euronoise 2021
25th to 27th of October
Madeira, Portugal - online

Self-calibrating Acoustic Sensor Networks with Per-Channel Energy Normalization

Vincent Lostanlen[1]

[1]Laboratoire des Sciences du Numérique de Nantes (LS2N), CNRS, Nantes, France
vincent.lostanlen@ls2n.fr

Abstract

The recent surge of machine learning models for wireless sensor networks brings new opportunities for environmental acoustics. Yet, these models are prone to statistical deviations, e.g., due to unforeseen changes in recording hardware or atmospheric conditions. In a supervised learning context, mitigating such deviations is all the more difficult that the area of coverage is vast. I propose to mitigate this problem by applying a form of adaptive gain control in the time-frequency domain, known as Per-Channel Energy Normalization (PCEN). While PCEN has recently been introduced for keyword spotting in the smart home, i show that it is also beneficial for outdoor sensing applications. Specifically, i discuss the deployment of PCEN for terrestrial bio-acoustics, marine bio-acoustics, and urban acoustics. Finally, i formulate three unsolved problems regarding PCEN, approached from the different perspectives of signal processing, real-time systems, and deep learning.

Keywords: acoustic signal detection, bioacoustics, far-field acoustics.

1 Introduction

The human ear exhibits a remarkable ability to decode acoustic events from a distant source. For example, a recent study has evaluated the intelligibility of shouted speech in a forest environment, and reported a classification accuracy of 75% for 17 French words at a distance of 90 meters [1]. What makes far-field recognition a challenging task is not solely that acoustic waves gradually decay in intensity as they spread away from the source, but also that they undergo absorption and reverberation depending on the propagation medium. These effects tend to alter the shape of the acoustic event of interest in terms of temporal envelope as well as spectral envelope. Moreover, the absorption spectrum of air is itself altered by meteorological variables such as temperature, pressure, and humidity [2]. Hence, in the context of acoustic sensor networks, guaranteeing the robustness of machine learning systems against missed or erroneous detections calls for distributed signal processing techniques which take these factors of variability into account.

Auditory neurophysiology provides useful domain-specific knowledge about the problem of far-field acoustic event detection, which can potentially be transferred to the domain of acoustical engineering. Whereas the recognition of speech involves high-level cognitive processes related to linguistic competence, meaningless stimuli such as dynamic ripples as less prone to individual learning effects; in this way, they shed light on the early stages of our auditory system [3]. At the level of the cochlea, two functional elements explaining the ability of human listeners to identify distant sounds are:

1. the band-pass selectivity of inner hair cell stereocilia, known as tonotopy [4]; and
2. the loudness adaptation of outer hair cells, known as electromotility [5].

Although tonotopy routinely appears in machine listening pipelines under the form of time–frequency decompositions, electromotility does not have such a well-established computational equivalent. For example, most applications of deep convolutional networks (convnets) in the time–frequency domain operate on the pointwise logarithm of the mel-frequency spectrogram or constant-Q wavelet scalogram, without any form of instance normalization. Note that batch normalization, a widespread technique in deep learning, does not qualify as an imitation of electromotility because its parameters are shared across recording conditions in the training set and are kept constant in the test set [6].

The situation changed in 2017 with the introduction of a nonlinear operator for post-processing spectrograms: per-channel energy normalization (PCEN) [7]. The key idea behind PCEN is to divide each "channel" (i.e., frequency band) in the mel-frequency spectrogram by a recursive estimate of the its expected value, under an assumption of local weak-sense stationarity. Contrary to batch normalization, the role of PCEN in the time–frequency domain may be compared to the role of acetylcholine as a regulator of electromotility in the cochlea [8]. PCEN has shown to outperform the pointwise logarithm on a task of far-field keyword spotting in the smart home and is now a component of the state-of-the-art deep learning model for automatic speech recognition: LEAF, which stands for Learnable Audio Frontend [9].

In this article, i demonstrate that the application scope of PCEN goes well beyond its original purpose of recognizing spoken queries (e.g., "OK Google") in a domestic environment. I review the usage of PCEN in recent publications from the scientific literature and outline the diversity of noise profiles against which PCEN is purposed: traffic noise in urban acoustics, insect noise in terrestrial bio-acoustics, vessel noise in marine bio-acoustics, and so forth. My point of view is that PCEN acts like a self-calibration mechanism for acoustic sensors: not only does it improve their area of coverage, it also reduces their dependency to spurious factors of variability; e.g., distance between sensor and source and weather conditions.

Section 2 illustrates the nonstationarity and nonuniformity of background noise in a sensor network named BirdVox as a motivating example for resorting to PCEN. Section 3 recalls the definition of PCEN and its known mathematical properties so far. Section 4 discusses the role of PCEN in some recent publications on outdoor acoustic sensor networks. Lastly, Section 5 formulates three unsolved problems with PCEN.

2 Motivating example

The BirdVox project[1] operates a network of nine acoustic sensors near Ithaca, NY, US, with the goal of monitoring the migration of three families of birds: thrushes *(Turdidae)*, warblers *(Parulidae)*, and sparrows *(Passerellidae)*. While aloft, these birds produce short vocalizations, known as flight calls, which carry a sort of "acoustic signature" of the species. Thus, the guiding idea behind BirdVox is to develop a supervised machine listening system for the automatic detection and classification of flight calls [10].

However, a major difficulty of this approach resides in the fact that supervised learning assumes the training set and the test set to be identically distributed. Yet, we observe that the mel–frequency spectra of the acoustic scene surrounding BirdVox sensors follows different empirical distributions depending on the geographical location of the sensor and the hour of day. Because human annotation for flight calls is particularly costly and time-consuming, collecting a training set which covers all recording conditions is not feasible in practice. Therefore, the role of PCEN will be to automatically calibrate all mel-frequency spectra to an identical distribution, so as to allow statistical generalization from training set to test set.

[1] For more information on the BirdVox project, visit: https://wp.nyu.edu/birdvox

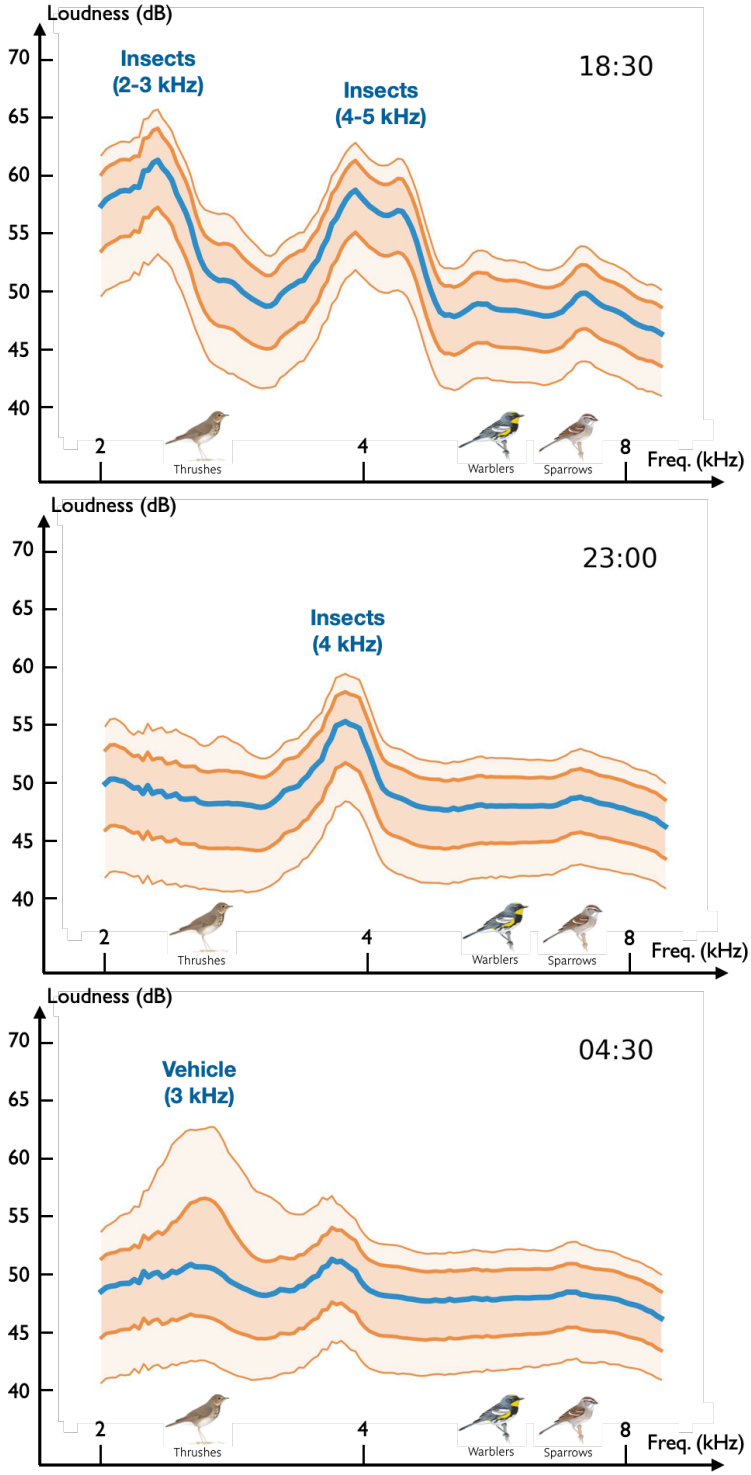

Figure 1 – An illustration of the nonstationarity of background noise in acoustic sensor networks. Each plot shows the distribution of short-term power spectral density for the same recording session, at various local times: 6:30 p.m. (top), 11 p.m. (middle), and 4:30 a.m. (bottom). The blue curve denotes the median value over a period of 30 minutes, while the shaded areas denote interdecile and intercentile ranges over the same period. The three bird drawings show the typical vocal ranges of three families of birds: thrushes *(Turdidae)*, warblers *(Parulidae)*, and sparrows *(Passerellidae)*.

Figure 1 illustrates the nonstationarity of background noise in the BirdVox sensor network; i.e., its dependency upon hour of day at a specific location. We find that different sources of noise interfere with the vocal ranges of the species of interest. Before dusk (6:30 p.m. local time), the stridulation of insects covers two narrow bands which are roughly one octave apart: i.e., at 2-2.5 kHz and 4-5 kHz respectively. At night (11 p.m.), the sound pressure level of background noise has reduced globally but leaves a passband around 4 kHz. Lastly, shortly before dawn (4:30 a.m.), we notice the presence of anthropogenic noise in the 2-4 kHz band: in this case, a passing train.

Conversely, Figure 2 illustrates the nonuniformity of background noise in the BirdVox sensor network; i.e., its dependency upon location at a specific hour of day. We observe that the power spectral density is approximately flat near Cayuga Lake (North of Ithaca) but presents evidence of traffic noise on the New York state route 34 (South of Ithaca) and evidence of biophonic noise near the Shindagin Hollow State Forest (South-East of Ithaca). These observations indicate that, even at the small scale of 1000 km^2 or so, machine listening in bioacoustics sensor networks faces the challenge of adapting to previously unseen recording conditions.

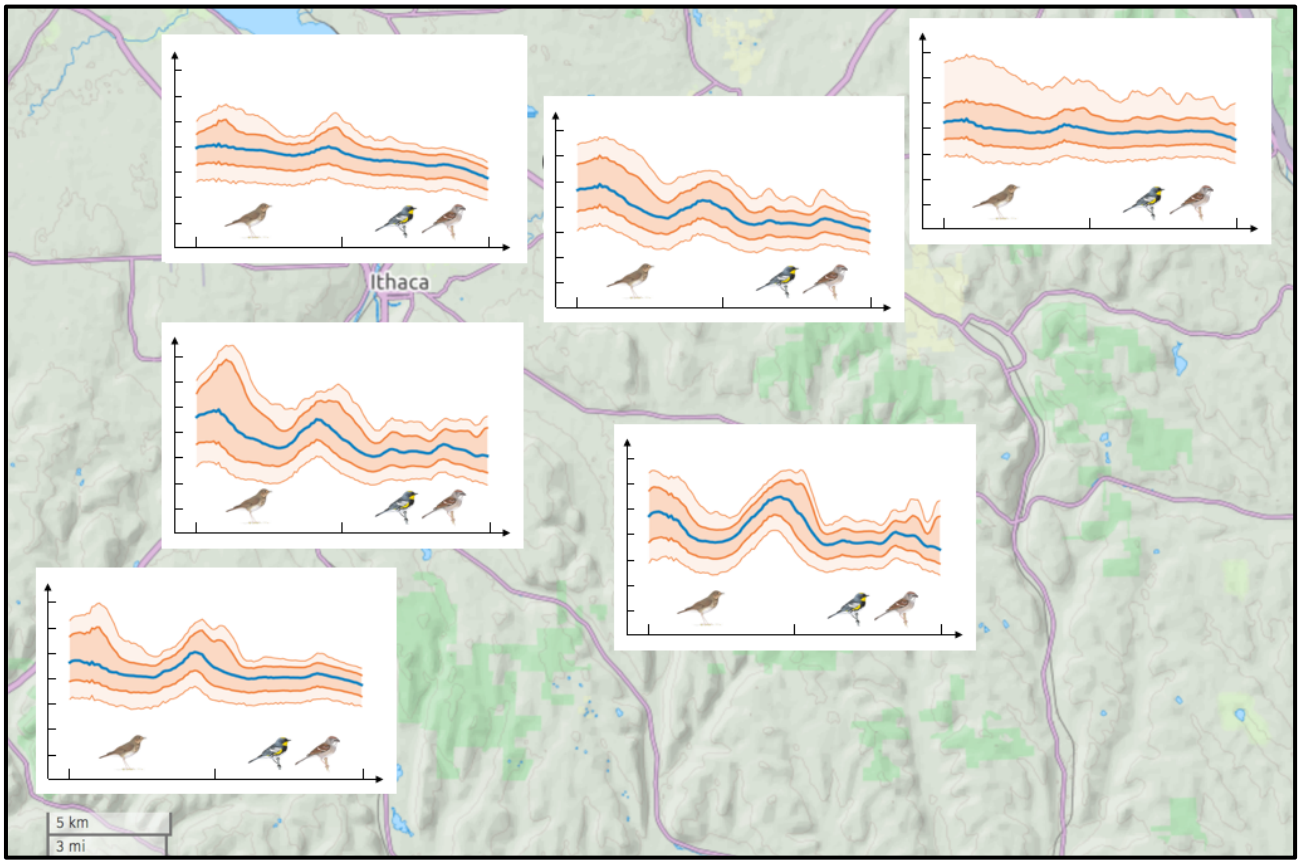

Figure 2 – An illustration of the nonuniformity of background noise in acoustic sensor networks. Each plot shows the distribution of short-term power spectral density for the same recording session at six different locations in Tompkins County, NY, US. The position of each line plot on the map represents its sensing location. Axis labels and legend are the same as in Figure 2.

3 Method

This section recalls the definition of PCEN and explains how it improves the robustness of spectrogram-based acoustic event detectors to spatiotemporal variations of background noise.

3.1 Definition

Let $\boldsymbol{E}(t, f)$ be the time–frequency representation of a monophonic audio signal. In full generality, the variable f may correspond to frequency in Hertz, as in the complex modulus of a short-term Fourier transform; to frequency in mels, as in a mel–frequency spectrogram; or to frequency in musical semitones, as in a constant-Q scalogram. The first operation in PCEN consists in applying a low-pass filter $\boldsymbol{\phi}_T(t)$ of decay constant equal to T, thus yielding a smoothed time–frequency representation $\mathbf{M}(t, f)$, defined as

$$\mathbf{M}(t, f) = \left(\mathbf{E} * \boldsymbol{\phi}_T\right)(t, f), \tag{1}$$

where the asterisk symbol $*$ denotes a convolution product and is implicitly broadcasted over different frequency "channels" f. The original publication on PCEN [7] proposes to define $\boldsymbol{\phi}_T(t)$ implicitly as an infinite impulse response (IIR) filter via a first-order autoregressive model of the form:

$$\mathbf{M}(t, f) = s\mathbf{E}(t, f) + (1 - \mathrm{s})\mathbf{M}(t - 1, f). \tag{2}$$

A previous publication [11] gives the formula linking the rate parameter s and the time scale parameter T. The next step in PCEN consists in the renormalization *per se* and involves various pointwise nonlinearities, yielding the PCEN-transformed spectrogram (or PCEN-gram for short):

$$\mathbf{PCEN}(t, f) = \left(\delta + \frac{\boldsymbol{E}(t,f)}{\left(\varepsilon + (\boldsymbol{E} * \boldsymbol{\phi}_T)(t,f)\right)^{\alpha}}\right)^{r} - \delta^r. \tag{3}$$

The parameters ε, α, δ, and r are application-specific and should be adjusted to the task at hand. The librosa implementation[2] defaults to the parameters of [7] and refers to [11] for practical recommendations on how to seek a good parameter setting manually. Another option is to define PCEN as a differentiable layer in a deep learning library such as PyTorch of TensorFlow, so that these parameters become trainable via gradient descent[3]. Furthermore, this approach allows all five parameters T, ε, α, δ, and r to be frequency-specific.

3.2 Properties

Intuitively, PCEN hinges on the fact that $\mathbf{M}(t, f)$ grows in proportion to $\mathbf{E}(t, f)$ if the frequency band f has stationary magnitudes at the time scale T, thus making the ratio $\mathbf{E}(t, f)/\mathbf{M}(t, f)$ of the order of one independent of noise level. Conversely, if the time–frequency region (t, f) coincides with the onset of an acoustic event, then we will have $\mathbf{E}(t, f) \gg \mathbf{M}(t, f)$ and thus $\mathbf{PCEN}(t, f) \gg 1$. Therefore, PCEN aims at canceling the amplitude fluctuations caused by background noise while preserving (and even enhancing) the local contrast near foreground onsets and offsets.

While the invention of PCEN in 2017 was driven by experimentation in speech processing and prior knowledge on computational auditory models, newer publications have supported its usefulness in machine learning for environmental acoustics. In particular, [11] has observed that PCEN converts noise from

[2] Official website of the librosa package for audio signal processing in Python: https://librosa.org/
[3] For a PyTorch implementation of PCEN, visit: https://github.com/daemon/pytorch-pcen

acoustic sensor networks into additive, white, quasi-Gaussian noise. Moreover, [11] has proven guarantees of numerical stability of PCEN with respect to the equalization of $\mathbf{E}(t, f)$, thus suggesting that variations in the absorption spectrum of air due to atmospheric conditions has almost no effect on the PCEN-gram.

4 Application scope

This section explains how the aforementioned definition and properties of PCEN can serve multiple application contexts, among which: terrestrial bioacoustics, marine bioacoustics, and urban acoustics.

4.1 Terrestrial bioacoustics

Going back to the problem statement of BirdVox (see Section 2), the replacement of a log-mel-spectrogram frontend by a PCEN frontend has significantly improved the ability of convnets to generalize to previously unseen recording conditions, such as held-out sensor locations or dawn vs. dusk. On a task of flight call detection (BirdVox-full-night dataset), [12] has reported reductions in miss rate between 15% to 110% depending on sensor locations. Note that, for this matter, the parameters of PCEN had to be adjusted between the original indoor application [7] and the outdoor application of BirdVox. Another example of publication which relies on deep learning with PCEN for terrestrial bio-acoustics is [13], which proposes to classify flight calls in terms of species, family, and genus.

Interestingly, PCEN is not solely beneficial as a pre-processing stage for machine learning in the time-frequency domain: it can also serve as the basis of feature engineering. A recent study on human vocal imitations of birdsong [14] has defined the following novelty curve for vocal activity detection:

$$\mathbf{Activity}(t) = \log\left(\frac{\max_f \mathbf{PCEN}(t,f) - \min_f \mathbf{PCEN}(t,f)}{\mathrm{median}_{t'}\left(\max_{f'} \mathbf{PCEN}(t',f') - \min_{f'} \mathbf{PCEN}(t',f')\right)}\right), \tag{4}$$

and managed to accurately segment both bird songs and their whistled imitations into syllabic units[4].

4.2 Marine bioacoustics

To the best of my knowledge, the only application of PCEN to marine bioacoustics to date is [15], which proposes to detect vocalizations from North Atlantic Right Wales *(Eubalaena glacialis)*. To this end, the paper proposes an analogy between PCEN and spectral flux at the limit case: $\varepsilon \to 0$, $\alpha \to 1$, $r \to 0$. Denoting by $\mathbf{PCEN_0}(t, f)$ this limit case, one obtains a simple PCEN-based definition of perceptual acoustic flux:

$$\mathbf{Flux}(t) = \max_f \mathbf{PCEN_0}(t,f) = \log\left(1 + \max_f \frac{\mathbf{E}(t,f)}{s\sum_{\tau=0}^{+\infty} \mathbf{E}(\mathrm{t}-\tau-1,f)}\right). \tag{5}$$

The definition above is more robust to underwater noise than traditional, logarithm-based spectral flux. Indeed, the performance metric (MTBFA@50) of PCEN-based $\mathbf{Flux}(t)$ for whale calls as a distance of 8 km is the same as the performance of the baseline for near-field sounds below 100 meters. Of course, training a convnet on log-mel-spectrogram features would outperform a simple hand-crafted feature such as spectral flux; but the take-home message of this paper is that, even in the absence of any feature learning, PCEN extends the detection radius of bioacoustics sensor networks, all other things being equal.

[4]Link to Python source code for PCEN-based vocal activity detection: https://github.com/BirdVox/oudyk_vihar2019

4.3 Urban acoustics

A growing number of publications apply PCEN to solve machine learning problems in urban acoustics. The earliest one is [16], which trained a convnet in the waveform domain to detect and classify sounds in the URBAN-SED dataset. The originality of this work is that it proposes to learn the time-domain parameters of the filterbank alongside those of PCEN (see Equation 3). In this sense, [16] is a forerunner of LEAF.
In [17], the authors present a new pretext task for self-supervised learning in urban acoustic sensor networks, named TriCycle. On the SONYC-UST dataset [18], combining PCEN with TriCycle achieved state-of-the-art results. Interestingly, PCEN improved accuracy both on the pretext task and the downstream task while reducing the sensitivity to sensor location—an observation in accordance with [12] (see Section 4.1).

In [19], the authors propose to apply PCEN as part of an outdoor keyword spotting system. The motivation behind this work resides in prototyping an accessible interface for urban crosswalks so that visually impaired pedestrians can interact by voice with the automatic signalization and cross the street safely.

In [20], the authors present a convnet for vehicle engine noise classification. This convnet takes as input a tensor named "Mod-PCEN" with three dimensions: time, frequency, and amplitude modulation rate. The third dimension corresponds to the frequency dimension of a short-term Fourier transform (STFT) which is performed over each frequency bin in the PCEN-gram.

A limitation of PCEN, in its original definition, is that the time scale parameter T remains fixed. To address this limitation, a recent publication [21] has proposed to extend the definition of PCEN to a multiscale setting. The key idea consists in varying T according to a geometric progression and stacking all PCEN-grams corresponding to each value of T. The resulting three-way tensor may serve as input to the first layer of a convnet. Intuitively, each slice of the tensor represents a different time scale of stationarity. On a task of urban sound classification, the authors have observed the multiscale PCEN outperforms single-scale PCEN across all choices of T. They have also noted that multiscale PCEN is more robust than single-scale PCEN to random fluctuations in the reverberation time of the acoustic scene at hand.

5 Future perspectives

This section describes three research directions which, in my opinion, have the potential to improve the theoretical understanding and practical usability of PCEN in the near future.

5.1 Probabilistic analysis

Empirical studies [11, 12] have suggested that the self-calibration role of PCEN can be attributed to two separate effects: the decorrelation of subband magnitudes on one hand; and their quasi-Gaussianization on the other hand. As of today, the former effect has received a theoretical justification: [11, Prop III.3] has proven that the PCEN of a stationary source-filter model $\boldsymbol{x}(t) = \boldsymbol{a}(t) \times (\boldsymbol{e} * \boldsymbol{h})(t)$ is stable to deformations of $|\widehat{\boldsymbol{h}}|(\omega)$; i.e., the spectrum of the filter $\boldsymbol{h}(t)$. However, the latter of these two effects (namely, quasi-Gaussianization) remains poorly understood as of today. Formally speaking, i ask under what conditions the PCEN of some random stationary process $\boldsymbol{X}$ approximates a multivariate Gaussian i.i.d. process. This question has practical importance because draws a conceptual link with unsupervised machine learning, perhaps in the spirit of variational inference with inverse autoregressive normalizing flows [22].

In this regard, a potential first step could be to assume that $\boldsymbol{X}$ is an instance of Gaussian stationary noise. Under some technical restrictions on the choice of time–frequency representation, its power spectrogram $\boldsymbol{E}_{\boldsymbol{X}}(t, f)$ follows a χ^2 ("chi-squared") distribution with two degrees of freedom [23]. Then, one could approximate the denominator of Equation (3), $\left(\varepsilon + (\boldsymbol{E}_{\boldsymbol{X}} * \boldsymbol{\phi}_T)(t, f)\right)^{\alpha}$, by a χ^2 distribution with $k \gg 2$

degrees of freedom; and ultimately $\mathbf{PCEN}_{\boldsymbol{X}}$ by the power transform of a Fisher-Snedecor distribution. Despite these initial ideas, numerical bounds for the Gaussian approximation of PCEN remain to be found and a discussion of the general (nonstationary) case is lacking. We leave them as open research questions.

5.2 Embedded implementations

The past few years have witnessed a surge of edge computing in low-cost acoustic sensor networks. Two examples are SONYC [18], which extracts self-supervised convnet features; and CENSE [24], which extracts third-octave spectrograms. Machine listening "on the edge", as opposed to "in the cloud", brings new opportunities in terms of privacy by design, of fault tolerance, and of lightweight connectivity. Thus, given that PCEN is by essence a distributed algorithm, it makes sense to incorporate it within the toolkit of embedded routines for audio signal processing on low-cost sensors.

Although PCEN was invented by researchers at Google [7] and integrated by researchers at Baidu as part of a study on "production speech models"[5], the product departments of these companies haven't openly communicated about it. Thus, it is unclear whether PCEN is actually being deployed on the client side of the Google Assistant, and likewise for the Baidu DuerOS conversational platform.

At the same time, there is a promising avenue of research in the field of solid-state electronics to implement PCEN into the new generation of keyword spotting hardware. In particular, [26] has prototyped a CMOS chip, named "normalized acoustic feature extractor" (NAFE), which comprises a mixed-signal (analog and digital) approximation of PCEN via an integrate-and-fire scheme. In combination with a spiking neural network (SNN), this chip achieves state-of-the-art results in keyword spotting, even in the presence of real-world sources of noise (traffic, restaurant, and so forth), while consuming less than 1 µW in total.

Future work should investigate whether PCEN can be made compatible with emerging technologies in machine listening, such as solar-powered batteryless sensors with wireless IoT connectivity [27].

5.3 Beyond short-term calibration: towards decentralized nonstationary PCEN

PCEN is currently defined as a short-term calibration mechanism, operating at the level of spectrogram frames. Although this definition has practical advantages, such as a low latency and a small memory footprint, PCEN also puts strong constraints on the choice of parameters: T, ε, α, δ, and r. Indeed, these parameters are supposed to be kept constant and shared across all sensors.

In future research, one could imagine relaxing these constraints and allowing PCEN parameters to vary not only depending on frequency, but depending on the noise profile surrounding the sensor at a given time and location. Thus, PCEN parameters would themselves become nonstationary and nonuniform: in turn, they would depend on stationary hyperparameters via the prediction of a neural network, which would govern the influence of long-term acoustic environment (several hours) upon the calibration of the short-term acoustic environment (one second or less). This idea bears a similarity with previous work on context-adaptive neural networks [12].

Making PCEN context-adaptive will introduce a feedback loop in Equation (3), in the sense that the prediction of a neural network from the input $\mathbf{PCEN}(t, f)$ will serve as a regressor for the prediction of the parameters T, ε, α, δ, and r at the frame $t + 1$. This feedback loop is comparable to a top-down effect in auditory neurophysiology (see Section 1).

[5] Unpublished manuscript (2017) by Baidu Research: https://arxiv.org/abs/1705.04400

6 Conclusion

PCEN is a simple but effective way of reducing the dependency of spectrograms to unwanted sources of variability. In this paper, i have argued that, in the context of acoustic sensor networks, it plays the role of a self-calibrating mechanism. Indeed, it can operate “on the edge” in a purely data-driven fashion, without communication nor synchronization between sensors. Recent publications in bio-acoustics and urban acoustics demonstrate that PCEN has a wide scope of applicability. That being said, our scientific understanding of PCEN is still in infancy. I have listed three unsolved problems regarding PCEN, approached from different perspectives: signal processing, real-time systems, and deep learning.

Acknowledgements

This work is partially supported by NSF award 1633259 (BIRDVOX) and an Atlanstic 2020 award: TrAcS (“Trainable Acoustic Sensors”). I thank Mathieu Lagrange for helpful discussions. I thank Jessie Barry, Ian Davies, Tom Fredericks, Jeff Gerbracht, Sara Keen, Holger Klinck, Anne Klingensmith, Ray Mack, Peter Marchetto, Ed Moore, Matt Robbins, Ken Rosenberg, and Chris Tessaglia-Hymes for designing autonomous recording units and collecting data. I acknowledge that the land on which the data were collected is the unceded territory of the Cayuga nation, which is part of the Haudenosaunee (Iroquois) confederacy.

References

[1] Meyer, J.; Meunier, F.; Dentet, L.; Do Carmo Blanco, N.; Sèbe, F. Loud and Shouted Speech Perception at Variable Distances in a Forest. *Annual Conference of the International Speech Communication Association (ISCA)*, Hyderabad, India, September 2-6, 2018.

[2] Bass, H. E.; Sutherland, L. C.; Zuckerwar, A. J.; Blackstock, D. T.; Hester, D. M. Atmospheric absorption of sound: Further developments. *The Journal of the Acoustical Society of America*, Vol 97(1), 1995, pp 680-683.

[3] Depireux, D. A.; Simon, J. Z.; Klein, D. J.; Shamma, S. A. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology*. Vol 85(3), 2001, pp 1220-1234.

[4] Mann, Z. F.; Kelley, M. W. Development of tonotopy in the auditory periphery. *Hearing research*, Vol 276(1-2), 2011, pp 2-15.

[5] Robles, L.; & Ruggero, M. A. Mechanics of the mammalian cochlea. *Physiological reviews*, Vol 81(3), 2001, pp 1305-1352.

[6] Bjorck, J.; Gomes, C.; Selman, B.; Weinberger, K. Q. Understanding batch normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, December 3-8, 2018.

[7] Wang, Y.; Getreuer, P.; Hughes, T.; Lyon, R. F.; Saurous, R. A. Trainable Frontend for Robust and Far-Field Keyword Spotting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, US, March 5-9, 2017.

[8] Frolenkov, G. I. Regulation of electromotility in the cochlear outer hair cell. *The Journal of physiology*, Vol 576(1), 2006, pp 43-48.

[9] Zeghidour, N.; Teboul, O.; de Chaumont Quitry, F.; Tagliasacchi, M. LEAF : A Learnable Frontend for Audio Classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, virtual conference, July 18-24, 2021.

[10] Lostanlen, V.; Salamon, J.; Farnsworth, A.; Kelling, S.; Bello, J. P. Birdvox-full-night: A Dataset and Benchmark for Avian Flight Call Detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, US, April 15-20, 2018.

[11] Lostanlen, V.; Salamon, J.; Cartwright, M.; McFee, B.; Farnsworth, A; Kelling, S.; Bello, J. P. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, Vol 26 (3), 2019, pp 39-43.

[12] Lostanlen, V.; Salamon, J.; Farnsworth, A.; Kelling, S.; Bello, J. P. Robust sound event detection in bioacoustic sensor networks. *PLOS ONE*, Vol 14(10), 2019, e0214168.

[13] Cramer, J.; Lostanlen, V.; Farnsworth, A.; Salamon, J.; Bello, J. P. Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, virtual conference, May 4-8, 2020.

[14] Oudyk, K.; Wu, Y.; Lostanlen, V.; Salamon, J.; Farnsworth, A.; Bello, J. P. Matching Human Vocal Imitations to Birdsong: An Exploratory Analysis. In *Proceedings of the International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)*, London, UK, August 29-30, 2019.

[15] Lostanlen, V.; Palmer, K.; Knight, E.; Clark, C.; Klinck, H.; Farnsworth; Wong, T.; Cramer, J.; Bello, J. Long-distance Detection of Bioacoustic Events with Per-Channel Energy Normalization. In *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York, NY, US, October 25-26, 2019.

[16] Zinemanas, P.; Cancela, P.; Rocamora, M. End-to-end convolutional neural networks for sound event detection in urban environments. In *Proceedings of the IEEE Conference of Open Innovations Association (FRUCT)*, Moscow, Russia, April 8-12, 2019.

[17] Cartwright, M.; Cramer, J.; Salamon, J.; Bello, J. P. TriCycle: Audio Representation Learning from Sensor Network Data Using Self-Supervision. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 20-23, 2019.

[18] Cartwright, M., Méndez Méndez, A. E.; Cramer, J.; Lostanlen, V.; Dove, G.; Wu, H. H; Salamon, Justin; Nov, Oded; Bello, J. SONYC Urban Sound Tagging (SONYC-UST): A Multilabel Dataset from an Urban Acoustic Sensor Network. In *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York, NY, US, October 25-26, 2019.

[19] Muhsinzoda, M.; Corona, C. C.; Pelta, D. A.; Verdegay, J. L. Activating accessible pedestrian signals by voice using keyword spotting systems. In *Proceedings of the IEEE International Smart Cities Conference (ISC2)*, Casablanca, Morocco, April 14-17, 2019.

[20] Becker, L., Nelus, A., Gauer, J., Rudolph, L., Martin, R. Audio Feature Extraction for Vehicle Engine Noise Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, virtual conference, May 4-8, 2020.

[21] Ick, C.; McFee, B. Sound Event Detection in Urban Audio with Single and Multi-Rate PCEN. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, virtual conference, June 6-11, 2021.

[22] Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; Welling, M. Improved Variational Inference with Inverse Autoregressive Flow. In *Advances in neural information processing systems (NeurIPS)*, Barcelona, Spain, Dec 5-10, 2016.

[23] R. Badeau. *Preservation of whiteness in spectral and time-frequency transforms of second-order processes*. Technical Report, Institut Mines-Télécom, 2016.

[24] Gontier, F.; Lostanlen, V.; Lagrange, M.; Fortin, N.; Lavandier, C.; Petiot, J. F. Polyphonic training set synthesis improves self-supervised urban sound classification. *The Journal of the Acoustical Society of America*, Vol 149(6), 2021, pp 4309-4326.

[25] Wang, D.; Kim, S. J.; Yang, M.; Lazar, A. A.; Seok, M. A. Background-Noise and Process-Variation-Tolerant 109nW Acoustic Feature Extractor Based on Spike-Domain Divisive-Energy Normalization for an Always-On Keyword Spotting Device. In *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, US, Feb 13-22, 2021.

[26] Lostanlen, V.; Bernabeu, A.; Béchennec, J. L.; Briday, M.; Faucou, S.; Lagrange, M. Energy Efficiency Is Not Enough: Towards a Batteryless Internet of Sounds. In *Proceedings of the International Workshop on the Internet of Sounds (IWIS)*, Trento, Italy, Sep 1-13, 2021.