LONG-DISTANCE DETECTION OF BIOACOUSTIC EVENTS
WITH PER-CHANNEL ENERGY NORMALIZATION

Vincent Lostanlen[1,2], Kaitlin Palmer[3], Elly Knight[4], Christopher Clark[1],
Holger Klinck[1], Andrew Farnsworth[1], Tina Wong[2], Jason Cramer[2], and Juan Pablo Bello[2]

[1] Cornell Lab of Ornithology, Ithaca, NY, USA
[2] New York University, Center for Urban Science and Progress, New York, NY, USA
[3] San Diego State University, San Diego, CA, USA
[4] University of Alberta, Edmonton, AB, USA

ABSTRACT

This paper proposes to perform unsupervised detection of bioacoustic events by pooling the magnitudes of spectrogram frames after per-channel energy normalization (PCEN). Although PCEN was originally developed for speech recognition, it also has beneficial effects in enhancing animal vocalizations, despite the presence of atmospheric absorption and intermittent noise. We prove that PCEN generalizes logarithm-based spectral flux, yet with a tunable time scale for background noise estimation. In comparison with pointwise logarithm, PCEN reduces false alarm rate by 50x in the near field and 5x in the far field, both on avian and marine bioacoustic datasets. Such improvements come at moderate computational cost and require no human intervention, thus heralding a promising future for PCEN in bioacoustics.

Index Terms— Acoustic noise, acoustic sensors, acoustic signal detection, spectrogram, underwater acoustics.

1. INTRODUCTION

The deployment of autonomous recording units offers a minimally invasive sampling of acoustic habitats [1], with numerous applications in ecology and conservation biology [2]. In this context, there is an extensive literature on tailoring spectrogram parameters to a specific task of detection or classification: the effects of window size, frequency scale, and discretization are now well understood [3, 4]. However, the important topic of loudness mapping, i.e. representing contrast in the time–frequency domain, has received less attention.

This article investigates the impact of distance between sensor and source on the time–frequency representation of acoustic events. In particular, we point out that measuring local contrast by a difference in pointwise logarithms, as is routinely done in machine learning for bioacoustics, suffers from numerical instabilities in the presence of atmospheric attenuation and intermittent noise. To address this problem, we propose to employ an adaptive gain control technique known as per-channel energy normalization (PCEN) [5].

We deliberately err on the side of design simplicity: rather than training a sophisticated classifier, we apply a constant threshold on the time series of max-pooled PCEN magnitudes. In doing so, our goal is not to achieve the lowest possible false alarm rate, but to argue in favor of replacing the logarithmic mapping of loudness by PCEN in all systems for long-distance sound event detections, including more powerful yet opaque ones such as deep neural networks [6, 7].

Section 2 discusses the theoretical benefits of such a replacement: it proves that PCEN extends temporal context beyond a single temporal frame, thus improving effective detection radius. Sections 3 and 4 present applications to avian and underwater bioacoustics respectively, thereby revealing complementary issues: while bird call detection focuses on mitigating atmospheric absorption at high audible frequencies (1–10 kHz), whale call detection focuses on mitigating the interference of amplitude-modulated noise from near-field passing ships at low audible frequencies (50–500 Hz).

2. SPECTROTEMPORAL MEASURES OF NOVELTY

2.1. Averaged spectral flux

Let $\mathbf{E}(\mathbf{x})[t, f]$ be the magnitude spectrogram of some discrete-time waveform $\mathbf{x}[t]$. In full generality, the ordinal variable f may either represent frequency on a linear scale, a mel scale, or a logarithmic scale. Given $\mathbf{E}$, an implementation of spectral flux composes three operators: loudness mapping, contrast estimation, and feature aggregation. In its most widespread variant, named *averaged spectral flux*, these three operators respectively correspond to pointwise logarithm, rectified differentiation, and frequential averaging:

$$\mathrm{SF}_{\mathrm{avg}}(\mathbf{x})[t] = \sum_f \frac{\max\left(\log\mathbf{E}(\mathbf{x})[t, f] - \log\mathbf{E}(\mathbf{x})[t-1, f], 0\right)}{N_{\mathrm{fr}}} \quad (1)$$

where N_{fr} is the number of frequency bands f in $\mathbf{E}$. The motivation underlying this design choice finds its roots in psychoacoustics, and notably the Weber-Fechner law, which states that the relationship between stimulus and sensation is logarithmic [8]. We may also remark that Equation 1 is invariant to gain. Indeed, multiplying the waveform $\mathbf{x}[t]$ by some constant $K \neq 0$ incurs a multiplication by K in each frequency band of $\mathbf{E}(\mathbf{x})$, and thus an additive bias of $\log K$ in $\log\mathbf{E}(x)$, which eventually cancels after first-order differentiation. In the case of a single point source at some distance d, the relative change in acoustic pressure K caused by a spherical wave propagation is proportional to $\frac{1}{d}$. Therefore, in a lossless medium without reflections, logarithm-based spectral flux is invariant to geometric spreading insofar as acoustic sources do not overlap in the time–frequency domain.
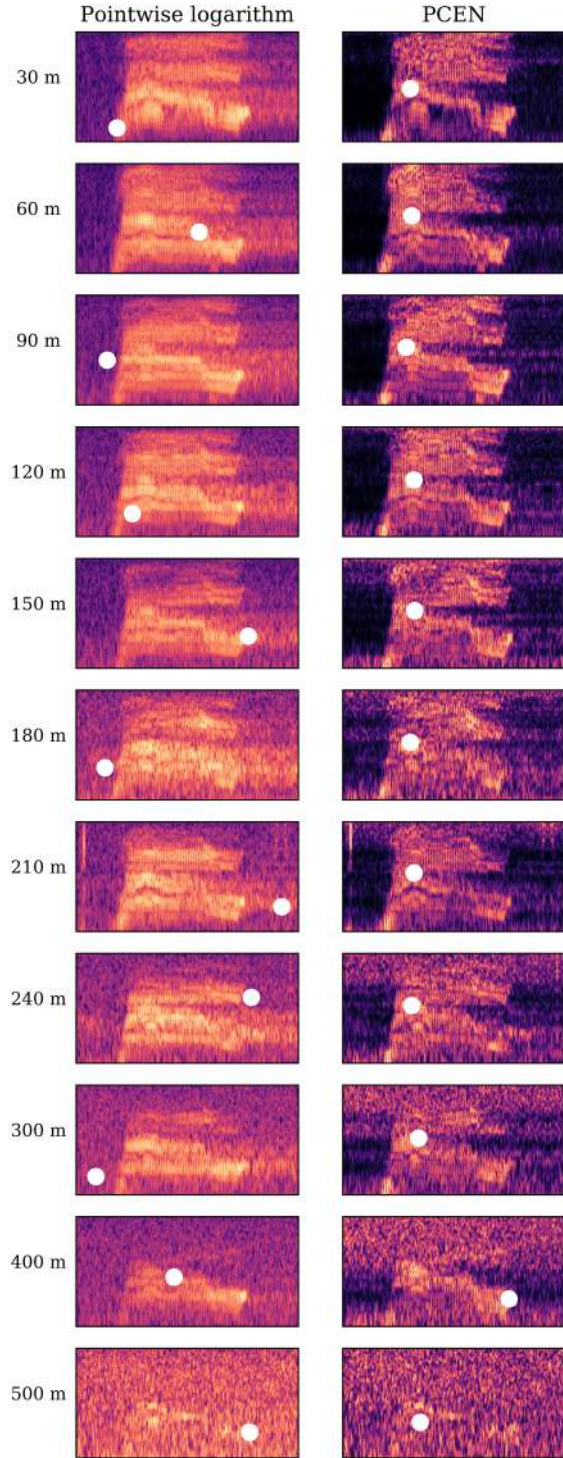
Figure 1: Effect of pointwise logarithm (left) and per-channel energy normalization (PCEN, right) on the same Common Nighthawk vocalization, as recorded from various distances. White dots depict the time–frequency locations of maximal spectral flux (left) or maximal PCEN magnitude (right). The spectrogram covers a duration of 700 ms and a frequency range between 2 and 10 kHz.

2.2. Max-pooled spectral flux

The situation is different in an absorbing medium. Indeed, heat conduction and shear viscosity, in conjunction with molecular relaxation processes, attenuate sine waves in quadratic proportion to their frequency [9]. Under standard atmospheric conditions, this attenuation is below $5\,\mathrm{dB\,km^{-1}}$ at $1\,\mathrm{kHz}$, yet of the order of $100\,\mathrm{dB\,km^{-1}}$ at $10\,\mathrm{kHz}$. As a result, bird calls spanning multiple octaves lose in bandwidth as they travel through air. A simple workaround is to replace the frequential averaging in Equation 1 by a max-pooling operator. This replacement yields the *max-pooled spectral flux*

$$\mathrm{SF_{max}}(\mathbf{x})[t] = \max_{f}\left(\log\mathbf{E}(\mathbf{x})[t,f] - \log\mathbf{E}(\mathbf{x})[t-1,f]\right), \quad (2)$$

which performs differentiation on a single frequency band, and is thus invariant to the low-pass filtering effect induced by absorption. However, as illustrated in Figure 1, the definition above suffers from numerical instabilities. Indeed, $\mathrm{SF_{max}}(\mathbf{x})$ discards all but two scalar values, corresponding to neighboring time–frequency bins in the spectrogram $\mathbf{E}(\mathbf{x})$.

2.3. Max-pooled per-channel energy normalization

In order to associate invariance and stability, this article proposes to increase the time scale of contrast estimation beyond a single spectrogram frame. To this end, we replace both the logarithmic mapping of loudness and the first-order differentiation by a procedure of per-channel energy normalization (PCEN). PCEN was recently introduced as a trainable acoustic frontend for far-field automatic speech recognition [5]. In full generality, PCEN results from an equation of the form

$$\mathrm{PCEN}(\mathbf{x})[t,f] = \frac{1}{r}\left(\frac{\mathbf{E}(\mathbf{x})[t,f]}{(\varepsilon + \mathbf{M}(\mathbf{x})[t,f])^{\alpha}} + \delta\right)^{r} - \frac{\delta^{r}}{r}, \quad (3)$$

where the gain control matrix $\mathbf{M}(\mathbf{x})$ proceeds from $\mathbf{E}(\mathbf{x})$ by first-order IIR filtering:

$$\mathbf{M}(\mathbf{x})[t,f] = s \times \mathbf{E}(\mathbf{x})[t-1,f] + (1-s) \times \mathbf{M}(\mathbf{x})[t-1,f]. \quad (4)$$

Note that the definition in Equation 3 differs from the original definition [5] by a factor of $\frac{1}{r}$. This is in order to allow the limit case $r \to 0$ to remain nonzero. Investigating the role of all parameters in PCEN is beyond the scope of this paper; we refer to the asymptotic analysis of [10] in this regard. Rather, we focus on the smoothing parameter $0 < s < 1$ as striking a tradeoff between numerical stability ($s \to 0$) and rapid adaptation to nonstationary in background noise ($s \to 1$). The following proposition, proven in Section 6, asserts that PCEN is essentially a generalization of spectral flux.

Proposition 2.1. *At the limit* $(s,\varepsilon,\alpha,r) \to (1,0,1,0)$ *in Equations 3 and 4, and for any finite value of* δ, $\mathrm{PCEN}(\mathbf{x})(t,f)$ *tends towards*

$$\log\left(\mathbf{E}(\mathbf{x})[t,f] + \mathbf{E}(\mathbf{x})[t-1,f]\right) - \log\mathbf{E}(\mathbf{x})[t-1,f], \quad (5)$$

which is a smooth approximation of the summand in Equation 1.

For the sake of simplicity, we adopt the PCEN parametrization that is prescribed by Proposition 2.1: we set $\varepsilon = 0$, $\alpha = 1$, $\delta = 1$, and $r = 0$. Derecursifying the autoregressive dependency in 4 and summarizing across frequencies yields the *max-pooled PCEN detection function*

$$\mathrm{PCEN_{max}}(\mathbf{x})[t] = \log\left(1 + \max_{f}\frac{\mathbf{E}(\mathbf{x})[t,f]}{s\sum_{\tau=0}^{+\infty}(1-s)^{\tau}\mathbf{E}(\mathbf{x})[t-\tau-1,f]}\right).$$
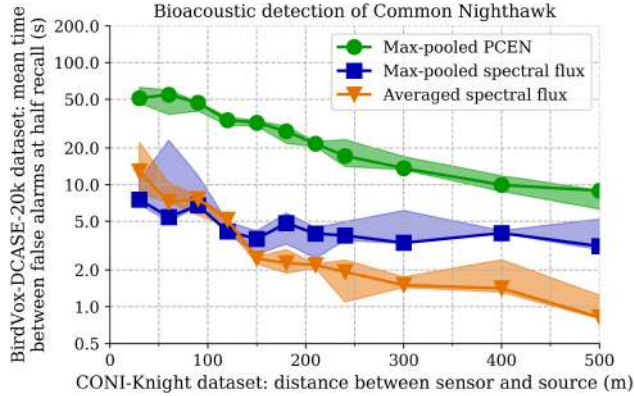$$(6)$$

Figure 2: Detection of Common Nighthawk calls: evolution of mean time between false alarms at half recall (MTBFA@50) as a function of distance between sensor of source. Shaded areas denote interquartile variations across individual birds. See Section 3 for details.
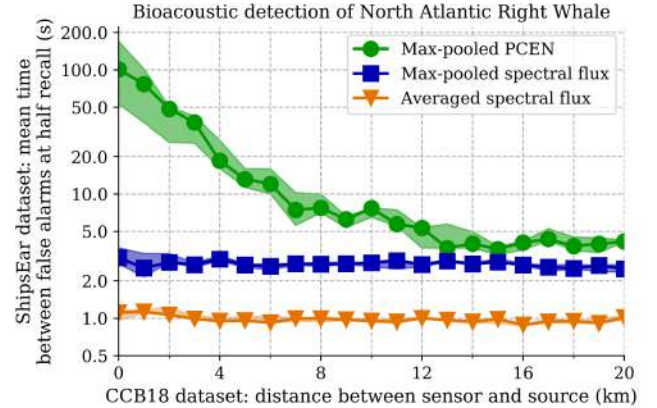

Figure 3: Detection of North Atlantic Right Whale calls: evolution of mean time between false alarms at half recall (MTBFA@50) as a function of distance between sensor of source. Shaded areas denote interquartile variations across days. See Section 4 for details.

3. APPLICATION TO AVIAN BIOACOUSTICS

3.1. CONI-Knight dataset of Common Nighthawk calls

We consider the problem of detecting isolated calls from breeding birds in a moderately cluttered habitat. To this end, we use the CONI-Knight dataset [11], which contains 64 vocalizations from five different adult male Common Nighthawks (*Chordeiles minor*), as recorded by 11 autonomous recording units in a regenerating pine forest north of Fort McMurray, AB, CA. The acoustic sensor network forms a linear transect, in which the distance between each microphone and the vocalizing individual varies from 30 m to 500 m. The dataset contains $11 \times 64 = 704$ positive audio clips in total, each lasting 700 ms. These clips were annotated by an expert, as part of a larger collection of continuous recordings which lasts seven hours in total. We represent each of these clips by their mel-frequency magnitude spectrograms, consisting of 128 bands between 2 kHz and 11.025 kHz, and computed with a Hann window of duration 12 ms (256 samples) and hop 1.5 ms (32 samples). These parameters are identical as in the state-of-the-art deep learning model for bird species recognition from flight calls [12]. We use the librosa implementation of PCEN [13] with $s = 0.09$, i.e. an averaging time scale of about $T = 100$ ms.

Figure 1 displays the mel-frequency spectrogram of one call at various distances, after processing them with either pointwise logarithm (left) or PCEN (right). Atmospheric absorption is particularly noticeable above 200 m, especially in the highest frequencies. Furthermore, we observe that max-pooled spectral flux is numerically unstable, because it triggers at different time–frequency bins from one sensor to the next. In comparison, PCEN is more consistent in reaching maximal magnitude at the onset of the call, and at the same frequency band.

3.2. Evaluation: mean time between false alarms at half recall

Our evaluation procedure consists in two stages: distance-specific threshold calibration and estimation of false alarm rate. In the first stage, we split the dataset of positive clips (i.e. containing one vocalization) into disjoint subsets of increasing average distance; sort the

values of the detection function over this subset in decreasing order; and set the detection threshold at the median value, thus yielding a detection recall of 50%. In the second stage, we run the detector on an external dataset of negative recordings, i.e. containing no vocalizations from the species of interest; apply the detection thresholds that were prescribed by the first stage; and count the number of false alarms, i.e. values of the detection function that are above threshold. Dividing the total duration of the dataset of negative recordings by this number of peaks above threshold yields the *mean time between false alarms at half recall* (MTBFA@50) of the detector, which grows in inverse proportion to false alarm rate. We repeat this operation over all available subsets to obtain a curve that decreases with distance, and which reflects the ability of the detection curve to generalize from near-field to far-field events.

3.3. Results and discussion

In the case of the Common Nighthawk, we choose the BirdVox-DCASE-20k dataset [14] as a source of negative recordings. A derivative of BirdVox-full-night [15], this dataset has been divided into 20k ten-second soundscapes from six autonomous recording units in Ithaca, NY, USA, and annotated by an expert for presence of bird calls. Among these 20k soundscapes, 9983 are guaranteed to contain no bird call, and *a fortiori* no Common Nighthawk call. These 9983 recordings amount to 27 hours of audio, i.e. over 30M spectrogram frames. For each detection function, we subtract the minimum value over each 10-second scene to the frame-wise value, in order to account for the nonstationarity in background noise at the scale of multiple hours.

Figure 2 summarizes our results. We find that max-pooled PCEN enjoys a five-fold reduction in false alarm rate with respect to average spectral flux. In addition, the false alarm rate at 300 m of max-pooled PCEN is comparable with the false alarm rate of averaged spectral flux at 30 m. As a post hoc qualitative analysis, we compute novelty curves for 200 recordings of outdoor noise from the ESC-50 dataset [16]: geophony (rain, wind), biophony (crickets), and anthropophony (helicopter, chainsaw). For max-pooled spectral flux, we find that the main causes of false alarms are pouring water (30% of total amount), crackling fire (17%), and water drops (10%).

4. APPLICATION TO MARINE BIOACOUSTICS

4.1. CCB18 dataset of North Atlantic Right Whale calls

We consider the problem of detecting isolated calls from whales in a noisy environment. To this end, we use the CCB18 dataset, which contains vocalizations from about 80 North Atlantic Right Whales (*Eubalaena glacialis*), as recorded by nine underwater sensors during five days in Cape Cod Bay, MA, USA. The distance between sensor and source is estimated by acoustic beamforming, similarly as in [17]. The dataset contains 40k clips in total, each lasting two seconds. These clips were annotated by an expert, as part of a larger collection of continuous recordings which lasts 1k hours in total. We represent each of these clips by their short-term Fourier transform (STFT) magnitude spectrograms, consisting of 128 bands between 8 Hz and 1 kHz, and computed with a Hann window of duration 128 ms and hop of 64 ms. We set $s = 0.33$, i.e. an averaging time scale of about $T = 1$ s. We choose the ShipsEar dataset as a source of negative recordings [18]. This dataset contains 90 ship underwater noise recordings from vessels of various sizes, most of them acquired at a distance of 50 m or less. These 90 recordings amount to 189 minutes of audio, i.e. 177k spectrogram frames.

4.2. Results and discussion

Figure 3 summarizes our results. First, we find that averaged spectral flux leads to poor false alarm rates, even in the near field. We postulate that this is because, in the CCB18 dataset, ship passage events occasionally introduce high received levels of noise. In other words, distance sets an upper bound, but no lower bound, on signal-to-noise ratio. Therefore, achieving 50% recall with averaged spectral flux requires to employ a low detection threshold, which in turn triggers numerous false alarms.

Secondly, we find that, across the board, replacing averaged spectral flux by max-pooled spectral flux allows a two-fold reduction in false alarm rate. We postulate that this improvement is due to the fact that whale calls are locally sinusoidal whereas near-field ship noise is broadband. Indeed, the max-pooled spectral flux of a chirp is above its averaged spectral flux, with a ratio of the order of N_{fr}; whereas the averaged and max-pooled spectral fluxes of a Dirac impulse are the same. Therefore, maximum pooling is particularly well suited to the extraction of chirps in noise [19].

Thirdly, we find that, in the near field, replacing spectral flux by PCEN leads to a 50-fold reduction in false alarm rate. We postulate that this is because ship noise has rapid amplitude modulations, at typical periods of 50 to 500 ms (i.e. engine speeds of 120 to 1200 rotations per minute). If this period approaches twice the hop duration (i.e. 128 ms in our case), short-term magnitudes $\log \mathbf{E}(x)[t-1, f]$ and $\mathbf{E}(x)[t, f]$ may correspond precisely to intake and expansion in the two-stroke cycle of the ship, thus eliciting large values of spectral flux. Nevertheless, in the case of PCEN, the periodic activation of one every other frame causes $\mathbf{M}(\mathbf{x})[t, f]$ to be of the order of $\frac{1}{2}\mathbf{E}(\mathbf{x})[t, f]$, assuming that the parameter T is large enough to encompass multiple periods. Therefore, $\text{PCEN}_{\max}(x)[t]$ peaks at $\log(\frac{3}{2})$ in the absence of any transient signal. This peak value is relatively low in comparison with the max-pooled PCEN of a near- or mid-field whale call.

Fourthly, we find that the false alarm rate of max-pooled PCEN increases exponentially with distance, until reaching comparable values as max-pooled spectral flux at a distance of 12 km. This decay is due, in part, to geometric spreading, but also to more complex acoustic phenomena, such as reflections and scattering with the surface as well as the ocean floor [20]. At these large distances, a successful detector should not only denoise, but also dereverberate whale calls. Max-pooled PCEN does not have any mechanism for dereverberation, and thus falls short of that objective. Thus, an ad hoc detection function is no longer sufficient, and the resort to advanced machine learning techniques appears as necessary. We must note, however, that deep convolutional networks in the time–frequency domain rely on the same functional blocks as max-pooled PCEN — i.e. rectified extraction of local contrast and max-pooling — albeit in a more sophisticated, data-driven fashion. Consequently, we believe that PCEN, whether parametrized by feature learning or by domain-specific knowledge, has a promising future in deep learning for environmental bioacoustics.

5. CONCLUSION

An adequate representation of loudness in the time–frequency domain is paramount to efficient sound event detection. This is particularly true in bioacoustic monitoring applications, where the source of interest may vocalize at a large distance to the microphone. Our experiments on the Common Nighthawk and the North Atlantic Right Whale demonstrate that, given a simple maximum pooling procedure across frequencies, per-channel energy normalization (PCEN) outperforms conventional (logarithm-based) spectral flux. Beyond the direct comparison between ad hoc detection functions at various distances, this study illustrates the appeal in replacing pointwise logarithm by PCEN in time–frequency representations of mid- and far-field audio signals. In the future, PCEN could be used, for example, as a similarity measure for spectrotemporal template matching; as an input to deep convolutional networks in the time–frequency domain [21]; or as a frequency-dependent acoustic complexity index for visualizing nonstationary effects in "false color spectrograms" [22] of open soundscapes.

6. APPENDIX: PROOF OF PROPOSITION 2.1

Proof. Applying Taylor's theorem to the exponential function yields

$$\frac{\delta^r}{r}\left[\left(1 + \frac{\mathbf{E}(\mathbf{x})[t,f])}{\mathbf{M}(\mathbf{x})[t,f]}\right)^r - 1\right] \approx \delta^r \log\left(1 + \frac{\mathbf{E}(\mathbf{x})[t,f]}{\mathbf{M}(\mathbf{x})[t,f]}\right) \quad (7)$$

with an error term proportional to $r\delta^r \log(1 + \frac{\mathbf{E}(\mathbf{x})[t,f]}{\mathbf{M}(\mathbf{x})[t,f]})^2$, which vanishes at the limit $r \to 0$ as long as $\mathbf{M}(\mathbf{x})[t, f]$ remains nonzero. On the left-hand side, we recognize $\mathbf{PCEN}(\mathbf{x})[t, f]$ with $\varepsilon = 0$ and $\alpha = 1$. On the right-hand side, the finite factor δ^r tends towards 1 for $r \to 0$. The limit $s \to 1$ allows to replace $\mathbf{M}(\mathbf{x})[t, f]$ by $\mathbf{E}(\mathbf{x})[t-1, f]$. We conclude with

$$\log\left(1 + \frac{\mathbf{E}(\mathbf{x})[t,f]}{\mathbf{E}(\mathbf{x})[t-1,f]}\right) = \log\left(\frac{\mathbf{E}(\mathbf{x})[t-1,f] + \mathbf{E}(\mathbf{x})[t,f]}{\mathbf{E}(\mathbf{x})[t-1,f]}\right)$$
$$= \log\left(\mathbf{E}(\mathbf{x})[t,f] + \mathbf{E}(\mathbf{x})[t-1,f]\right) - \log\mathbf{E}(\mathbf{x})[t-1,f]. \quad (8)$$

Interestingly, the distinction between Equation 1 and Equation 5 mirrors the distinction between the rectified linear unit (ReLU) $y \mapsto \max(y, 0)$ and the softplus $y \mapsto \log(1 + \exp(y))$ in deep learning. ∎

7. ACKNOWLEDGMENT

8. REFERENCES

[1] J. Shonfield and E. Bayne, "Autonomous recording units in avian ecological research: current use and future applications," *Avian Conservation and Ecology*, vol. 12, no. 1, pp. 42–54, 2017.

[2] M. G. Efford, D. K. Dawson, and D. L. Borchers, "Population density estimated from locations of individuals on a passive detector array," *Ecology*, vol. 90, no. 10, pp. 2676–2682, 2009.

[3] J. S. Ulloa, A. Gasc, P. Gaucher, T. Aubin, M. Réjou-Méchain, and J. Sueur, "Screening large audio datasets to determine the time and space distribution of screaming piha birds in a tropical forest," *Ecological informatics*, vol. 31, pp. 91–99, 2016.

[4] E. C. Knight, S. P. Hernandez, E. M. Bayne, V. Bulitko, and B. V. Tucker, "Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks," *Bioacoustics*, pp. 1–19, 2019. [Online]. Available: https://doi.org/10.1080/09524622.2019.1606734

[5] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proc. IEEE ICASSP*, 2017, pp. 5670–5674.

[6] P. Zinemanas, P. Cancela, and M. Rocamora, "End-to-end convolutional neural networks for sound event detection in urban environments," in *Proceedings of the Conference of Open Innovations Association (FRUCT)*, April 2019, pp. 533–539.

[7] M. Cartwright, J. Cramer, J. Salamon, and J. P. Bello, "TriCycle: Audio representation learning from sensor network data using self-supervision," in *Proc. IEEE WASPAA*. IEEE, Oct 2019.

[8] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE ICASSP*, vol. 6. IEEE, 1999, pp. 3089–3092.

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

[10] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, Jan 2019.

[11] E. C. Knight and E. M. Bayne, "Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software," *Bioacoustics*, pp. 1–16, 2018. [Online]. Available: https://doi.org/10.1080/09524622.2018.1503971

[12] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 141–145.

[13] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thomé, C. Raffel, D. Lee, K. Lee, O. Nieto, E. Battenberg, D. Ellis, R. Yamamoto, J. Moore, Z. Wei, R. Bittner, K. Choi, P. Friesch, F.-R. Stöter, S. K. Golu, S. Waloschek, S. Kranzler, R. Naktinis, D. Repetto, C. F. Hawthorne, C. Carr, W. Pimenta, P. Viktorin, P. Brossier, J. F. Santos, J. Wu, E. Peterson, and A. Holovaty, "librosa/librosa: 0.6.3," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1252297

[14] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-DCASE-20k: A dataset for bird audio detection in 10-second clips," 2018. [Online]. Available: https://zenodo.org/record/1208080

[15] ——, "BirdVox-full-night: A dataset and benchmark for avian flight call detection," in *Proc. IEEE ICASSP*, April 2018.

[16] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[17] C. Clark, R. Charif, D. Hawthorne, A. Rahaman, G. Givens, J. George, and C. Muirhead, "Acoustic data from the Spring 2011 bowhead whale census at Point Barrow, Alaska," *Journal of Cetacean Research and Management*, 2019.

[18] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," *Applied Acoustics*, vol. 113, pp. 64–69, 2016.

[19] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. DAFx*, 2013.

[20] R. P. Hodges, *Underwater acoustics: analysis, design, and performance of sonar*. Wiley, 2010.

[21] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *Under review*, 2019.

[22] M. Towsey, L. Zhang, M. Cottman-Fields, J. Wimmer, J. Zhang, and P. Roe, "Visualization of long-duration acoustic recordings of the environment," *Procedia Computer Science*, vol. 29, pp. 703–712, 2014.