(n)\} \tag{2}$$

where $\mathcal{H}$ indicates the Hilbert transform and $j = 1, \ldots, M$. Then, equation (2) can be expressed in polar coordinates, i.e., $a_j(n) = A_j(n)e^{i\phi_j(n)}$ where $A_j(n)$ is the instantaneous amplitude of the signal, and $\phi_j(n)$ is the phase from which can be derived the instantaneous frequency $f_j(n) = \frac{f_s}{2\pi}\big[\phi_j(n+1) - \phi_j(n)\big]$. Finally, the spectral features are derived considering the mean normalized frequencies (MNF) calculated as in [27] and the amplitude calculated as a mean of all instantaneous amplitude, obtaining a spectrogram spanning over 10 min.

Fig. 2 compares the three extracted features in relation to a queenless hive state for a time interval of 10 minutes. It is evident how the HHT-based method is capable of expressing the frequency behaviour of the analyzed bee hives.

2.2. Classification

For classification, a first approach employing SVMs and various feature combinations is carried out (Section 3.4). As indicated in [17], SVMs are good classifiers for this problem when used with a radial basis function (RBF) kernel. Here, all SVMs are computed with the RBF kernel, with penalty parameter (C) of 1 and the gamma parameter of 1/(number of features). The input data consists of various combinations of features which are extracted from 10min audio recordings. The samples are normalized using z-score normalization across each training and test sets as described in Section 3.2.

Recent research in the field of computational sound scene analysis shows a clear dominance of data-driven deep learning methods such as CNNs over other traditional machine learning methods [18]. Therefore, we adopt a CNN classifier designed to further explore how applicable these models are for beehive state recognition.

When using CNNs it is important to consider the amount of data needed to train such large networks. To meet these constraints, we segment the original 10min audio samples into 1 min segments. Feature extraction is performed on this new set of shorter samples. To mitigate the loss of temporal context that the shorter segments bring and which was deemed important in [19] for representing beehive sounds, we adapt a procedure to obtain long-term features introduced in [20], where each spectrogram is transformed in a stack of averaged slices over time.

To increase the generalization of the models, we carry out data augmentation on the dataset by creating 3 versions of each sample with a random pitch shift between -1 and 1 semitone. The normalization of the data is performed frequency-wise with z-score normalization along the training set samples. The normalization parameters computed for the training set are then used to apply the same transformation to both validation and test sets.

The general network architecture, presented in Table 1, consists of four convolutional layers (two layers of 16 filters of size 3×3 and two layers of 16 filters of size 3×1) with max pooling, followed by three dense layers (256 units, 32 units and 1 unit). All layers use a leaky rectifier as activation function with the exception of the output layer which uses the sigmoid function.

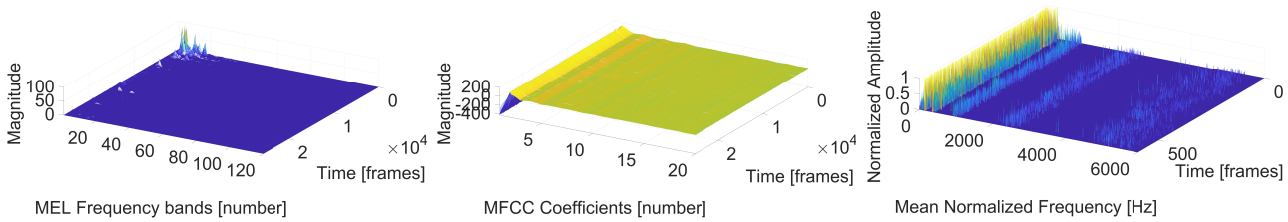

Fig. 2. Comparison of feature extraction: (a) Mel spectra considering the Mel frequency bands, (b) MFCCs considering the obtained coefficients, and (c) HHT-based features considering the extracted mean normalized frequency as function of magnitude/amplitude and time.

Table 1. CNN architecture.

Layer	Size
Input	time frames×freq bands
Conv 1	16 (3×3) filters
Conv 2	16 (3×3) filters
Conv 3	16 (3×1) filters
Conv 4	16 (3×1) filters
Dense 1	256 units
Dense 2	32 units
Dense 3	1 unit

3. EVALUATION

Several experiments are carried out using real-world audio data, with an aim to evaluate the performance of the proposed systems.

3.1. Data

Audio data from the NU-Hive [22] project acquired in honey bee hives is used for training and evaluating the proposed system. In particular, the data from two hives has been used and for each hive a period of one day where the queen bee was present and one day without queen bee has been considered, for a total amount of 576 files of 10 min duration each ($\sim$ 96 hours). As reported in [22], the data was acquired continuously with $f_s = 32$ kHz and the microphones are MEMS type positioned inside the hive, avoiding propolization.

3.2. Experimental setup

Given the interest in constructing a system for a real-world scenario, we evaluate how well the classifiers are able to generalize to unseen hives. Thus, besides randomly splitting the dataset between train and test sets, we also implement a "hive-independent" splitting scheme. This means having training samples belonging only to certain hives, and testing using samples from other, unseen hives.

For the random scheme, a test size of 5% of the total amount of data is used and, when applying the SVM classifier, all remaining data (95%) is used in a single training set. For the CNN implementation, the remaining data is further split in half between the training and validation sets. For the hive-independent scheme, the data is split in two, according to which hive they belong to; one is kept for training and the other for testing. In the case of the CNNs, a validation set, used for early stopping purposes, is obtained by randomly selecting 10% of the training set data.

3.3. Evaluation metrics

The results of each experiment are evaluated using the area under the curve score (AUC) [28]. Each experiment is run twice in different splits, following the same setup and parameters. We report the results on each run and the average AUC over the two.

3.4. SVM Experiments

Several experiments using SVMs with different sets of features are set up and run with the two split configurations reported in Section 3.2:

SVM_MFCCs20: each input sample is a vector of 20 MFCCs resulting from averaging the 10min MFCC coefficients over time.

SVM_HHTdwns20: the HHT spectrogram obtained for a 10min audio recording (see Section 2.1) is aggregated over time and its maximum frequency limited to 6000 Hz. This process results in a vector with 6000 frequency bands representing one 10min recording. The frequency bands are further downsampled into 20 HHT bands, in order to reduce feature dimension.

SVM_MFCCs20_HHTdwns20: a combination of both representations described above is used. In total, a sample corresponding to a 10 min audio recording is represented by a vector of size 40 which is the concatenation of both feature vectors.

SVM_MEL120dwns20: each Mel spectrogram of a 10 min audio recording is averaged over time resulting in a 120-dimensional vector which is further downsampled into 20 bands.

SVM_LOG_MEL120dwns20: the log-Mel spectrograms of the 10 min audio recordings are averaged over time, and the frequency bands downsampled into 20 bands.

3.5. CNN Experiments

Similar to the SVM experiments, the designed CNN model is trained with different features with both random and "hive-independent" splits of the data. As described in Section 2.2, the proposed CNN approach uses spectrograms of 1 min audio data as input. These are computed by processing the audio with a sampling rate of 22.05 kHz, and applying a window size of 2048 samples and hop length of 512 samples. After, the resulting spectrograms are further transformed, as described in Section 2.2, to highlight long-term contextual aspects. In specific, the spectrogram is segmented along the time dimension into 30 slices, each containing approximately 86 time frames ($\sim$2sec). The slices are further averaged over time and stacked together creating a matrix with 30 columns and the same original number of frequency bands. The experiments are:

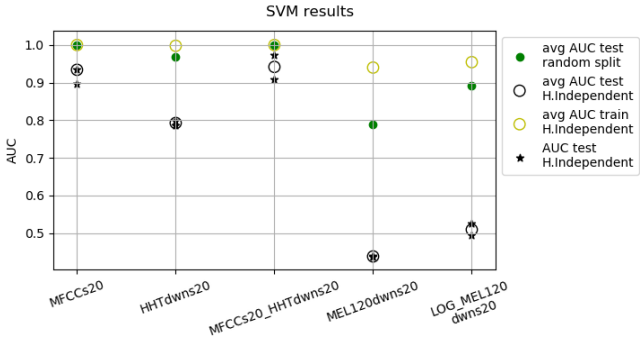

Fig. 3. [SVM results] The ⋆ represents the AUC score on the test set for each fold of the hive-independent setup and SVM experiment. The ○ and ○ represent the average AUC score over the two folds in both train and test sets, respectively. The ● reports the average AUC score of the test sets across the two folds of the random split setup.

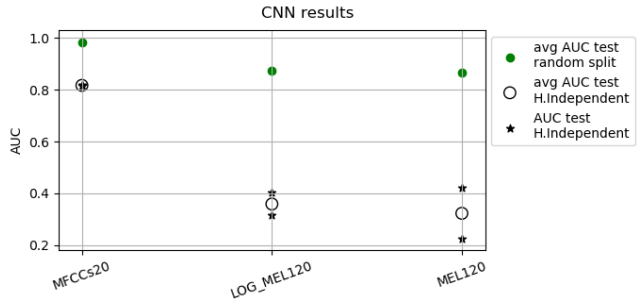

Fig. 4. [CNN results] The ⋆ represents the AUC score on the test set for each fold of the hive-independent setup and CNN experiment. The ○ represents the average AUC score on test sets over the two folds and the ● reports the average AUC score of the test sets across the two folds of the random split setup.

CNN_MFCCs20: Uses as input data 20 MFCCs.

CNN_MEL120: Input samples are obtained from Mel spectra with 120 frequency bands.

CNN_LOG_MEL120: This configuration uses the log-Mel spectra computed with 120 frequency bands.

The network described in Table 1 is trained over 100 epochs on batches of 145 samples with the RMSprop [29] optimizer. Early stopping with a patience value of five and dropout of 50% in the three last layers is employed during training.

3.6. Results

3.6.1. SVM results

The results of the SVM experiments are reported in Fig. 3 with the average AUC scores of the test sets over two folds in both the random split and the "hive-independent" split setups. Additionally, the individual AUC scores of the test set in the "hive independent" split setup are also included. Observing the results for the averaged test AUC score in the random split setup (●), they are consistent with the reported results in [17] that indicate a perfect classification when using SVMs with RBF kernels and MFCCs on a random split setup. Here we further conclude that also the HHTs are good representations that yield similar classification results in this setup.

Given this, and as indicated in [17] and [19], the challenge when working with beehive sounds through a machine learning approach lies on learning classifiers that are able to generalize to different hives. This aspect is confirmed here with the reported lower values of the averaged test AUC scores in the "hive-independent" setup (○) for every experiment conducted. Despite the lower scores for the "hive-independent" setup when compared with random splits, results indicate that the SVMs are successful in generalizing to unseen hives for most feature combinations.

An interesting result is presented in Fig. 3 **[MFCCs20_ HHTdwns20]** where the averaged test AUC score (~ 0.94) in this setup is better than the one reported in **[MFCCs20]** (~ 0.91). These results indicate that the combination of these two features results in better predictions in a "hive-independent" setup than the predictions carried out with classifiers learning from each feature individually. It is also of notice in **[MEL120dwns20]** and **[LOG_MEL120dwns20]** the inadequacy of using these representations together with SVMs for this classification task on the "hive-independent" setup.

3.6.2. CNN results

The resulting AUC scores for the experiments with CNNs are shown in Fig. 4. For the three CNN experiments carried out in a random split setup, the averaged AUC scores (●) show the ability of the procedure in this classification task for both MFCCs and Mel spectra. The challenge in predicting the beehive state in hives different than the ones where the model was trained is evident when observing the resulting averaged test AUC score in a "hive-independent" setup, (○); overall the results in this setup decrease when comparing with the random split. Once again, as Fig. 4 **[MFCCs20]** shows, the use of MFCCs as feature is especially useful in this problem, while Mel spectra when used as features do not appear to generalize well to unseen hives given the resulting AUC lower than 0.5.

4. CONCLUSIONS

This work explored the potential of traditional and neural network-based machine learning methods exploiting MFCCs, Mel spectrograms, and the HHT as feature extractors to determine the presence of the queen bee inside the hive. Several experiments on a real-world scenario have demonstrated the potential of the work exploiting the use of HHT as spectral representations and the design of representations suitable for modelling long-term temporal context. Results using SVMs show their ability to generalize to unseen hives and also the dominance of MFCCs as representations of the data when compared to other features. Better results were obtained when combining HHTs and MFCCs, which is an interesting point for further investigation. The CNNs do not appear to generalize as well to unseen hives in the tested configurations, however they achieve good results in a hive-dependent scenario which indicates the feasibility of the application of deep learning methods to this unique problem, at least in a more controlled supervised scenario.

Future work will further evaluate the methods in a hive independent scenario, for which the dataset must be augmented with new hives. A deeper investigation of CNNs in combination with the HHT and MFCC features is planned; we will also investigate the application of this procedure to the identification of other states of the honey bee hive, including swarming or pest presence. The dataset[1] and python code[2] developed for this work are publicly available.

[1] https://zenodo.org/record/2563940#.XGVwpDP7SUk

[2] https://github.com/madzimia/Audio_based_identification_beehive_states.git

5. REFERENCES

[1] Alexandra-Maria Klein, Bernard E Vaissière, James H Cane, Ingolf Steffan-Dewenter, Saul A Cunningham, Claire Kremen, and Teja Tscharntke, "Importance of pollinators in changing landscapes for world crops," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 274, no. 1608, pp. 303–313, 2007.

[2] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Acoustic detection of human activities in natural environments," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 686–695, 2012.

[3] H. Frings and F. Little, "Reactions of honey bees in the hive to simple sounds," *Science*, pp. 122–125, 1957.

[4] A. Michelsen, W. H. Kirchner, and M. Lindauer, "Sound and vibrational signals in the dance language of the honeybee, apis mellifera," *Behavioral Ecology and Sociobiology*, vol. 18, no. 3, pp. 207–212, Jan. 1986.

[5] W. H. Kirchner, "Acoustical communication in honeybees," *Apidologie*, vol. 24, no. 3, pp. 297–307, 1993.

[6] M. Hrncir, F. G. Barth, and J. Tautz, "Vibratory and airborne sound-signals in bee communication," in *In Insect Sounds and Communication: Physiology, Behaviour, Ecology, and Evolution*, S. Drosopoulos and M. Claridge, Eds., pp. 421–436. CRC Press, 2006.

[7] J. H. Hunt and F. J. Richard, "Intracolony vibroacoustic communication in social insects," *Insectes Sociaux*, vol. 60, pp. 403–417, 2013.

[8] T. Cejrowski, J. Szymański, H. Mora, and D. Gil, "Detection of the bee queen presence using sound analysis," in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds., pp. 297–306. Springer International Publishing, 2018.

[9] A. Robles, T. Saucedo-Anaya, E. Gonzlez-Ramrez, and C. Galvn Tejada, "Frequency analysis of honey bee buzz for automatic recognition of health status: A preliminary study," *Research in Computing Science*, vol. 142, pp. 89–98, June 2017.

[10] D.G. Dietlein, "A method for remote monitoring of activity of honeybee colonies by sound analysis," *Journal of Apicultural Research*, vol. 24, no. 2, pp. 176–183, 1985.

[11] S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, "Monitoring of swarming sounds in bee hives for prevention of honey loss," in *International Workshop on Smart Sensors in Livestock Monitoring*, Sept. 2006.

[12] S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, "Monitoring of swarming sounds in beehives for early detection of the swarming period," *Computers and Electronics in Agriculture*, vol. 65, pp. 72–77, 2008.

[13] A. Qandour, I. Ahmad, D. Habibi, and M. Leppard, "Remote beehive monitoring using acoustic signals," *Acoustics Australia / Australian Acoustical Society*, vol. 42, no. 3, pp. 204–209, Dec. 2014.

[14] J. J. Bromenshenk, "Honey bee acoustic recording and analysis system for monitoring hive health," 2007, US Patent 7549907.

[15] Dan Stowell, "Computational bioacoustic scene analysis," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. P. W. Ellis, Eds., pp. 303–333. Springer, 2018.

[16] P. Amlathe, "Standard machine learning techniques in audio beehive monitoring: Classification of audio samples with logistic regression, K-nearest neighbor, random forest and support vector machine," M.S. thesis, Utah State University, 2018.

[17] I. Nolasco, "Audio-based beehive state recognition," M.S. thesis, Queen Mary University of London, 2018.

[18] "DCASE Challenge 2017, Task1 results," `http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results`.

[19] I. Nolasco and E. Benetos, "To bee or not to bee: Investigating machine learning approaches for beehive sound recognition," in *2018 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018, pp. 133–137.

[20] V. Bisot, R. Serizel, and S. Essid, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[21] N. Huang, "Introduction to the Hilbert-Huang transform and its related mathematical problems," in *Hilbert-Huang Transform and Its Applications*, pp. 1–26. World Scientific, Sept. 2005.

[22] S. Cecchi, A. Terenzi, S. Orcioni, P. Riolo, S. Ruschioni, and N. Isidoro, "A preliminary study of sounds emitted by honey bees in a beehive," in *Audio Engineering Society Convention 144*, May 2018.

[23] Jake Burton, David Frank, Mahdi Saleh, Nassir Navab, and Helen L. Bear, "The speaker-independent lipreading play-off; a survey of lipreading machines," in *IEEE International Image Processing Applications and Systems (IPAS)*, 2018.

[24] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Acoustic features for environmental sound analysis," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. P. W. Ellis, Eds., pp. 13–40. Springer, 2018.

[25] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[26] L. Marple, "Computing the discrete-time "analytic" signal via fft," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, Sept 1999.

[27] A. Abdelouahad, A. Belkhou, A. Jbari, and L. Bellarbi, "Time and frequency parameters of semg signal force relationship," in *2018 4th International Conference on Optimization and Applications (ICOA)*, April 2018, pp. 1–5.

[28] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds., pp. 147–179. Springer, 2018.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.