forum acusticum 2023

# USING A NEURAL NETWORK BASED VOCALIZATION DETECTOR FOR BROILER WELFARE MONITORING

**Pieter Thomas**[1*] **Tomasz Grzywalski**[1] **Yuanbo Hou**[1]
**Patricia Soster de Carvalho**[2,3] **Maarten De Gussem**[3,4] **Gunther Antonissen**[2]
**Frank Tuyttens**[5,6] **Paul Devos**[1] **Dick Botteldooren**[1]
[1]WAVES-Acoustics, Department of Information Technology, Ghent University, Belgium
[2]Department of Pathobiology, Pharmacology and Zoological Medicine, Ghent University, Belgium
[3]Poulpharm BVBA, Izegem, Belgium
[4]Vetworks BVBA, Poeke, Belgium
[5]Department of Veterinary and Biosciences, Faculty of Veterinary Medicine, Ghent University,Belgium
[6]Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium

## ABSTRACT

The poultry industry in Flanders, Belgium, is characterized by highly efficient production systems. To assure sustainable food production, where broiler quality of life is key, data collection systems and (automated) interpretation of broiler health and welfare are of capital importance. This paper describes the realization and deployment of an acoustic detector for broiler vocalizations, as part of a larger set of behavior and welfare monitoring tools developed within the ICON-WISH project. The vocalization detector is based on a convolutional neural network. For training, a labelled library with vocalizations is built (>2k samples), based on a large set of broiler audio recordings covering the full broiler life-span. Four different types of vocalizations (pleasure notes, distress calls, short peeps and warbles) are identified in function of broiler age to account for spectral changes. Based on this library, the neural network achieves a balanced accuracy of 87.9%. To indicate its potential, the detector is applied in a real-life medium scale housing, in which several groups of broilers are exposed to different environmental conditions (heat stress a.o.). The occurrence and type of vocalizations is analyzed, and the potential to identify broiler stress is investigated.

**Keywords:** *broiler welfare, vocalization detection, sound recognition*

[*]***Corresponding author***: *pieter.thomas@ugent.be*


## 1. INTRODUCTION

In 2019, about 220M broiler chickens were raised in 486 commercial broiler farms in Flanders, Belgium, with a projected global compound annual growth rate of 7% by 2025. This increase of production due to further intensification shows diminishing returns because of health and welfare issues.

Current methods for assessing broiler chicken behavior, welfare and detecting diseases rely on subjective and labor-intensive data collection by expensive domain experts, conducted during a short (several hours), not always representative time-frame. Such an approach leads to delayed interventions and incomplete/incorrect data interpretation and is not cost effective.

The ICON-WISH project aims to increase productivity by developing Precision Livestock Farming systems that automatically collect data and interpret broiler behavior and welfare both under experimental and commercial farm

conditions. This will reduce the productivity losses, improve objectivity and allow for real-time continuous observations and immediate interventions.
In the scope of this project, a broiler vocalization detector has been developed as part of a larger set of sensory equipment (including video and motion detection) to detect broiler behavior, stress and (in a further stage) health (respiratory diseases). This paper describes the construction of such a vocalization detector based on a convolutional neural network, trained with a broiler vocalization database. The vocalization detector is then evaluated in a medium scale housing to evaluate the effects of heat stress.

## 2. BROILER VOCALIZATIONS

Broilers have the ability to communicate through vocal signs, which makes sound analysis a useful tool for monitoring their behavior and biological responses to external stimuli [1]. Four different types of vocalizations are identified. Distress calls are characterized by repetitive and high-energy vocalizations [2-4], whereas short peeps are identified as low-energy and short duration vocalizations with descending energy [4]. Pleasure notes, on the other hand, are described as vocal expressions with low energy that tend to swing upward in pitch, with an ascending frequency and a short duration [4]. Warble notes can be either ascending or descending in frequency, and are characterized by a repetitive, bow-like vocalization with low energy [4].

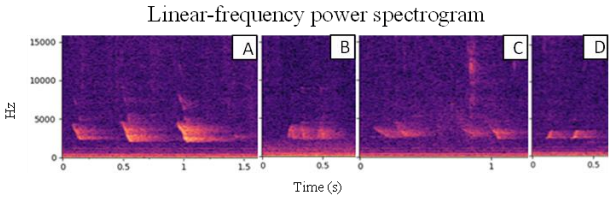


**Figure 1**. Linear-frequency power spectrogram of distress call (A), warbles (B), short peep (C), and pleasure notes (D).

Fontana et al. [5] found that the average peak frequency of the vocalization is inversely proportional with broiler age and weight. The mean peak frequency at the first day is at 3.6kHz, to decrease to 1.5kHz after 36days. This frequency dependency is taken into account in the model by generating a labelled library for each week of the broiler life-span.

## 3. BROILER VOCALIZATION DETECTOR

In order to build a broiler vocalization detector a database of vocalizations has been developed. The development of the database consisted of three main stages: (1) acquisition of audio samples from broiler pens, (2) extraction of audio samples containing broiler vocalizations using an existing general-purpose neural network model for audio event detection and (3) manual classification of broiler sounds into one of 5 predefined classes (4 types of broiler vocalizations and other, non-broiler related, sounds). Next, a custom neural network was designed and trained (in a supervised manner) to perform the automatic classification of broiler sounds. The following chapters describe the details of this process.

### 3.1 Construction of a broiler vocalization database

*3.1.1 Audio stream acquisition*

The base audio is obtained from recordings at Poulpharm (Izegem, Belgium) during the whole lifetime (42 days) of broilers. The study involved a group of 10 broilers that were housed in the same pen (size 1,1m x 2,1m) in a room isolated from other broilers and the outside environment (Figure 2). A microphone was placed in the center of the pen at 90cm height to record the chickens' vocalizations at 48kHz sample rate. Due to a power failure, however, the last days of recordings were lost.
As such, this resulted in 37 days of continuous recordings that could not be assessed directly by a human expert. Hence potential relevant samples were extracted in a first stage.

**Figure 2**. Recording setup at Poulpharm (Izegem, Belgium) for broiler vocalizations. The position of the microphone is indicated.

*3.1.2 Sample extraction*

Similar to previous work [6], this paper uses the pure-convolution-based pre-trained audio neural networks (PANNs) [7] to tag the audio stream recorded on site. PANNs with 80.75 M parameters are trained on the large-scale AudioSet [8] and results in 527 classes of audio events, which covers most of the real-life audio events. Specifically, PANNs have six convolutional blocks. Each convolutional block contains two convolutional layers with a kernel size of (3 × 3). Batch normalization and ReLU activation functions are used to accelerate and stabilize the training. Next, a linear dense layer is applied to the high-level representations learned by the convolutional blocks, followed by an event classification dense layer with the sigmoid activation function to recognize the 527 classes of audio events.

Labels assigned by PANNs that could indicate vocalization of chickens were selected, e.g. chicken, bird, etc. When the probability of these labels at any given second exceeded a threshold of 0.2, this second was tagged. Then sound samples were cut from the audio stream by including one second before and one second after the tagged interval. Overlapping samples were merged, and only sound samples containing a fragment with a probability of at least 0.3 were kept to further reduce the amount of data. This resulted in samples with a length between 3 and 45 seconds, containing one or several possible vocalizations.

*3.1.3 Sample labelling*

Subsequently, auditive evaluation and visual inspection of the linear-frequency spectrogram were conducted to manually label each sample. The samples were then classified into four distinct types: distress calls, short peeps, warbles and pleasure notes. In cases where none of these vocalizations were detected, the sample was labeled as "other sound". If a file contained two or multiple types of vocalizations, it was excluded from the analysis and classified as a "combination of sounds." The vocalizations were classified on a weekly basis (from 1 to 5), enabling to accurately categorize each vocalization according to its type and age.

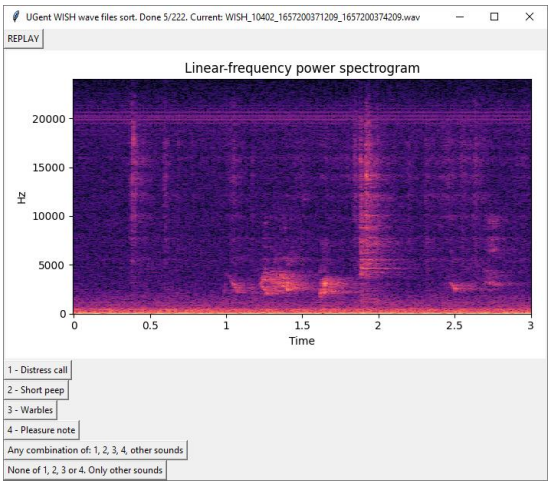


**Figure 3**. Desktop application used for manual categorization of broiler vocalizations.

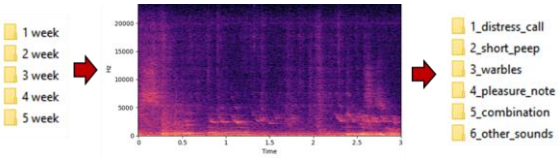


**Figure 4**. Broiler vocalization evaluated per week and classified into distress call, short peep, warbles, pleasure notes, combination or other sounds.

Manual labeling of the extracted samples yielded a database of broiler vocalizations of a total of 2559 audio recordings of a total duration of 3 hours and 10 minutes. Detailed breakdown of the database is depicted in Table 1.

**Table 1.** Composition of the broiler vocalizations database (number of audio samples and total duration).

| Sound type | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 |
|---|---|---|---|---|---|
| Distress calls | 188 (1377s) | 48 (224s) | 136 (900s) | 55 (285s) | 59 (282s) |
| Pleasure notes | 685 (3183s) | 67 (251s) | 51 (184s) | 0 | 0 |
| Short peeps | 51 (218s) | 212 (960s) | 154 (799s) | 151 (595s) | 54 (214) |
| Warbles | 24 (66s) | 85 (214s) | 96 (286s) | 49 (127s) | 52 (125s) |
| Other sounds | 83 (314s) | 22 (84s) | 59 (187s) | 128 (421s) | 50 (154s) |

### 3.2 Architecture and performance of the recognizer

*3.2.1 Neural network architecture*

The proposed neural network accepts on input a time-frequency representation on the input signal in the form of a log-mel spectrogram. In particular, the input signal, after being resampled to 16kHz, is converted to a spectrogram using 512-point Short-Time Fourier Transform with window of the same size, 160 samples (10ms) hop length and Hann window. The STFT is then processed with 64 mel filters that span the frequency range from 50Hz to 8kHz. Finally, the mel spectrogram is converted to decibel scale and linearly scaled to fit the (-1, 1) values range.

The neural model has been designed as a fully-convolutional neural network with 11 2D convolutional layers and one 1D convolutional layer. Each 2D convolutional layer uses ELU activation and is followed by a batch normalization that speeds up model convergence. The output 1D convolutional layer has an effective stride of 24 spectrogram frames (240ms). It consists of six neurons: the first five generate logits related to the broiler vocalization classification and the final neuron estimates normalized broiler age. The addition of broiler age as a training target helps the model to discover and understand the tight dependency between the chicken age and vocalization pitch which improves its accuracy.

The effective receptive field of the model in time axis is 194 frames. Because convolutional layers do not use padding (they are all valid convolutions), this is also the minimal length of the spectrogram required to obtain a prediction. The same length of the spectrogram was also used during the pre-training and fine-tuning. For detailed breakdown of the parameters of convolutional layers please refer to Table 2.

**Table 2.** Configuration of the neural network's convolutional layers and spatial resolution of output activation matrixes for a minimum-sized input of 194 spectrogram frames and 64 frequency bands. All dimensions are provided in following order: frequency, time.

| Layer type | Filters | Kernel | Strides | Output shape |
|---|---|---|---|---|
| Conv 2D | 64 | 3, 3 | 1, 1 | 62, 192 |
| Conv 2D | 64 | 3, 3 | 1, 1 | 60, 190 |
| Conv 2D | 96 | 4, 4 | 2, 2 | 29, 94 |
| Conv 2D | 96 | 3, 3 | 1, 1 | 27, 92 |
| Conv 2D | 128 | 3, 5 | 2, 3 | 13, 30 |
| Conv 2D | 128 | 3, 3 | 1, 1 | 11, 28 |
| Conv 2D | 128 | 3, 4 | 2, 2 | 5, 13 |
| Conv 2D | 128 | 3, 3 | 1, 1 | 3, 11 |
| Conv 2D | 128 | 3, 3 | 1, 1 | 1, 9 |
| Conv 2D | 128 | 1, 3 | 1, 2 | 1, 4 |
| Conv 2D | 128 | 1, 4 | 1, 1 | 1, 1 |
| Conv 1D | 6 | 1 | 1 | 1 |

The model consists of a total of 1.16 million trainable parameters and is able to processes 60 seconds of audio signal in one second on a modern multicore CPU.

*3.2.2 Pre-training on AudioSet*

Before the actual training on broiler sounds, the proposed neural network was first pre-trained on AudioSet [8]. The pre-training used a sample of 100k recordings from the AudioSet database that is focused on bird and fowl sounds. In particular, the sample included all available AudioSet recordings that were labeled with a "Bird" or "Fowl" label or any of their 20 sub-labels, which accounted for about 35k of recordings. The additional 65k recordings were selected to uniformly represent the remaining 505 AudioSet labels.

During the pre-training the final 1D convolutional layer of the neural network was temporarily replaced with a 527-neuron layer with sigmoid activation. The pre-training was conducted for 100 epoch using Adam optimizer with a learning rate of 0.001 and Mean Squared Error loss. Training batches included 32 samples. During training, a Mean Average Precision was controlled on a small (3%) held-out validation dataset and the best model weights were saved.

*3.2.3 Broiler vocalizations pre-processing*

Before being used for neural model training, broiler vocalization recordings have been cleaned from background noise using a state-of-the-art nonstationary noise suppression algorithm [9]. This effectively removed ventilation hum and background babble from other poultry farm animals and staff. The main motivation for removing the background noise was to obtain clean vocalizations and to prevent model from finding false relations in the input signals not related to welfare of the experimental broilers.

*3.2.4 Fine-tuning for broiler vocalization detection*

In this stage the model, previously pre-trained on AudioSet, was fine-tuned to detect broiler vocalizations using the broiler vocalization database. The fine-tuning was split into two steps. In the first step only the output 1D convolutional layer was trained while weights of the 2D convolutional layers remained fixed.

In the next step five copies of the model (and its weights) were created and combined into one meta-model with 5 branches (Figure 5). The five individual models became branches of the meta-model. All branches were identical except for the number of initial 2D convolutional layers whose weights remain fixed (frozen) during fine-tuning. In particular, the first branch had no weights fixed, in the second branch the first convolutional layer's weights were fixed, in the third branch two convolutional layers' weights were fixed and so on.

The idea of using the branched model with different number of fixed layers in each branch is similar to inception [10]. Like inception module, our branched model exploits different views on the same data because in each branch the predictions are being made based on features of different complexity: branch number one can optimize all features for the target task, but branch number five must build predictions based on medium-level features that cannot be changed. Eventually the branched model underwent a standard training on the broiler vocalizations which finished the training process.


**Figure 5**. Neural network model training in 3 stages. Blue color indicates frozen weights (not updated during training).

In both steps training was conducted for 50 epochs using batch size of 16 and Adam optimizer with learning rate of 0.001. The optimized loss function was a weighted sum of broiler vocalization classification loss (categorical crossentropy) and broiler age estimation loss (mean squared error), with the latter having the weights of 0.5 to emphasize its smaller importance. In both cases the controlled parameter was broiler vocalization classification balanced accuracy on a held-out validation set which dictated when the final model weights snapshot was taken.

*3.2.5 Evaluation procedure*

In order to evaluate the effectiveness of the proposed solution an extensive evaluation procedure has been conducted. First the broiler vocalization database was split into training, validation and testing subsets by selecting 25 random recordings from each class as testing set and another 25 random recordings from each class as validation set. Next, for each recording a ground truth normalized broiler age was assigned by taking the difference between this recording's beginning timestamp and first recording's beginning timestamp and dividing it by the experimental broiler lifespan (35.68 days).

The test set was used only once after the whole training pipeline concluded to measure model's effectiveness. The whole experiment was repeated ten times with random split between training, validation and testing and results on the test set from each repetition were aggregated and jointly summarized. As such, the final results are presented on a test dataset of 1250 recordings (5 classes, 25 recordings per class, 10 experiments).

*3.2.6 Performance of the detector*

The proposed method achieved a broiler vocalization classification balanced accuracy of 87.9%. Detailed information about the recognizer performance is presented in Table 3.

**Table 3.** Accuracy of the broiler vocalization classifier.

| Broiler sound category | Precision | Recall | F1-score |
|---|---|---|---|
| Distress calls | 91.7% | 97.2% | 94.4% |
| Pleasure notes | 93.6% | 88.0% | 90.7% |
| Short peeps | 78.7% | 84.4% | 81.5% |
| Warbles | 87.3% | 87.6% | 87.4% |
| Other sounds | 89.2% | 82.4% | 85.7% |
| Macro average | 88.1% | 87.9% | 87.9% |

Figure 6 shows the recognizer's confusion matrix.

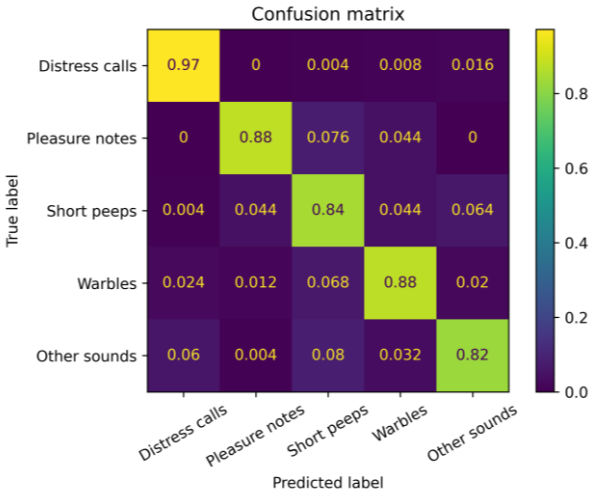


**Figure 6**. Confusion matrix of the broiler vocalization recognizer.

Considering the broiler age estimation, the recognizer's mean absolute error was 2.02 days (6% of broiler life span). The histogram of the broiler age estimation errors is depicted in Figure 7.

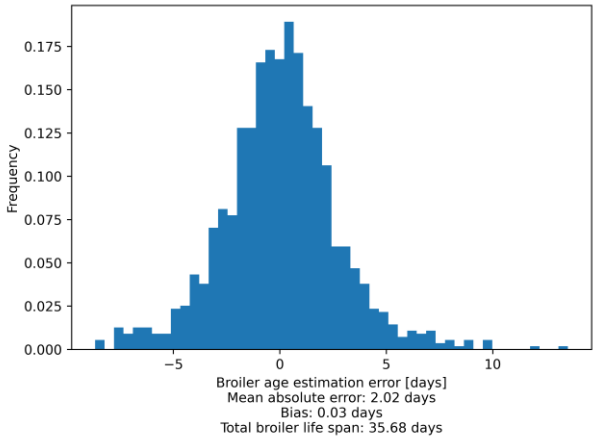


**Figure 7**. Histogram of broiler age estimation error. Values below 0 indicate that the estimated broiler age was lower than the actual age.

## 4. PREDEPLOYMENT IN A MEDIUM-SCALE HOUSING

To validate the use of the proposed model, a study in a medium scale housing at ILVO, Belgium was set up. Currently, a sequence of four trials is being developed. Per round, the study includes a total of 560 Ross 308 broiler chickens, randomly assigned to two treatments with two replications of 140 broilers each one (4 pens of size 9m x 4m). Treatment conditions comprised maintenance of standard temperature in one compartment, while the other compartment was subjected to heat stress of 32°C for six hours per day, between days 28 to 33 and days 35 to 40.
Two vocalization detectors are installed per pen, with microphone at approx. 1.5m height, to monitor variation in broiler vocalization between different compartments as a proxy for the broiler welfare.

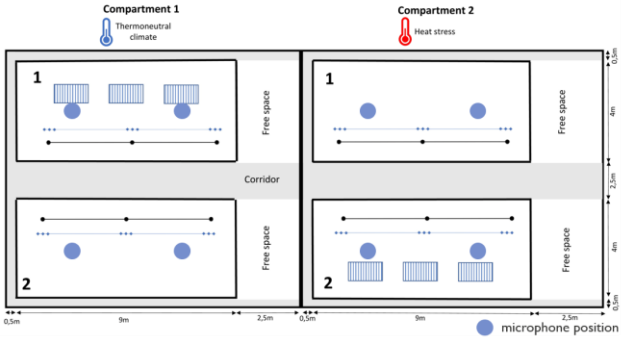


**Figure 8**. Schematic overview of the experimental setup. Two compartments (with and without heat stress) each contain 2 pens with 140 broilers each. The position of the acoustic vocalization detector (microphone) is indicated.

## 5. CONCLUSIONS

The broiler vocalization recognizer achieved a considerable accuracy on 4 classes of vocalizations with very limited confusion, which should allow for automatic long-term monitoring of poultry farms. However, assessment of its applicability in a real-life situation is still ongoing. Although the authors have taken measures to ensure sufficient generalization ability of the model, there is still a risk that it will fail to produce consistent results in all expected operational conditions. If this risk materializes it might be necessary to extend the dataset with recordings registered in other conditions or to simulate such conditions by generating synthetic samples. Additionally, the recognizer accuracy might be further improved by using more sophisticated network architectures and training schemes, but only if they won't increase the overall computational complexity which has to remain relatively low to keep the solution cost-effective.

## 6. ACKNOWLEDGEMENTS

The imec.icon project WISH is a research project bringing together academic researchers and industry partners. Project WISH is co-financed by imec and receives financial support from Flanders Innovation & Entrepreneurship (project nr. HBC.2021.0664) and/or Innoviris.

## 7. REFERENCES

[1] De Moura, D.J., Naeaes, I. De A., De Souza Alves, E.C., De Carvalho, T.M., Do Vale, M.M., De Lima, K.A.O. "Noise analysis to evaluate chick thermal comfort." *Sci. Agric. 65*, 438–443, 2008

[2] Herborn, K.A., McElligott, A.G., Mitchell, M.A., Sandilands, V., Bradshaw, B., Asher, L. "Spectral entropy of early-life distress calls as an iceberg indicator of chicken welfare." *J. R. Soc. Interface 17*, 2020

[3] Mao, A., Giraudet, C.S.E., Liu, K., De Almeida Nolasco, I., Xie, Z., Xie, Z., Gao, Y., Theobald, J., Bhatta, D., Stewart, R., McElligott, A.G., "Automated identification of chicken distress vocalizations using deep learning models", *J. R. Soc. Interface 19*, 2022

[4] Marx, G., Leppelt, J., Ellendorff, F. "Vocalisation in chicks (gallus gallus dom.) during stepwise social isolation." *Appl. Anim. Behav. Sci. 75*, 61–74, 2001

[5] Fontana, I., Tullo, E., Carpentier, L., Berckmans, D., Butterworth, A., Vranken, E., Norton, T., Berckmans, D., Guarino, M., "Sound analysis to model weight of broiler chickens", *Poultry Science 96*, issue 11, 2017

[6] Hou, Y., Kang, B., Van Hauwermeiren, W., Botteldooren, D. (2022). "Relation-guided acoustic scene classification aided with event embeddings." *in proc. of the international joint conference on neural networks (IJCNN)*, Padua, Italy.

[7] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M. D. "PANNs: large-scale pretrained audio neural networks for audio pattern recognition." *IEEE/ACM transactions on audio, speech, and language processing*, 28, 2880-2894, 2020.

[8] Gemmeke, J. F., Ellis, d. P., Freedman, D., Jansen, A., Lawrence, W., Moore, r. C., ... Ritter, M. "AudioSet: an ontology and human-labeled dataset for audio events." in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 776-780). 2017.

[9] Sainburg, T., Thielk, M., Gentner, T. Q. "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires." in F. E. Theunissen (ed.), *plos computational biology*, vol. 16, issue 10, 2020.

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. "Going deeper with convolutions (version 1)." arxiv. 2014 https://doi.org/10.48550/arxiv.1409.4842