# Joint Scattering for Automatic Chick Call Recognition

Changhong Wang
*Nantes Université, École Centrale Nantes*
*CNRS, LS2N, UMR 6004*
Nantes, France
changhong.wang@ls2n.fr

Emmanouil Benetos
*Centre for Digital Music*
*Queen Mary University of London*
London, UK
emmanouil.benetos@qmul.ac.uk

Shuge Wang, Elisabetta Versace
*School of Biological and Behavioural Sciences*
*Queen Mary University of London*
London, UK
{shuge.wang; e.versace}@qmul.ac.uk

*Abstract*—Animal vocalisations contain important information about health, emotional state, and behaviour, thus can be potentially used for animal welfare monitoring and behavioural neuroscience studies. Motivated by the spectro-temporal patterns of chick calls in the time–frequency domain, in this paper we propose an automatic system for chick call recognition using the joint time–frequency scattering (JTFS) transform. Taking full-length recordings as input, the system first extracts chick call candidates by an onset detector and silence removal. After computing their JTFS features, a support vector machine classifier groups each candidate into different chick call types. Evaluating on a dataset comprising 3013 chick calls collected in laboratory conditions, the proposed recognition system using the JTFS features improves the frame- and event-based macro F-measures by 10.2% and 11.7%, respectively, than that of a mel-frequency cepstral coefficients baseline.

*Index Terms*—Audio signal processing, bioacoustics, scattering transform.

## I. INTRODUCTION

Livestock farming is central for human sustainment. As farming technologies are booming, the large-scale and breeding-intensive poultry industries require systems to automatically monitor the welfare of animals. Livestock vocalisations play a crucial part in such systems, for example, assessing laying hens' thermal comfort [1], finding avian influenza-infected chickens to prevent the spread of diseases [2], and detecting abnormal sound of broilers [3] as an early warning tool. Yet, to the authors' knowledge, there is limited computational research on chick vocalisations and not any prior work on automatic chick call recognition.

Several decades ago, researchers [4] grouped chicks' vocalisations into contact (or distress) calls and pleasure calls, analysed their characteristics through displaying them on the spectrogram, and explored the common features of the sound signals that stimulate the production of each type of call. It was reported that contact calls are composed of descending frequencies only, are much louder, reach lower frequencies, and are given at a slower rate; while pleasure calls perform the opposite, i.e., are composed of ascending frequencies, are much softer, start from higher frequencies, and are produced at a faster rate. Variations of chick sound patterns under

Most of the work took place when the first author was at Queen Mary University of London.

successive changes in social isolation were analysed in [5]. Four types of chick calls were all labelled manually by visual inspection of the spectrogram: pleasure calls, contact calls, short peeps, and warbles.

As a pilot study towards fully automatic chick call recognition, we focus on two most frequent call types: pleasure call and contact call. Inspecting both types of calls in the time–frequency domain as shown in Fig. 1, we notice that pleasure calls (0-3 s) are characterised by upward frequency changes, low energy, and short duration while contact calls (3-6 s) exhibit the opposite, i.e., downward frequency changes, high energy, and long duration, which matches the findings in [4]. These spectro-temporal patterns are similar to the portamento musical playing technique explored in [6]. The latter technique also exhibits continuous frequency changes along with temporal modulations, which were shown in [6] that can be captured by the joint time–frequency scattering (JTFS). The JTFS is an instance of the scattering transform [7], a signal representation that is locally invariant to translations and stable to deformations. By adding a frequency scattering along the log-frequency axis, the JTFS exhibits extra invariance to frequency transpositions. These invariance properties make it a desirable representation for chick call recognition.
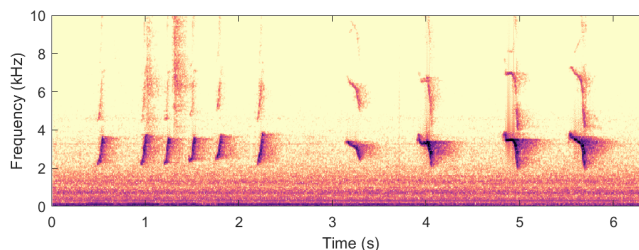


Fig. 1. Examples of pleasure (0-3 s) and contact (3-6 s) chick calls.

To the authors' knowledge, this is the first study to automatic recognition of chick calls using the joint time–frequency scattering transform. We propose a recognition system comprising two stages: (1) detection stage: we extract chick call candidates using an onset detector, followed by silence removal; (2) recognition stage: with the JTFS features extracted, a support vector machine classifier groups each candidate into different chick call types. The results show that the

JTFS features outperform a mel-frequency cepstral coefficients baseline, with the frame- and event-based macro F-measures improved 10.2% and 11.7%, respectively. We introduce each stage in Sections II and III, respectively. Section IV presents the evaluation dataset, metrics, and results, and Section V concludes the paper.

## II. CHICK CALL DETECTION

The number of calls emitted by the chicks is a good indicator of their state [5]. Automatic detection of chick calls can either be a preprocessing stage for chick call recognition or as a separate system specifically optimised for chick call counting. We investigate the former case by firstly detecting chick call onsets and then removing silence to extract chick call segments. We detect the onsets of chick calls using the *SuperFlux* [8] algorithm, which calculates the difference per frequency band in the magnitude spectrogram, applies a maximum filter along the frequency axis, sums up all positive changes over all bands, and selects the final onsets using peak-picking. The application of the maximum filter before summing up the positive changes reduces the number of false positives originated from frequency modulations without missing onsets.

Fig. 2 top displays the reference (dotted lines) and detected onsets (solid lines) of example chick calls. As can be seen, the intervals between two onsets include a large proportion of silence. We extract chick call segments from the inter-onset intervals by removing silence where the energy of the signal is below a certain threshold. A threshold of -32 dB is used according to experimenting on a grid of threshold values. Fig. 2 bottom shows the comparison of the reference call segments and the output segments of our detection system.
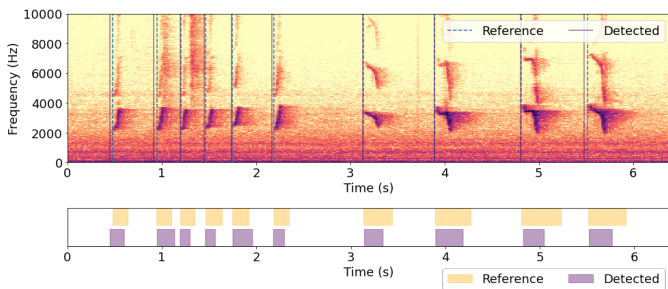


Fig. 2. Visualisation of onset detection and segmentation results for example chick calls. Top: spectrogram with reference onsets in dotted lines and detected onsets in solid lines; bottom: comparison of reference and detected call segments. For this example, the frame-based precision $\mathcal{P}$=96%, recall $\mathcal{R}$=67%, and F-measure $\mathcal{F}$=79%.

## III. CHICK CALL RECOGNITION

### A. Joint Time–Frequency Scattering

Similar to convolutional neural networks (CNNs), the scattering transform comprises operations of wavelet convolutions, modulus nonlinearities, and average pooling. Decomposing an audio waveform $\boldsymbol{x}(t)$ by a wavelet filterbank $\boldsymbol{\psi}_\lambda(t)$ and taking

complex modulus, we obtain the first-order wavelet modulus transform, which is also known as the scalogram $\mathbf{X}(t, \lambda)$:

$$\mathbf{X}(t, \lambda) = |\boldsymbol{x} * \boldsymbol{\psi}_\lambda|(t), \qquad (1)$$

where $t$ is the time variable and $\lambda$ is the log-frequency variable of $\boldsymbol{\psi}_\lambda(t)$. Averaging $\mathbf{X}(t, \lambda)$ by a lowpass filter $\boldsymbol{\phi}_T(t)$ with averaging scale $T$, the first-order scattering transform $\mathbf{S}_1(t, \lambda)$ is defined as

$$\mathbf{S}_1(t, \lambda) = \Big(|\boldsymbol{x} * \boldsymbol{\psi}_\lambda| * \boldsymbol{\phi}_T\Big)(t). \qquad (2)$$

Although $\mathbf{S}_1(t, \lambda)$ is invariant to time shifts, the averaging comes at the detriment of fast temporal modulations with time structures smaller than $T$. To recover these temporal modulations and to capture the variation along log-frequency axis, we decompose the scalogram with a joint time–frequency wavelet convolution, complex modulus, and averaging, according to [9]. Therefore, the obtained representation has all desirable properties: stability to time warps, and invariance to time shifts and frequency transpositions.

Convolving the scalogram $\mathbf{X}(t, \lambda)$ with a spectro-temporal wavelet filterbank $\boldsymbol{\Psi}_{v_\mathrm{t}, v_\mathrm{f}, \theta}(t, \lambda)$, taking complex modulus, and averaging by a two-dimensional (2-D) lowpass filter $\boldsymbol{\Phi}_{T,F}(t, \lambda)$, the joint time–frequency scattering (JTFS) transform [9] is defined as:

$$\mathbf{S}_2(t, \lambda, v_\mathrm{t}, v_\mathrm{f}, \theta) = \Big(\Big|\mathbf{X} \overset{t,\lambda}{*} \boldsymbol{\Psi}_{v_\mathrm{t}, v_\mathrm{f}, \theta}\Big| \overset{t,\lambda}{*} \boldsymbol{\Phi}_{T,F}\Big)(t, \lambda). \quad (3)$$

$T$ and $F$ are the temporal and spectral averaging scales of $\boldsymbol{\Phi}_{T,F}(t, \lambda)$, respectively. The symbol $\overset{t,\lambda}{*}$ denotes a 2-D convolution over the time variable $t$ and the log-frequency variable $\lambda$. The 2-D joint wavelet filterbank $\boldsymbol{\Psi}_{v_\mathrm{t}, v_\mathrm{f}, \theta}(t, \lambda)$ is derived from two 1-D wavelet filterbanks, the temporal filterbank $\boldsymbol{\psi}_{v_\mathrm{t}}(t)$ and the spectral filterbank $\boldsymbol{\psi}_{v_\mathrm{f}, \theta}(\lambda)$, by:

$$\boldsymbol{\Psi}_{v_\mathrm{t}, v_\mathrm{f}, \theta}(t, \lambda) = \boldsymbol{\psi}_{v_\mathrm{t}}(t)\boldsymbol{\psi}_{v_\mathrm{f}, \theta}(\lambda). \qquad (4)$$

$v_\mathrm{t}$ and $v_\mathrm{f}$ are the log-frequency variables of $\boldsymbol{\psi}_{v_\mathrm{t}}(t)$ and $\boldsymbol{\psi}_{v_\mathrm{f}, \theta}(\lambda)$, and measures the temporal and spectral variabilities, respectively. $\theta = \pm 1$ is the orientation variable that reflects the oscillation direction (up or down) of the spectro-temporal pattern. $\theta = -1$ flips the centre frequency of $\boldsymbol{\psi}_{v_\mathrm{f}, \theta}(\lambda)|_{\theta=1}$ from $2^{v_\mathrm{f}}$ to $-2^{v_\mathrm{f}}$.

Fig. 3 shows the log-frequency spectrograms and the frame-wise JTFS features of a pleasure call and a contact call. The directions of frequency changes are captured by the clear slopes, as shown by the bottom figures. For pleasure calls, the energy concentrates on the upward (left) side, while that of the contact call appears on the downward (right) side. Since chick calls with upward and downward frequency changes over time belong to different classes, i.e., pleasure calls and contact calls, we use the JTFS of both directions.

### B. Recognition System

We extract the JTFS features for chick calls by setting appropriate hyperparameters. An averaging scale $T = 2^{14}$ (in samples, corresponding to 372 ms) is used to roughly cover the average chick call duration. Filters per octave of
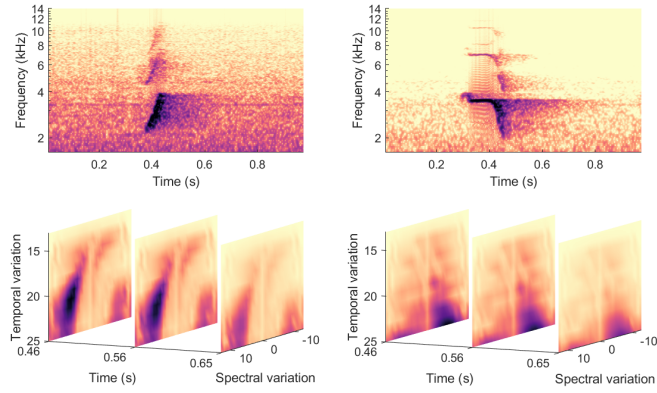
Fig. 3. The joint time–frequency scattering (JTFS) representation of example pleasure (left) and contact (right) calls. Top: log-frequency spectrogram; bottom: frame-wise JTFS features.

the temporal filters in the first-order wavelet decomposition $Q_1^{(\text{t})} = 16$ are used to offer a fine resolution scalogram. $Q_2^{(\text{t})} = 2$ and $Q_1^{(\text{f})} = 2$ are the filters per octave of the temporal and spectral wavelet filterbank used in the joint wavelet convolution due to the less oscillatory nature of the signal at these decompositions. $M = [0, 50]$ Hz is the extracted range of temporal modulation that contains the core information of chick calls. To account for temporal context, we calculate the mean and standard deviation of the JTFS features of 5 frames centred at the current frame. The frame size and dimension of the resulting JTFS features are 92 ms and 850, respectively. With the features calculated, we propose a chick call recognition system with two classification schemes:

- Scat-Only: in this scheme, a machine learning classifier takes the JTFS features of the whole recording as input and outputs frame-wise labels; neighbouring frames with the same label are then fused into chick call events. We postprocess the obtained chick call events by gap filling and minimum duration pruning. We fill the gaps between neighbouring events when the gaps are shorter than the shortest event in the training set; and prune the events that have smaller duration than the minimum duration event in the training set. The minimum duration is automatically calculated subject to the call type and the train-test split during recognition.

- Seg-Scat: this scheme builds upon the detection system introduced in Section II. A machine learning classifier takes the JTFS features of the extracted chick call segments as input, outputs frame wise labels, and assigns one label to each segment based on the majority vote of its frame labels.

Both classification schemes use support vector machines (SVMs) [10] with Gaussian kernels as classifiers due to their good generalisability based on a limited amount of training data [11]. In the recognition process, we split the dataset (see Section IV-A) into training and test sets by leaving one chick subject out. Within each split, we run a 3-fold cross-validation,

sampling the training set in a way that ensures approximately the same ratio of positive and negative class instances for a given chick call type in each fold. The SVM hyperparameters to be optimized are the error penalty parameter $C$ and the width of the Gaussian kernel $\gamma$. We use consistent parameter grids of $10^{\{0:1:2\}}$ and $10^{\{-4:1:-2\}}$ for $C$ and $\gamma$, respectively, during training and select the best ones for testing. For the Scat-Only scheme, we calculate the JTFS features of full-length recordings, while for the Scat-Only method, the JTFS features are calculated for the detected chick call candidates. All features are z-score normalised with the training-set statistics. Our source code[1] is based on the ScatNet[2] toolbox and is publicly available for reproducibility.

## IV. Evaluation

### A. Dataset

As a pilot study[3], we collected data in laboratory conditions, using domestic chicks from the Ross 308 strain of the species *Gallus gallus* as an animal model. We placed one chick at a time in an arena within 12 hours after their hatching. For each chick, we recorded their movements and sounds for around 10 minutes by a Microsoft LifeCam Studio Webcam and an AKG P170 microphone with a Behringer U-Phoria UMC204HD audio interface. The camera was placed approximately 1 m above the centre of the arena and the microphone 1 m above the outer wall of the arena. All data was recorded at a sampling rate of 44.1kHz/16bits. We collected one recording for one chick at a time.

As a proof-of-concept study in this paper, we evaluate the proposed system on a dataset comprising recordings of 4 chicks. This is due to the available annotations we have currently: start time, end time, and call type annotations for full-length recordings of 4 chicks. All the annotations were created by an expert in chick behaviour from the Prepared Minds Lab[4] at Queen Mary University of London using Sonic Visualiser [12]. The expert was previously trained to a 91.1% agreement level with another expert in the same task on an example chick. For ambiguous labelling cases, decisions were made following discussions between three experts. Three types of chick calls were annotated: pleasure, contact, and uncertain calls. Uncertain calls are those calls that all the three experts were not certain about. The total duration of the dataset is 44 minutes. Table I lists the number of each type of calls in the dataset, where the classes are highly imbalanced. As described in Section III-B, we conduct a subject-independent evaluation of our recognition system, with no overlap of chick subject in the training and test sets.

### B. Baseline and Metrics

To the authors' knowledge, there is not yet any prior work on fully automatic recognition of chick calls. We compare

TABLE I
NUMBER OF PLEASURE, CONTACT, AND UNCERTAIN CALLS
PRODUCED BY EACH CHICK IN THE CHICK CALL DATASET.

| Chick | Pleasure | Contact | Uncertain | Total |
|---|---|---|---|---|
| 1 | 119 | 315 | 9 | 443 |
| 2 | 617 | 492 | 146 | 1255 |
| 3 | 47 | 682 | 60 | 789 |
| 4 | 36 | 455 | 35 | 526 |
| Total | 819 | 1944 | 250 | 3013 |

TABLE II
ONSET DETECTION AND SEGMENTATION RESULTS FOR EACH
CHICK IN THE CHICK CALL DATASET ($\mathcal{P}$, $\mathcal{R}$, AND $\mathcal{F}$ ARE IN %;
T/A=TOTAL OR AVERAGE).

| Chick | #calls | Onset detection (*mir_eval*) | | | Segmentation (frame-based) | | | Segmentation (event-based) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ |
| 1 | 443 | 74 | 93 | 83 | 56 | 80 | 66 | 62 | 77 | 69 |
| 2 | 1255 | 95 | 98 | 97 | 75 | 70 | 72 | 68 | 67 | 67 |
| 3 | 789 | 88 | 99 | 93 | 87 | 82 | 84 | 78 | 86 | 81 |
| 4 | 526 | 77 | 97 | 86 | 71 | 81 | 75 | 65 | 81 | 72 |
| T/A | 3013 | 83.5 | 96.8 | 89.8 | 72.3 | 78.3 | 74.3 | 68.3 | 77.8 | 72.3 |

the proposed recognition system with that taking the mel-frequency cepstral coefficients (MFCCs) as input, which were commonly used for animal vocalisation analysis [2], [13], [14]. Similarly to the main experiments (see Section III-B), we conduct two baseline classification schemes, i.e., MFCC-Only and Seg-MFCC. The former calculates the MFCCs of full recordings, outputs frame labels of chick call type, and fuses neighbouring frames with the same labels into chick call events. The latter first extracts chick call candidates using the detection method in Section II and then computes the MFCCs for each segment. The frame size and dimension of the MFCC features are 25 ms and 24, respectively. When comparing the results from the MFCCs to that from the JTFS, we resample the former to the same frame size of the latter (92 ms). We conduct both frame- and event-based evaluation of the detection and the recognition results. The former compares the output chick call labels with the ground truth in a frame-wise manner, while the latter compares the predicted chick call events with the ground truth by their onset and duration. An event is considered to be correctly detected when its onset falls in a tolerance window of $\pm 100$ ms around that of the reference call and its duration is more 50% of the reference call duration. We use precision $\mathcal{P}$, recall $\mathcal{R}$, and F-measure $\mathcal{F}$ as the metrics for each evaluation method [15].

### C. Results

To investigate the influence of the detection stage on the final recognition stage, we first present the evaluation of both onset detection and segmentation. We evaluate the onset detection result using the *mir_eval* library [16], which calculates a maximum match of the reference and the detected onset times subject to a window constraint of $\pm 150$ ms. The segmentation result is evaluated using both frame- and event-based evaluation methods introduced in Section IV-B. Table II shows the detection and segmentation results for each chick and those on the whole dataset. We achieve average F-measure of 89.8% for the onset detection stage; for the segmentation result, average F-measures of 74.3% for frame-based evaluation, and 72.3% for event-based evaluation are obtained.

Table III displays the recognition results on the recordings of the four chicks. We compare these results from four fronts: the JTFS features versus the MFCCs; the two classification schemes using the JTFS; the recognition results using detected versus annotated chick call segments; and the results of

different call types. Comparing the Scat-Only with the MFCC-Only method, we observe that they achieve comparable results in the frame-based evaluation, both with macro F-measures of 32.1%; while in the event-based evaluation, the former underperforms the latter, with macro F-measure of 12.7% against 23.8%. For the comparison between the JTFS and the MFCCs using detected segments, the Seg-Scat outperforms Seg-MFCC, with macro F-measures improved 10.2% in the frame-based evaluation and 11.7% in the event-based evaluation, respectively. Narrowing the scope with the recognition results using the JTFS, i.e., Scat-Only and Seg-Scat, the former underperforms the latter in both frame- and event-based evaluation. The Seg-Scat increases macro F-measure by 6.2% as compared to that of the Scat-Only in the frame-based evaluation and by 26.2% in the event-based evaluation. It is expected that the Scat-Only yields much lower event-based macro F-measures in contrast to the Seg-Scat, where the latter uses chick call segments, either detected or annotated.

As shown in Table III, both 'Seg-' methods, i.e., Seg-Scat and Seg-MFCC, exhibit much better performance using the annotated chick call segments as compared to using the detected chick call segments. In the frame-based evaluation, macro F-measures increase 20.9% and 10.5% for the two methods, respectively; the corresponding improvement in the event-based evaluation are 19.7% and 10.1%, respectively. This verifies the potential of the JTFS for chick call recognition. Inspecting the F-measures of each type of calls, we notice that all four methods exhibit better performance on contact call classification than that on pleasure and uncertain call classification. This may be attributed to the small amount of samples for the latter two types of calls or to their characteristics (e.g., lower amplitude and shorter duration). Yet, the JTFS is more robust to data imbalance as compare to the MFCCs. For example, the Seg-Scat achieves frame-based F-measure of 60.3% against 16.5% from the Seg-MFCC for recognising pleasure calls using annotated segments, although both methods have comparable results on contact call classification.

### V. CONCLUSION

In this paper, we have proposed a fully automatic system for chick call recognition using the joint time–frequency scattering

TABLE III
FRAME- AND EVENT-BASED RECOGNITION RESULTS IN TERMS OF F-MEASURE (%) USING FULL-LENGTH RECORDINGS, ANNOTATED CHICK CALL SEGMENTS, AND DETECTED CHICK CALL SEGMENTS AS INPUTS, RESPECTIVELY, FOR THE PROPOSED AND THE BASELINE METHODS: (1) SCAT-ONLY: RECOGNISE CHICK CALLS DIRECTLY FROM FULL-LENGTH RECORDINGS USING THE JTFS ONLY; (2) SEG-SCAT: SEGMENT AUDIO INTO CHICK CALL CANDIDATES AND CLASSIFY EACH SEGMENT USING THE JTFS; (3) MFCC-ONLY: RECOGNISE CHICK CALLS DIRECTLY FROM FULL-LENGTH RECORDINGS USING THE MFCCS ONLY; (4) SEG-MFCCS: SEGMENT AUDIO INTO CHICK CALL CANDIDATES AND CLASSIFY EACH SEGMENT USING THE MFCCS. 'MARCO' MEANS THE MACRO F-MEASURE OF THE THREE TYPES OF CALLS.

| Recognition system input | Classification scheme | Frame-based | | | | Event-based | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Pleasure | Contact | Uncertain | Macro | Pleasure | Contact | Uncertain | Macro |
| Full-length recordings | Scat-Only | 10.5 | 79.5 | 6.3 | 32.1 | 7.0 | 29.8 | 1.4 | 12.7 |
| | MFCC-Only | 4.8 | 83.8 | 7.8 | 32.1 | 2.5 | 61.0 | 8.0 | 23.8 |
| Annotated segments | Seg-Scat | 60.3 | 95.3 | 22.0 | **59.2** | 63.0 | 91.5 | 21.3 | **58.6** |
| | Seg-MFCC | 16.5 | 91.0 | 8.3 | 38.6 | 18.8 | 87.3 | 5.8 | 37.3 |
| Detected segments | Seg-Scat | 17.8 | 82.0 | 15.0 | **38.3** | 19.5 | 83.1 | 14.0 | **38.9** |
| | Seg-MFCC | 4.5 | 77.0 | 2.8 | 28.1 | 3.6 | 75.0 | 2.9 | 27.2 |

(JTFS). The system first extracts chick call candidates from full-length recordings by onset detection and silence removal; and then classifies each candidate based on their JTFS features. Compared with a mel-frequency cepstral coefficients baseline, the proposed system using the JTFS of detected chick call segments improves macro F-measures by 10.2% and 11.7% in the frame- and event-based evaluation and is more robust to data imbalance. Comparison of the recognition performance using detected and annotated chick call segments verifies the potential of the JTFS for chick call recognition.

Three limitations exist in the current study. We collected data in a laboratory which is much cleaner than that recorded in real living conditions of chicks. For the latter case, we could process the recordings with a source separation technique [17] to separate the sound of multiple chicks or a denoising method to remove noise, prior to the application of the JTFS features. The other two limitations include the small amount of data, in terms of both duration and number of chicks, and the highly imbalanced number of examples for each class. Lack of data is the main reason of not having applied deep learning techniques in this paper. For future work, we may use semi-supervised or unsupervised learning for the task due to the availability of unlabelled data. Unsupervised learning may also potentially discover new chick call patterns. We can also compare the proposed system to existing vocalisation detection systems for other animals, for example, the broiler stress detection system in [14]. Once fully developed, the proposed recognition system could be used for other research in computational bioacoustics, such as acoustic playback studies on reciprocal chick calls, neurophysiological studies to examine the electrophysical recordings in chicks exposed to specific calls, and analysis of other bioacoustic signals exhibiting similar spectro-temporal patterns.

## REFERENCES

[1] X. Du, L. Carpentier, G. Teng, M. Liu, C. Wang, and T. Norton, "Assessment of laying hens' thermal comfort using sound technology," *Sensors*, vol. 20, no. 2, p. 473, 2020.

[2] K. Cuan, T. Zhang, J. Huang, C. Fang, and Y. Guan, "Detection of avian influenza-infected chickens based on a chicken sound convolutional neural network," *Computers and Electronics in Agriculture*, vol. 178, p. 105688, 2020.

[3] L. Liu, B. Li, R. Zhao, W. Yao, M. Shen, and J. Yang, "A novel method for broiler abnormal sound detection using WMFCC and HMM," *Journal of Sensors*, vol. 2020, 2020.

[4] N. Collias and M. Joos, "The spectrographic analysis of sound signals of the domestic fowl," *Behaviour*, vol. 5, no. 1, pp. 175–188, 1953.

[5] G. Marx, J. Leppelt, and F. Ellendorff, "Vocalisation in chicks (*Gallus gallus* dom.) during stepwise social isolation," *Applied Animal Behaviour Science*, vol. 75, no. 1, pp. 61–74, 2001.

[6] C. Wang, V. Lostanlen, E. Benetos, and E. Chew, "Playing technique recognition by joint time-frequency scattering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[7] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[8] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *International Conference on Digital Audio Effects (DAFx)*, vol. 7, 2013.

[9] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, 2019.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[11] F. Albu and D. Martinez, "The application of support vector machines with Gaussian kernels for overcoming co-channel interference," in *Proceedings of the IEEE Signal Processing Society Workshop*, 1999, pp. 49–57.

[12] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1467–1468.

[13] J. C. Bishop, G. Falzon, M. Trotter, P. Kwan, and P. D. Meek, "Livestock vocalisation classification in farm soundscapes," *Computers and Electronics in Agriculture*, vol. 162, pp. 531–542, 2019.

[14] N. Jakovljević, N. Maljković, D. Mišković, P. Knežević, and V. Delić, "A broiler stress detection system based on audio signal processing," in *IEEE Telecommunications Forum (TELFOR)*, 2019, pp. 1–4.

[15] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[16] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[17] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.