# An Open-Access System for Long-Range Chainsaw Sound Detection

Nikolaos Stefanakis* ,†, Konstantinos Psaroulakis*, Nikonas Simou*, Christos Astaras‡
* FORTH-ICS, Heraklion, Crete, Greece, GR-70013
† Hellenic Mediterranean University, Department of Music Technology
and Acoustics, Rethymno, Greece, GR-74100
‡ Forest Research Institute, ELGO-DIMITRA, Vasilika, Thessaloniki GR-57006

*Abstract*—**A pipeline for automatic detection of chainsaw events in audio recordings is presented as the means to detect illegal logging activity in a protected natural environment. We propose a two-step process that consists of an activity detector at the front end and a deep neural network (DNN) classifier at the back end. At the front end, we use the Summation or Residual Harmonics method in order to detect patterns with harmonic structure in the audio recording. Active audio segments are consequently fed to the classifier that decides upon the absence or presence of a chainsaw event. As acoustic feature, we propose the widely-used amplitude spectrogram, passing it through the recently proposed Per-Channel Energy Normalization (PCEN) process. Results based on real-field recordings illustrate that the proposed end-to-end system may efficiently detect low-SNR chainsaw events at a very low false detection rate.**

## I. Introduction

Human activity is considered today as the primary reason for habitat loss for a large number of Earth's plant and animal species. It contributes to the permanent loss of species but also to the weakening of the ecosystems that are of significant importance for the overall health of the planet and as a consequence, to the quality of the human life. Estimates suggest the Earth has lost about half of its forests in 8000 years of human activity, with much of this occurring in recent decades.

Acoustic sensors in natural environments may be of great use towards their protection. One common application is the use of acoustic sensors for monitoring sounds produced by the various species while in their natural environment [1], [2]. The other use is detecting illegal human activity inside protected areas such as trespassing, illegal hunting, logging, grazing etc [3]–[5]. These tasks become possible with the use of low-cost, power-efficient electronic devices known as Autonomous Recording Units (ARUs). ARUs are capable of recording continuously for several days or weeks and considering that several recorders can be used simultaneously, huge amounts of audio data can be gathered in relatively short periods of time. As a consequence, it is infeasible for human experts to hear or visually inspect the entire collection of recordings. Thus, automatic or semi-automatic processing of the sound files is necessary for analyzing the information in a timely manner.

A limited number of different approaches have been proposed for recognizing chainsaw activity from audio recordings. Most of them were developed based on conventional classification techniques [6]–[11] and were evaluated based on small-scale datasets, with limited consideration on the effect of background noise in the recording.

An issue of primary importance in classifying sound sources located far from the acoustic sensor is robustness to noise and also, robustness to variations in the level of the acoustic pattern of interest [12]. In this paper, this task is investigated from the perspective of DNNs. Robustness to noise with DNNs has been studied from several researchers from the perspective of feature enhancement [13]–[15] and also, from the perspective of feature normalization [2], [12], [16]. Also, several works have demonstrated the importance of data augmentation [15], [17]–[19] in helping the network to generalize to variations of the acoustic patterns and more generally, in avoiding overfitting.

In this paper, we present an open-access end-to-end system [20] for the detection of chainsaw events in real-field audio recordings. The system consists of a Voice Activity Detector (VAD), at the front end, and a binary DNN classifier, at the back end. As input to the classifier we present a spectrogram domain acoustic feature that is passed through the recently proposed Per-Channel Energy Normalization (PCEN) process [2], [16]. Our results demonstrate the effectiveness of PCEN for the intended task and the significant advantage in classification performance that can be achieved using data augmentation. Overall, the proposed pipeline may efficiently detect low-SNR chainsaw events at a very low false-positive rate, which is an important requirement when considering the huge amounts of audio data involved in such applications.

## II. Methodology

### A. Voice Activity Detection

Due to the fact that chainsaw sound has a harmonic structure similar to voiced speech, we exploit the Summation of Residual Harmonics (SRH) method [21] for Voice Activity Detection (VAD), which is well-known for it's ability to provide reliable voicing decisions in noisy conditions. Note

This work was funded via a Single RTDI State Aid Action ”Research – Create – Innovate” grant (T1EDK-04488), which is co-financed by Greece and the European Union (European Regional Development Fund) as part of the Operational Program ”Competitiveness, Entrepreneurship and Innovation” of the National Strategic Reference Framework (NSRF) 2014-2020.

that while this VAD approach was initially designed for speech signals, it is used for the first time to our knowledge in the context of environmental sound classification in this paper. The outcome of the SRH approach that we utilize is a time-varying metric of the voicing activity that, for the needs of this paper, is called Voicing Strength (VS). The metric becomes available as a time-series in the form $v(\tau)$, where $\tau$ is the time-frame index. VS is calculated using the relevant function provided in the COVAREP toolbox [22]. The reader is referred to the original publication [21] for more details regarding the VAD approach.

To demonstrate the robustness of the proposed VAD in chainsaw detection, we plot in Fig. 1(a) the amplitude spectrogram of a portion of a real field recording containing chainsaw activity. In Fig. 1(b) one can see the variation of VS as a function of time while subfigure (c) illustrates the variation of the energy of a high-pass-filtered version of the signal as a function of time. The harmonic structure of the chainsaw sound is more or less visible in (a) and the chainsaw activity is clearly correlated with a rise of $v(\tau)$ in (b). On the other hand, the variations of the energy follow less clearly the chainsaw activity, especially in the time interval between 2 and 5 s where the chainsaw sound is weaker. This justifies the use of VS as a criterion for chainsaw activity, at the same time showing that energy criteria that are conventionally utilized for audio segmentation are not trustworthy at so low SNR conditions. Finally, it is worth to note from Fig. 1(a) that chainsaw harmonics are visible only up to 1kHz. Working with real-field recordings we have observed that this is a very regular phenomenon. It can be explained by the fact that high-frequency chainsaw harmonics are weaker than the low-frequency ones and are thus often masked by background noise. Moreover, air absorption [23] causes sound waves to attenuate in quadratic proportion to their frequencies, which makes high-frequency components even less detectable when propagating at long ranges. This observation is exploited in order to pose a limit to the highest frequency of analysis in the construction of the activity detector and also in the choice of the acoustic features extracted for classification.

### *B. VAD-based Segment Selection*

Using VS as a metric to detect portions with harmonic structure in a long (e.g. 24 hr) audio recording is an essential part of the workflow followed for chainsaw detection in this paper. Our VAD-based segment selection algorithm (VAD-SS) is utilized as the front-end of our chainsaw detection system, followed by a binary DNN classifier at the back-end. Moreover, the presented VAD-SS process is exploited for automated spotting of interesting sound events from the original recordings and has thus been used for creating the dataset that is required for training the back-end. Key to automatic segmentation is the observation that chainsaw activity is expected to produce high VS values not at single isolated time-frames, but along multiple continuous time-frames. Following this skeptic, we form the requirement that there are at least $N_{ca}$ consecutive time-frames with VS values greater than a threshold $T_{srh}$ in order for an audio portion to be selected for further processing. In more detail, the following process is followed for extracting audio segments from a long continuous recording;

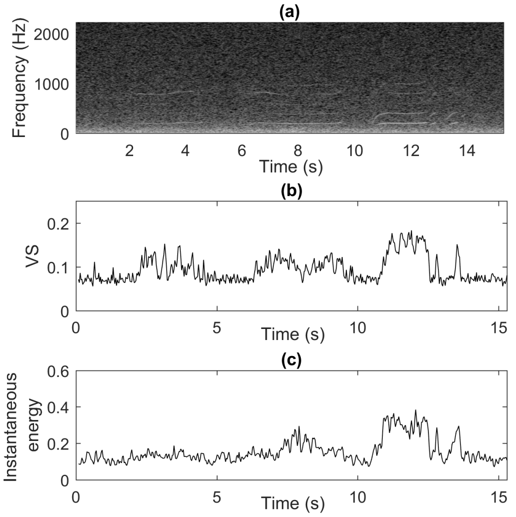

Fig. 1. The amplitude spectrogram of an audio segment with chainsaw activity is shown in logarithmic scale in (a). The corresponding VS values as a function of time are shown in (b) and the instantaneous energy as a function of time in (c).

1) Calculate VS in terms of $v(\tau)$ across the entire duration of the recording
2) Construct the collection of all the active time-frame indexes, where an active time-frame is defined as any time-frame where condition $v(\tau) \geq T_{srh}$ holds. Consecutive active time-frames are merged together to form a single so-called utterance, starting at time $\tau_u^{start}$ and ending at time $\tau_u^{end}$, where $u$ is the utterance index.
3) Keep only the utterances which are formed by more than $N_{ca}$ consecutive time-frames, i.e., $\tau_u^{end} - \tau_u^{start} > N_{ca}$ holds.
4) If there are any utterances with lengths smaller than an integer number $N_d$, extend their lengths to become equal to $N_d$. Perform this operation by simply reducing $\tau_u^{start}$ and increasing $\tau_u^{end}$ an equal number of time-frames.
5) Finally, if there is overlap between two consecutive utterances, merge them in order to form a single longer utterance.

Following this operation, several utterances of different lengths are produced from each audio recording and the length of each utterance is at least $N_d$ time-frames. The reason for using a different value for $N_d$ and $N_{ca}$ is that a chainsaw revving may last several multiples of $N_{ca}$, but due to the high level of noise or due to the non-stationary nature of the particular chainsaw instance, only a small portion of the event passes the $v(\tau) \geq T_{srh}$ criterion. Additionally, the tactic to extend the acoustic representation along time makes it easier for the listeners to annotate the data, possibly by allowing them to better perceive the transitions at the beginning or end of the underlying event.

For the needs of data collection VAD-SS was performed with $T_{srh}$ in the range $[0.8, 0.96]$, $N_{ca} = 3$ and $N_d = 23$. Additional parameters that should be reported concern those used to invoke the "pitch_srh" function in COVAREP toolbox, that was necessary for calculation of VS; The number of harmonics was set to 4 and the minimum and maximum $F0$ values were set to 40 and 760 Hz respectively. Finally, the frame-length and hop-size for calculation of the SRH were 180 and 60 ms respectively, which means that the length of an extracted utterance is ensured to be at least 1.5 s.

### C. Acoustic Features

Several spectro-temporal representations were tested as features for chainsaw detection with DNNs. An observation common to all cases was that frequencies above 2 kHz do not add significant gain to the classification performance and more over, lead to an unnecessary increment of the feature dimension. This observation is also discouraging for using mel-spectrogram as an acoustic feature since the mel scale is approximately linear up to the considered frequency limit [24].

In this paper, we present results for two different acoustic features which are based on the spectrogram of the raw audio signal. Transformation from the time domain to the time-frequency (TF) domain is performed with a window size of 720 samples (90 ms) and hop length of 120 samples (30 ms), while the size of FFT used is equal to 1024. Along the frequency dimension, we take into account only frequency bins with indexes from 9 to 215, corresponding to frequencies from 63 to 1680 Hz. From this basic TF representation, the first acoustic feature, termed Amplitude Spectrum (AS) from now on, is derived by applying maximum pooling - keeping only one value every 2 bins - and then normalizing the feature with respect to the sum of all it's elements. The second acoustic feature, termed PCEN spectrogram from now on, is derived by applying the PCEN method [12] to the previously described TF representation and then again using maximum pooling to keep only the frequency bin with the maximum magnitude every two bins. PCEN was applied on a per-utterance level, using the parameter values that were recommended for outdoor applications in [25] and more particularly, $\epsilon = 10^{-6}$, $a = 0.8$, $\delta = 10$ and $r = 0.25$. Moreover, for deriving the smooth version of the TF energy at the denominator of the expression (see e.g. Eq.(1) in [12]) an auto-regressive process with a factor $a = 0.15$ was used.

### D. DNN Classifier and Presentation of Classification Results

Various DNN architectures were tested, among which the most promising appeared to be the one based on Long Short Term Memory (LSTM) units. The proposed model consists of a 512 unit LSTM layer followed by a second LSTM layer with 256 units. The output of the second LSTM layer is fed to a fully connected layer of 128 units using a relu activation function, followed by a dropout layer with 0.3 probability and an output softmax layer. In order to optimize training, we apply a learning rate decay of $0.4$ and tolerance of $5$ epochs, using early stopping with a tolerance of $10$ epochs to avoid overfitting. Both are implemented according to the validation loss which is calculated every time over a random $5\%$ split of the available training data. The implementation was made in Python using the keras toolbox.

Our DNN operates on a fixed feature size and more particularly on a size of 46 (time-frames) × 103 (frequency bins), which corresponds to an audio segment of 1.44 s, slightly shorter than the minimum utterance duration produced by the VAD-SS. While the classifier outputs one decision per-segment, decisions for multiple successive segments can be merged together, when assigned to the positive class, in order to mark a longer chainsaw time interval. The marked chainsaw segments are automatically extracted as *.wav* files of variable duration and with a naming convention that is indicative of the temporal location inside the recording. Due to the very low false-detection-rate achieved by the end-to-end system (shown in the last Section), this makes it easy for the user to verify the presence or absence of chainsaw activity in very long audio recordings with relatively little effort.

## III. Dataset

The recordings were obtained from 13 SWIFT ARUs developed by Cornell University's Lab of Ornithology, which use an omnidirectional analog microphone (PUI Audio Inc., Part Nr: POW-1644L-B-LW100-R). The recordings were obtained in PCM format using 8kHz sampling rate and 16 bits resolution. The ARUs were distributed in different protected areas in Greece (see Table I).

For generating a basic dataset required for training, recordings from each ARU were randomly selected and subjected to the VAD-SS process. The utterances extracted by the segmentation step were then manually labeled into various categories including that of "chainsaw". It is worth noting that, apart from the chainsaw sounds, several instances were triggered by insects, grazing animals such as goats and sheep and also by dogs and birds. Additional instances were triggered by aeroplanes and also by cars and trucks that happened to pass within the acoustic range of each ARU. Finally, there were several cases of human voice. Since this paper considers a binary classification problem, chainsaw sounds represent the Positive class while all other types of sounds populate the Negative class. The number of utterances detected for the Positive (Npos) and Negative (Nneg) class can be seen for each ARU in Table I. It is noted that the total duration of chainsaw events in the basic dataset was 1.95 h while that of all other sounds equal to 8.39 h.

Apart from the basic dataset, an augmented dataset is also considered in this paper. The augmented dataset was constructed by enriching the basic dataset using the following approaches;

**Resampling**: rather than using pitch shifting and time stretching, well known in the context of data augmentation for DNN classification [2], [18], we use Matlab's built-in function *resample* in order to downsample and upsample each audio

| Index | Name | Npos | Nneg | Location |
|---|---|---|---|---|
| 1 | SW2 | 880 | 1253 | Evagelistria |
| 2 | RP6 | 800 | 1130 | Rhodope |
| 3 | RP11 | 636 | 804 | Rhodope |
| 4 | RP14 | 205 | 348 | Rhodope |
| 5 | PR | 155 | 1204 | Prespes |
| 6 | RP10 | 129 | 789 | Rhodope |
| 7 | RP15 | 113 | 457 | Rhodope |
| 8 | RP3 | 76 | 4047 | Rhodope |
| 9 | RP1 | 0 | 1026 | Rhodope |
| 10 | RP2 | 0 | 128 | Rhodope |
| 11 | RP4 | 0 | 1251 | Rhodope |
| 12 | SW1 | 0 | 1722 | Maroneia |
| 13 | EVR | 0 | 1005 | Evros |
| Total | | 2994 | 15164 | |

TABLE I
NUMBER OF DETECTED UTTERANCES PER ACOUSTIC PATTERN AND ARU.

segment by 6%. Note that the upsampling and downsampling operations alter both the duration and the pitch of the input audio segment.

**Mixing with external chainsaw sounds**: the sound produced by a chainsaw unit during operation differs depending on the brand name or the model. To help the classifier gain robustness to such variability, we searched for additional chainsaw recordings in open-access datasets. In particular, chainsaw recordings were downloaded from freesound (https://freesound.org/) and from the dataset provided in [26]. Moreover, some additional audio samples were captured by the first author using a smartphone device. Most of these recordings were captured near the chainsaw unit and can thus be characterized as ”clean” chainsaw recordings. To make this data exploitable for the task under investigation, the clean chainsaw sounds were converted to an appropriate audio format (PCM, 8kHz, 16bit) and were then convolved with an FIR filter whose frequency response was a descending function of the frequency, providing an attenuation of 0 dB at 0 Hz and -12 dB at 4kHz. This was done in order to simulate losses due to air absorption [23]. The resulting audio files were mixed with portions of background noise - randomly selected from all 13 ARUs - using two different mixing weights, one corresponding to an SNR of 0 dB and another to an SNR of -10 dB. This augmentation approach enriched the dataset with 2.68 hours of additional chainsaw events. Moreover, the background noise segments participating in the mixture were also presented for training as representatives of the Negative class.

## IV. EVALUATION

A ”leave-one-out” approach is considered for validation, in which case all the utterances extracted from one ARU are put into test, while the data associated with the rest ARUs is considered to be available for training. As stated in [2], this validation approach will better reflect the system's ability to adapt to variations of background noise in time (e.g. dawn vs. dusk) and space (i.e., different sensor location), as well as to variations in the characteristics of different ARUs (e.g. frequency response). The test was repeated for ARUs 1 to 8 while ARUs 9 to 13 - that did not contain any chainsaw events - were used only for training.The results are averaged across all 8 ARUs. Doing so, the performance metrics are neutralized with respect to the number of Positive or Negative samples associated to each ARU.

| | AS | | | PCEN | | |
|---|---|---|---|---|---|---|
| Specificity | 95% | 90% | AUC | 95% | 90% | AUC |
| Basic | 54.3 | 66.7 | 0.837 | 58.2 | 69.6 | 0.890 |
| Augmented | 73.3 | 82.3 | 0.914 | 77.7 | 85.3 | 0.942 |

TABLE II
SENSITIVITY (IN %) AND AUC ACHIEVED FOR EACH METHOD USING THE BASIC AND THE AUGMENTED DATASET. SENSITIVITY RESULTS ARE SHOWN FOR VALUES OF THE PROBABILITY THRESHOLD THAT RESULT TO 95% AND 90% SPECIFICITY.

The first part of the evaluation involves the DNN classification performance. The decisions are here taken per-segment based on a probability threshold $p_{thr}$, so that a sample is assigned to the Positive (resp. Negative) class whenever $p \geq p_{thr}$ (resp. $p < p_{thr}$) holds, where $p$ is the probability of the Positive class, provided by the softmax layer of the DNN. For the needs of evaluation, for each method, we fix $p_{thr}$ to the value that results to an average specificity equal to 95% and 90%. The average sensitivity and the Area Under the Curve (AUC) for each method, with and without data augmentation, can be seen in Table II. The results indicate that data augmentation improves average sensitivity around 19% for both AS and PCEN (for 95% specificity). Also, as expected, PCEN achieves a better average sensitivity compared to AS in all cases.

Additional results are also presented for VAD-SS alone and for the end-to-end system that consist of both the VAD-SS and the DNN classifier. To calculate the True Detection Rate (TDR), we used Praat in order to manually annotate chainsaw events in randomly selected audio portions (chosen from all 8 ARUs), marking the beginning and end of each chainsaw event in the recording. TDR is obtained here as the ratio of the duration of the audio content detected by the algorithms to that detected by the human expert. In an additional experiment, False Detection Rate (FDR) is calculated by dividing the duration of non-chainsaw events detected by the algorithm, to the total duration of the audio content. Calculation of FDR is based on eight 12h recordings - one for each ARU put into test - that did not contain any chainsaw events and that did not participate in any way in the extraction of the data used for training. The parameters used for VAD were $T_{srh} = 0.78$, $N_{ca} = 3$ and $N_d = 23$ while for the classifiers, the value of $p_{thr}$ was equal to 0.7 and 0.5 for SA and PCEN respectively.

The TDR and FDR results are shown in Table III. It can be seen that, in average, VAD-SS lets 86.6% of the actual chainsaw activity to pass through. Following the DNN classification step, TDR drops to 73.2% and 73.3% when using AS and PCEN respectively. On the other hand, in terms of average FDR, VAD-SS extracts 313.9 seconds per hour

|  | VAD-SS | VAD-SS+AS | VAD-SS+PCEN |
|---|---|---|---|
| TDR (%) | 86.6 | 73.2 | 73.5 |
| FDR (s/hour) | 313.9 | 11.4 | 9.5 |

TABLE III
TDR AND FDR RESULTS FOR THE FRONT-END ALONE AND FOR EACH END-TO-END SYSTEM.

when there is no chainsaw activity. However, passing the DNN classification step, the FDR drops to 11.4 and 9.5 s/hour when using AS and PCEN respectively. While this again reflects the superiority of the PCEN classifier against AS, at the same time it illustrates the ability of the proposed pipeline to achieve a very low FDR, which is very crucial considering the huge amounts of audio content involved in such applications.

We note that due to the fact that our evaluation is based on real audio recordings, it was not possible to know the SNR in the test set. To provide however an impression about the challenging acoustic conditions that the system is facing, we uploaded in [27] an audio dataset with the recordings that were used for evaluating the end-to-end system. Each audio recording is uploaded with a corresponding .textrgrid file, so that the interested reader can open these files in Praat, navigate to the marked regions and listen to the actual chainsaw events in each recording.

## V. Conclusion

Acoustic detection of chainsaw may greatly assist human experts towards monitoring of illegal logging activity in protected areas. In the context of building a system to automatically detect chainsaw events, it is shown that the Summation of Residual Harmonics method [21], originally developed for applications related to speech, provides an efficient front end also for acoustic activity detection in a natural environment. Combined with a binary DNN classifier at the back end, we managed to construct a pipeline that detects 73.5 % of low-SNR chainsaw events at a FDR of 9.5 s/hour. Our results demonstrate the advantage gained by using PCEN as a feature pre-processing step and highlight the importance of data augmentation in the classification performance. The presented system can be freely downloaded in the form of python code [20].

## References

[1] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1–9, 2014.

[2] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, 2019.

[3] S. Ntalampiras and I. Potamitis, "Detection of human activities in natural environments based on their acoustic emissions," in *Proc. of EUSIPCO*. IEEE, 2012, pp. 1469–1473.

[4] M. Ghiurcau, C. Rusu, R. Bilcu, and J. Astola, "Audio based solutions for detecting intruders in wild areas," *Signal Processing*, vol. 92, no. 3, pp. 829–840, 2012.

[5] I. Karmiris, C. Astaras, K. Ioannou, I. Vasiliadis, D. Youlatos, N. Stefanakis, A. Chatziefthimiou, T. Kominos, and A. Galanaki, "Estimating livestock grazing activity in remote areas using passive acoustic monitoring," *Information*, vol. 12, no. 8, pp. 290, 2021.

[6] L. Czúni and P. Varga, "Time domain audio features for chainsaw noise detection using WSNs," *IEEE Sensors Journal*, vol. 17, no. 9, pp. 2917–2924, 2017.

[7] M.e Ghiurcau, C. Rusu, and R. Bilcu, "A modified tespar algorithm for wildlife sound classification," in *Proc. of IEEE International Symposium on Circuits and Systems*. IEEE, 2010, pp. 2370–2373.

[8] M. Ghiurcau and C. Rusu, "About classifying sounds in protected environments," in *Proc. of International Symposium on Electrical and Electronics Engineering (ISEEE)*. IEEE, 2010, pp. 84–87.

[9] V. Andrei, H. Cucu, and L. Petrică, "Considerations on developing a chainsaw intrusion detection and localization system for preventing unauthorized logging," *Journal of Electrical and Electronic Engineering*, vol. 3, no. 6, pp. 202–207, 2015.

[10] J. Colonna, B. Gatto, E. Dos Santos, and E. Nakamura, "A framework for chainsaw detection using one-class kernel and wireless acoustic sensor networks into the amazon rainforest," in *Proc. of IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2016, vol. 2, pp. 34–36.

[11] L. Grama, E. Buhuş, and C. Rusu, "Acoustic classification using linear predictive coding for wildlife detection systems," in *International Symposium on Signals, Circuits and Systems (ISSCS)*, 2017.

[12] Y. Wang, P. Getreuer, T. Hughes, R. Lyon, and R. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.

[13] I. McLoughlin, H. Zhang, Y. Xie, Z.and Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[14] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 559–563.

[15] I. Choi, K. Kwon, S. Bae, and N. Kim, "Dnn-based sound event detection with exemplar-based approach for noise reduction," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016, pp. 16–19.

[16] V. Lostanlen, K. Palmer, E. Knight, C. Clark, H. Klinck, A. Farnsworth, T. Wong, J. Cramer, and J. Bello, "Long-distance detection of bioacoustic events with per-channel energy normalization," *arXiv preprint arXiv:1911.00417*, 2019.

[17] K. Piczak, "Environmental sound classification with convolutional neural networks," in *25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[18] J. Salamon and J. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[19] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[20] "Chainsaw sound detection system," https://github.com/spl-icsforth/An-open-access-system-for-long-range-chainsaw-sound-detection.

[21] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. of 12th Annual Conference of the International Speech Communication Association*, 2011.

[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 960–964.

[23] Louis C Sutherland and Gilles A Daigle, "Atmospheric sound propagation," *Handbook of Acoustics*, vol. 28, pp. 305–329, 1998.

[24] S. Stevens, J. Volkmann, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[25] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, S. Farnsworth, A.and Kelling, and J. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.

[26] L. Todor, V. Zoicas, L. Grama, and C. Ru, "The SPG (signal processing group) sound database," *Novice Insights*, vol. 15, no. 1, pp. 62–65, 2014.

[27] "A dataset of environmental audio recordings containing chainsaw events," https://zenodo.org/record/5824433#.YgeaVN_P1PY.