ISSN 1870-4069

Frequency Analysis of Honey Bee Buzz for Automatic Recognition of Health Status: A Preliminary Study

Antonio Robles-Guerrero[1], Tonatiuh Saucedo-Anaya[2], Efrén González-Ramírez[1], Carlos E. Galván-Tejada[1]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Mexico

[2] Universidad Autónoma de Zacatecas, Unidad Académica de Física, Zacatecas, Mexico

aroblesp@uaz.edu.mx, gonzalez_efren@hotmail.com, ericgalvan@uaz.edu.mx, tsaucedo@fisica.uaz.edu.mx

Abstract. The study of honey bee health has received special attention in the last years. Researchers has been monitoring physical variables to determinate the status of the colony. This is a first approach in the development of a real time monitoring system to provide useful information to beekeepers that will help them to prevent colony losses. This study presents an analysis of the sound from two colonies of bees in the Mel frequency domain. The first is a healthy colony with queen and the second one is a hive with no queen and with a reduced population. Sound samples were acquired for each colony and characterized using Mel Frequency Cepstral Coefficients (MFCC). To summarizes the information, statistical descriptors was obtained for each Mel coefficient. An exploratory analysis of samples revealed two different hive characteristics; the presence and lack of a queen bee. For honey bee buzz recognition, a Logistic Regression Model was used. The preliminary results show that it is possible to classify both characteristics obtaining high classification rates using a reduced set of features.

Keywords: honey bee, remote sensing, beehive monitoring, queenless state.

1 Introduction

Pollinators are essential for diet diversity, biodiversity, and the maintenance of natural resources. The honey bee is the most important pollinator. Approximately 73% world cultivated crops depend on some variety of bees [1]. The colony health is influenced by external factors, such as, increase of pathologies,

pollution, pesticides, among others. Monitoring honey bee health is an important task for beekeepers. Early detection of health status can be crucial to ensure the survival of the colony.

A queen bee plays an important role in a colony, she controls workers by releasing pheromones and produces eggs. Lost of the queen can result in the dead of the whole colony in a few months, unless a new queen is introduced.

In the last years researchers have been looking for non invasive methods for continuous monitoring and automatic detection of honey bee health status. Special attention has received the monitoring of physical variables, such as, temperature, humidity, sound, vibrations, colony weight, and gas contents [22]. The sound in bee hives has been analyzed for detecting the swarming period. Swarming is characterized by an increase of the power spectral density before it takes place. In [14], a method for predicting the swarming period is proposed based on labeling the sounds. [12], concluded that the noise generated by bees has a high probability of correspondence to the physiological state. There are patented devices to determinate the honey bee health by comparing a captured hive sound with known acoustic fingerprints of a healthy colony [5].

Changes in sound due to Varroa mite infestation have been investigated by [19], where his prediction accuracy is claimed to be better than any random guessing, although results are not validated. The queenless state has been investigated by using spectrograms [16]. Analysis results were classified by using a Kohonen Self Organising Map and artificial neural networks. Although [16], found the frequency characteristics for each condition their results were not satisfactory.

The aim of this research work is to propose a methodology based on MFCC and Machine Learning for automatic recognition of the status of a honey bee colony based on sound recording, The proposed methodology can be implemented in dedicated devices to detect the presence or absence of the queen bee avoiding invasive inspection of the colony. Furthermore, more conditions can be analyzed by using the same estrategy.

The rest of the paper is organized as follows: in section 2 a detailed description of the data set acquisition and the methodology for feature extraction is presented. In section 3, the process for feature selection and model validation is described. The paper ends in section 4 by presenting conclusion and future work.

2 Materials and Methods

2.1 Honeybee Monitoring System

Based on previous works [10,7,11], a monitoring system was developed based on a Raspberry Pi 2 model B, figure 1. Sound samples were acquired by using omnidirectional electret microphones placed inside the hives. Microphones were protected by a metallic mesh to avoid them begin covered with wax. The main idea was keeping the system as simple as possible with only a microphone by

hive. The signal was acquired and converted to a digital signal by using a dspic microcontroller with a 12-bits resolution ADC. The system was 10000 mAh battery powered allowing an autonomy of about 24 hrs.

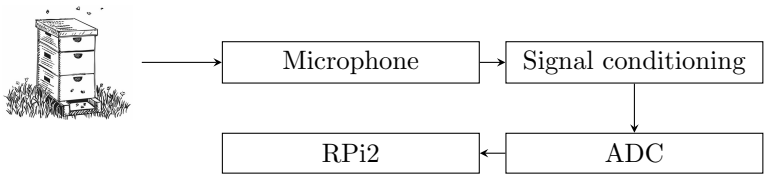


Fig. 1. System overview.

2.2 Data Set Description

Sound samples were extracted from two colonies of Carniolan honey bee (*Apis mellifera carnica*). The first colony has a queen with a large population, and the second one is a queenless colony with a reduced population. Colony population were compared only by visual examination, however a quantitative method for measuring the population is necessary. The beehives were protected with insulation material against low temperatures (low temperature is quite common in Zacatecas city).

Various researchers have reported the range of frequencies of the acoustic signals produced by a honey bee colony are in the range from 100 to 1 kHz [2,18], and that most of the sound have frequencies around 300, 410 and 510 Hz [9]. The sampling frequency for this investigation was set up to 4 kHz to get a good quality representation of the sound activity without increasing the storage requirements; this is the double of the minimum of Nyquist frequency required. [18] showed that frequency of the sound changes slowly along the day. To evaluate the evolution of the frequencies, 3 min of sound of the honey bee buzz were recorded every 15 min for 24hrs. The experiment for data acquisition was carried out during 45 days (beginning of mid-April to May).

2.3 Feature Extraction

MFCC is one of the most important technique for feature extraction in speak recognition. Although MFCC is used for human sound perception, in this work it is proposed for sound bee characterization because of its effectiveness reported in others areas (i.e. sound genre classification [21] and environmental sounds recognition [8]). The MFCC transforms the raw signal into a compact series of parameters representing the original signal. Figure 2 shows the process of feature extraction: (i) the wave form is first passed through a pre-emphasis filter, (ii) the signal is divided into frames of short duration (typically 25 ms), (iii) each frame

is multiplied by a hamming windows, (iv) the Fast Fourier transform (FFT), is calculated for each frame, (v) the power spectrum is warped according the Mel-scale, (vi) the spectrum is segmented according to a triangular filter bank, and finally (vii) the coefficients are computed by applying an Discrete Cosine Transformation (DCT), to the logarithm of the filter bank output.

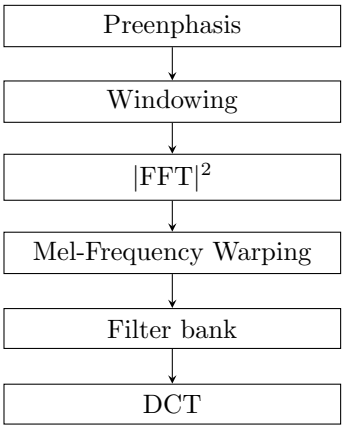


Fig. 2. MFCC feature extraction methodology.

3 Results

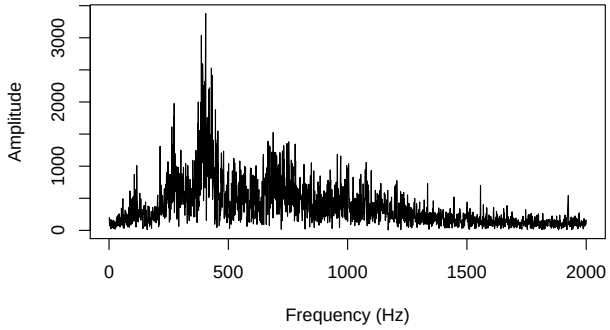
3.1 Feature Extraction and Preprocessing

The samples shown in Figure 3, correspond to sound recorded during the afternoon of a spring day in May. Figure 3(a) shows the frequency bands of the healthy colony; most of the sound activity is present around 400 Hz. On the other hand, in figure 3(b), the queenless colony present a different pattern; the emitted sound is distributed in more frequency bands.

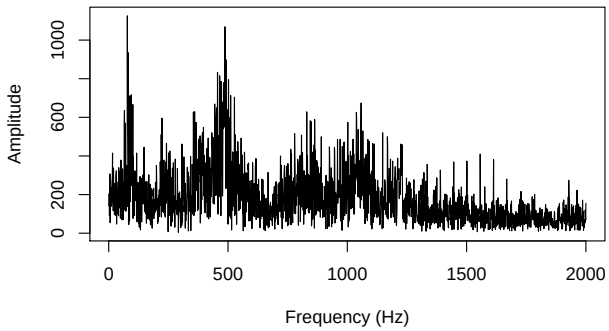
The MFCC were computed from 88 instances of sound captured from the colony with queen and 98 from the queenless colony. These data come from 24 hrs of recording. The difference between recorded instances were due to the non-equal battery life of the two recorders. The parameters of the MFCC calculation were: window size of 25 ms (in this period of time the signal is considered quasi stationary), and with 10 ms of overlapping, the pre-emphasis value was set to 0.94. Those values are typically used in speech recognition.

After the MFCC extraction were carried out, and in order to reduce the data set size and computational cost for each Mel coefficient, the following statistical descriptors were computed:

– mean,



(a) Healthy colony.



(b) Unhealthy colony.

Fig. 3. Comparison of the frequency spectrum of the healthy and unhealthy colonies.

- trimmed mean (20%),
- kurtosis,
- standard deviation,
- skewness,
- median,
- variance,
- coefficient of variation,
- quantiles (2.5, 25, 50, 75, 97.5).

The resulting data set is composed of 168 features and 186 instances. Two classes were chosen in this work: healthy and unhealthy colony.

After feature extraction standardization was performed over data; it scales data in function of the mean (μ) and the standard deviation (σ):

$$z = \frac{x - \mu}{\sigma}. \tag{1}$$

This process reduces the effects of the different distributions [6].

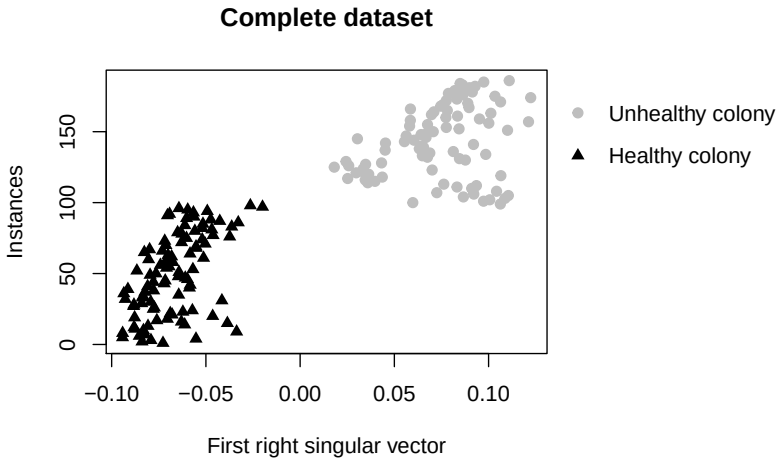


Fig. 4. SVD of the complete data set.

3.2 Feature Selection

An exploratory analysis was conducted on R Project software [20]. The first analysis carried out was a Singular Value Decomposition (SVD). SVD computes the set of eigenvalues and eigenvectors of a matrix. It is a common technique in the analysis of multivariate data that can reveal structures in the data set that may be useful for classification. The SVD analysis of the data set is shown in figure 4. The black triangle (▲) corresponds to the colony with queen and the gray circle (•) is the queenless colony. The samples are grouped forming well defined clusters, evidencing two conditions clearly distinguishable.

Not all the features are useful. In order to find features that best describe the conditions of the hives, a SVD analysis was conducted on each statistical descriptor (figure 5). Mean, trimmed mean, median, coefficient of variation and quantiles, form similar clusters. In the rest of the descriptor the cluster are mixed, and they might not be useful to make a good prediction.

By visual examination it was determined that one statistical descriptor is enough to make a good prediction; the mean values of the MFCC were selected. The free package *randomForest* (v4.6-12) [17] R Project Software was used to evaluate the importance of each feature in mean descriptor. Random Forest is a combination of tree predictors such that each of them depends on the values of a random vector sampled independently with the same distribution in the forest [4]. The result of the random forest evaluation is shown in Figure 7. The most important features are mean of Mel coefficient 4 and 10.

3.3 Validation

In order to validate the model the dataset was divided into two parts: a training set (70%), that was used to train a Logistic Regression model and a test set

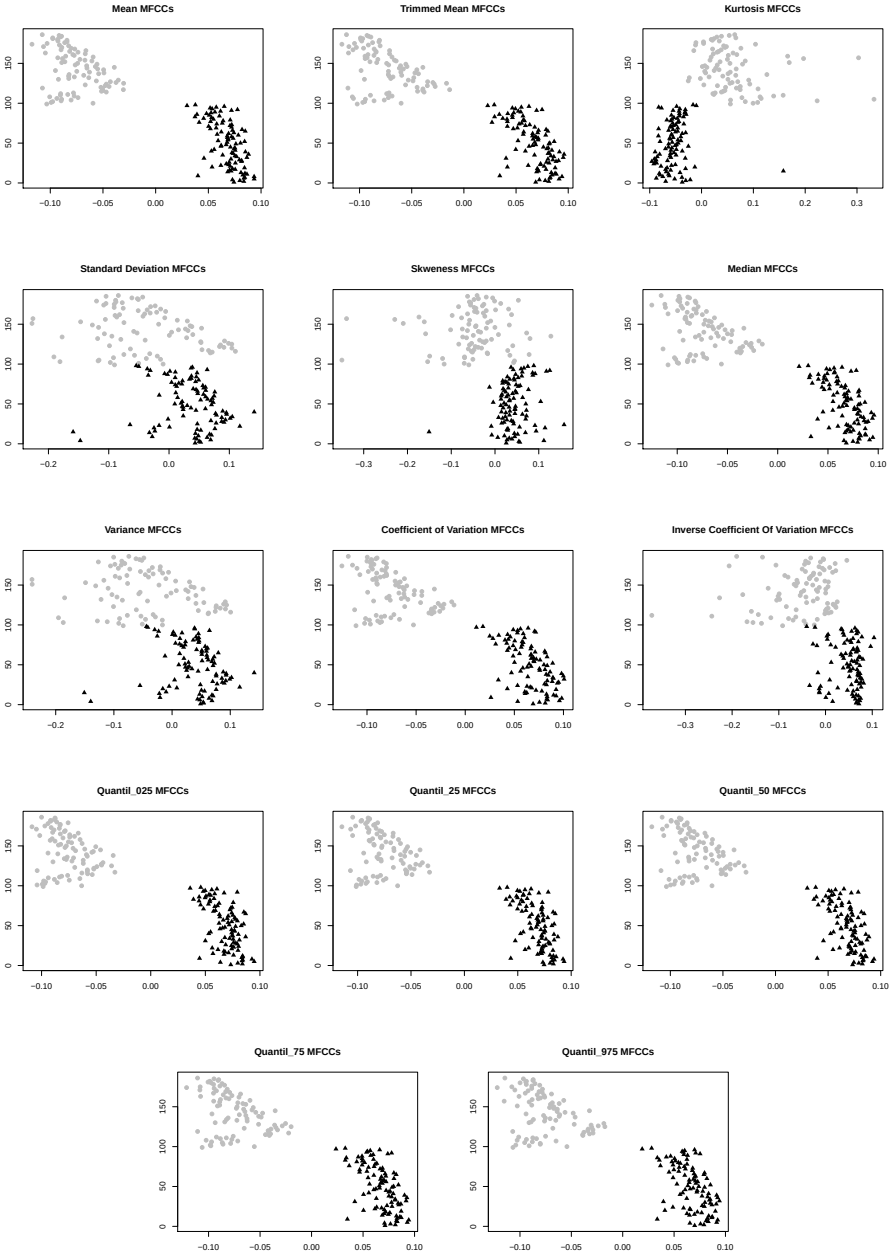


Fig. 5. Singular value decomposition of each statistical descriptor, the black triangle (▲) corresponds to healthy colony and gray circle • to unhealthy colony.

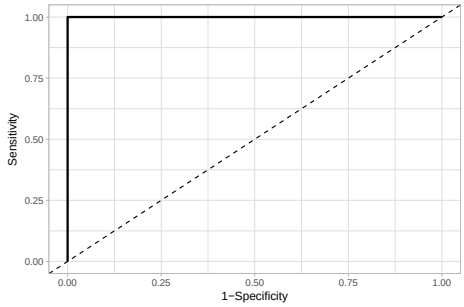


Fig. 6. ROC curve using two features.

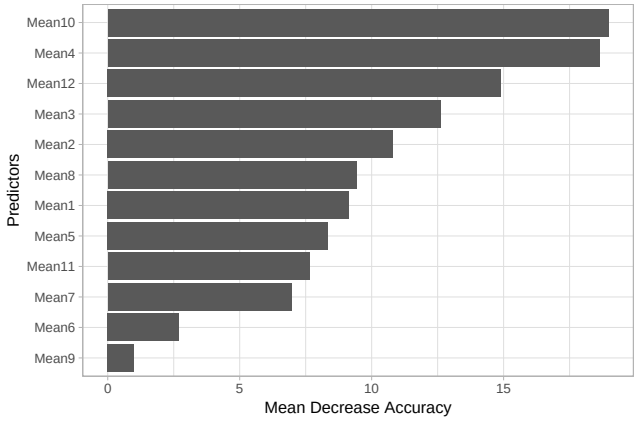


Fig. 7. Features in order of importance.

(30%), that was used to measure the model performance predictor. To evaluate the model performance a ROC curve was obtained. Receiver Operative Characteristics (ROC), analysis is a well known model performance measure for machine learning algorithms [13,15,3]. The ROC curve is a plot of true positives against false positive. The plot shows the number of correctly classified samples versus the number of incorrectly classified negative samples. A perfect classifier is reflected by a curve which lies in the upper left corner with an unitary area under the ROC curve. In figure 6, it is shown the ROC curve of the model. To achieve a perfect classification only two features were enough; mean values of Mel coefficients 4 and 10.

4 Conclusion and Future Work

The proposed methodology based on Mel Frequency Cepstral Coefficients and machine learning algorithms suggests that it can be effectively used for honey

bee status recognition by sound analysis. However, data analysis from more hives are needed in order to confirm these results.

The FFT of the sound signal from a beehive with queen shows a characteristic pattern around 400 hz that is different from that obtained from a beehive with no queen. The patter found in a queenless colony shows a different frequency distribution. Data for this research were obtained during a month and a half period.

In the model validation only two features were necessary to achieve a perfect prediction, however, more research with more beehives is needed for a better evaluation of the predictor performance

Future work will include the reduction of the number of instances and the recording time to have optimal values. This will reduce the storage space required and the computational cost on a dedicated device. It is also important, as future work, to find otjer patterns of specific health status of honey bees; such as, pre-swarming behavior, varroa mite infection, size population, among others. With this information it will be possible to create a database and a system able to identify the health status of a colony.

Acknowledgments. We thank for the partial support in the development of this project to CONACYT.

References

1. Abrol, D.P.: Pollination Biology: Biodiversity conservation and agricultural production. Springer Netherlands, Dordrecht, 1 edn. (2011), http://www.sciencedirect.com/science/article/pii/B9780125839808500193 http://link.springer.com/10.1007/978-94-007-1942-2
2. Bencsik, M., Bencsik, J., Baxter, M., Lucian, A., Romieu, J., Millet, M.: Identification of the honey bee swarming process by analysing the time course of hive vibrations. Computers and Electronics in Agriculture 76(1), 44–50 (2011), http://dx.doi.org/10.1016/j.compag.2011.01.004
3. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997), http://linkinghub.elsevier.com/retrieve/pii/S0031320396001422
4. Breiman, L.: Random Forests. Mach. Learn. 45(1), 5–32 (2001), https://doi.org/10.1023/A:1010933404324
5. Bromenshenk, J.J., Henderson, C., Seccomb, R., Rice, S., Etter, R.: Honey bee acoustic recording and analysis system for monitoring hive health (2009), http://www.google.com/patents
6. Brownlee, J.: Machine Learning Mastery with R. Melbourne, Australia (2017)
7. Chen, W.S., Wang, C.H., Jiang, J.A., Yang, E.C.: Development of a monitoring system for honeybee activities. In: Proceedings of the International Conference on Sensing Technology, ICST. vol. 2016-March, pp. 745–750. Taiwan (2016)
8. Chu, S., Narayanan, S., Kuo, C.C.J.: Environmental Sound Recognition With Time-Frequency Audio Features. IEEE Transactions on Audio, Speech, and Language Processing 17(6), 1142–1158 (aug 2009), http://ieeexplore.ieee.org/document/5109766/

9. Dietlein, D.G.: A method for remote monitoring of activity of honeybee colonies by sound analysis. Journal of Apicultural Research 24(3), 176–183 (1985)
10. Edwards-Murphy, F., Magno, M., O'Leary, L., Troy, K., Whelan, P., Popovici, E.M.: Big brother for bees (3B) - Energy neutral platform for remote monitoring of beehive imagery and sound. In: Proceedings - 2015 6th IEEE International Workshop on Advances in Sensors and Interfaces, IWASI 2015. pp. 106–111 (2015)
11. Edwards-Murphy, F., Srbinovski, B., Magno, M., Popovici, E.M., Whelan, P.M.: An automatic, wireless audio recording node for analysis of beehives. 2015 26th Irish Signals and Systems Conference, ISSC 2015 pp. 1–6 (2015)
12. Eskov, E.K., Toboev, V.A.: Changes in the structure of sounds generated by bee colonies during sociotomy. Entomological Review 91(3), 347–353 (2011)
13. Fawcett, T.: An introduction to ROC analysis environments support. Pattern Recognition Letters 27(8), 861–874 (2006), http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X
14. Ferrari, S., Silva, M., Guarino, M., Berckmans, D.: Monitoring of swarming sounds in bee hives for early detection of the swarming period. Computers and Electronics in Agriculture 64(1), 72–77 (2008)
15. Hand, D.J., Till, R.J.: A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45(2), 171–186 (2001)
16. Howard, D., Duran, O., Hunter, G., Stebel, K.: Signal Processing the acoustics of honeybees (APIS MELLIFERA) to identify the "queenless" state in Hives. Proceedings of the Institute of Acoustics 35, 290–297 (2013)
17. Liaw, A., Wiener, M.: Classification and regression by randomForest. R News 2(3), 18–22 (2002), http://cran.r-project.org/doc/Rnews/
18. Pérez, N., Jesús, F., Pérez, C., Niell, S., Draper, A., Obrusnik, N., Zinemanas, P., Spina, Y.M., Letelier, L.C., Monzón, P.: Continuous monitoring of beehives' sound for environmental pollution control (2016)
19. Qandour, A., Ahmad, I., Habibi, D., Leppard, M.: Remote Beehive Monitoring Using Acoustic Signals. Acoustics Australia 42(3), 204–209 (2014)
20. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017), https://www.r-project.org/
21. Tzanetakis, G., Cook, P.: Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing 10(5), 293–302 (2002)
22. Zacepins, A., Brusbardis, V., Meitalovs, J., Stalidzans, E.: Challenges in the development of Precision Beekeeping. Biosystems Engineering 130, 60–71 (2015)