Per-Channel Energy Normalization: Theory and Practice

Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee,
Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello

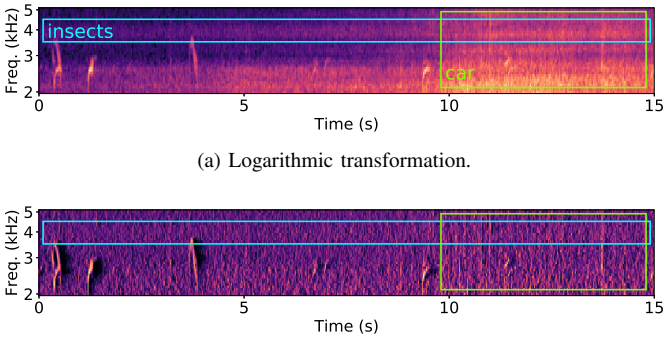
***Abstract*—In the context of automatic speech recognition and acoustic event detection, an adaptive procedure named per-channel energy normalization (PCEN) has recently shown to outperform the pointwise logarithm of mel-frequency spectrogram (logmelspec) as an acoustic frontend. This article investigates the adequacy of PCEN for spectrogram-based pattern recognition in far-field noisy recordings, both from theoretical and practical standpoints. First, we apply PCEN on various datasets of natural acoustic environments and find empirically that it Gaussianizes distributions of magnitudes while decorrelating frequency bands. Secondly, we describe the asymptotic regimes of each component in PCEN: temporal integration, gain control, and dynamic range compression. Thirdly, we give practical advice for adapting PCEN parameters to the temporal properties of the noise to be mitigated, the signal to be enhanced, and the choice of time-frequency representation. As it converts a large class of real-world soundscapes into additive white Gaussian noise (AWGN), PCEN is a computationally efficient frontend for robust detection and classification of acoustic events in heterogeneous environments.**

***Index Terms*—Acoustic noise, acoustic sensors, acoustic signal detection, signal classification, spectrogram.**

I. Introduction

FREQUENCY transposition is a major factor of intra-class variability in many sound classification tasks, including automatic speech recognition (ASR) [1], acoustic event detection (AED) [2], and bioacoustic species classification [3]. Tuning auditory filters to the perceptual mel scale provides a time-frequency representation, named mel-frequency spectrogram, in which the frequency transpositions of any periodic audio signal become vertical translations [4]. In the presence of a single source, this property allows convolutional operators in the time-frequency domain [5], such as convolutional neural networks [1] and time-frequency scattering [6], to extract pitch contours as spectrotemporal patterns, regardless of their fundamental frequency – a property known as equivariance [7], [8].

Yet, there is often more than one active source in real-world audio recordings, especially outdoors [9]. Even after narrowing down the classification task to the identification of the most salient source only (thereafter called foreground), the presence of background noise hinders equivariance along the mel-frequency axis [10]. Indeed, on one hand, intra-class variability causes frequency transposition of the foreground while leaving the background unaffected. On the other hand, equivariance is only possible if foreground and background happen to be transposed simultaneously. To reconcile these two assumptions, the background must result from a stochastic process which is stationary along the mel-frequency axis. White noise fits the required property of stationarity [11], but does not reflect real-world data: in the absence of any further processing, magnitudes in the mel-frequency spectrogram $\mathbf{E}(t,f)$ of an acoustic scene are typically not stationary, but sparse and strongly correlated, both along time t and mel frequency f [12]. Notwithstanding, the robustness of deep neural networks to adversarial additive perturbations is theoretically optimal if noise is additive, white, and Gaussian (AWGN) [13].



(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

Fig. 1. A soundscape comprising bird calls, insect stridulations, and a passing vehicle. The logarithmic transformation of the mel-frequency spectrogram (a) maps all magnitudes to a decibel-like scale, whereas per-channel energy normalization (b) enhances transient events (bird calls) while discarding stationary noise (insects) as well as slow changes in loudness (vehicle). Data provided by BirdVox. Mel-spectrogram and PCEN computed with default librosa 0.6.1 parameters and $T = 60$ ms (see Section IV).

Per-channel energy normalization (PCEN) [14] has recently been proposed as an alternative to the logarithmic transformation of the mel-spectrogram (logmelspec), with the aim of improving robustness to non-AWGN background noise. PCEN contains both dynamic range compression (DRC, also present in logmelspec) and adaptive gain control (AGC) with temporal integration. AGC is a prior stage to DRC involving a low-pass filter ϕ_T of support T, thus yielding

$$\mathbf{PCEN}(t,f) = \left(\frac{\mathbf{E}(t,f)}{(\varepsilon + (\mathbf{E} \overset{t}{*} \phi_T)(t,f))^{\alpha}} + \delta \right)^{r} - \delta^{r} \quad (1)$$

where α, ε, r, and δ are positive constants. While DRC reduces the variance of foreground loudness, AGC is intended to suppress stationary background noise, making PCEN suitable in cyberphysical systems for far-field ASR at the industrial scale [15] and a promising avenue for AED research [16].

V. Lostanlen, A. Farnsworth, and S. Kelling are with the Cornell Lab of Ornithology. J. Salamon, M. Cartwright, B. McFee, and J. P. Bello are with New York University.

In this article, we show empirically that PCEN Gaussianizes and decorrelates subband magnitudes in $\mathbf{E}(t,f)$ across mel-frequencies f, thus bringing them closer to AWGN and improving equivariance to frequency transposition. As an illustration, Figure 1 compares logmelspec and PCEN on a complex acoustic scene: while PCEN enhances chirped events, it converts background noise into a spectrotemporal texture that is devoid of long-range interactions.

Despite a growing body of evidence that PCEN improves keyword spotting and speech-to-text systems [17], little is known about the influence of each parameter in Equation 1. In addition to motivating PCEN with empirical findings, this article aims at elucidating its parameter space by means of theoretical and practical insights combined.

II. Statistical analysis

A. Datasets

The SONYC project has deployed 51 acoustic sensors in New York City, NY, USA, to monitor noise pollution [18]. To date, the data collected by SONYC exceeds 14k hours; in order to make the computation of magnitude histograms tractable, we extract 66 ten-second recordings. To maximize acoustic diversity in the construction of the SONYC subset, we collected ten different example recordings from YouTube for 22 urban sound classes – e.g. *car horn*, *crowd*, *jackhammer*, etc. – and extracted one-second VGGish features [19] from both the YouTube recordings and the SONYC recordings. For each class, we ranked the SONYC recordings by their mean squared frame-wise Euclidean distance to the example recordings. Out of the class-wise top-50, we manually curated three recordings, amounting to a subset of $22 \times 3 \times 10 = 660$ seconds of audio (7.3M coefficients).

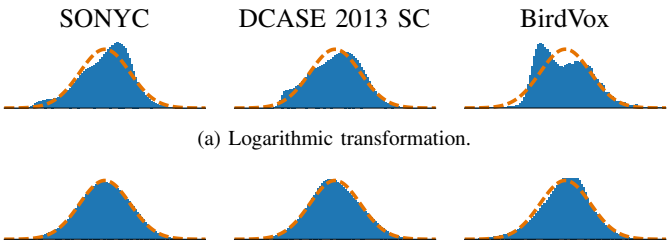
The DCASE 2013 Scene Classification (SC) dataset was recorded in various periurban locations – both indoor and outdoor – near London, UK, by a person wearing a binaural microphone [20]. It consists of 100 half-minute recordings from ten different soundscape classes (*open air market*, *restaurant*, *bus*, etc.), amounting to $100 \times 30 = 3000$ seconds of audio (33M coefficients).

The BirdVox project uses nine acoustic sensors near Ithaca, NY, USA, for monitoring avian migration [21]. Out of the 7k hours of audio in the full BirdVox data, we manually curate 15 one-minute recordings; the resulting subset amounts to $15 \times 60 = 900$ seconds of audio (10M coefficients).

B. Gaussianization of magnitudes

Figure 2 displays a histogram of all magnitudes in the matrix of mel-spectrogram coefficients, after either logarithmic transformation or PCEN. We observe that, for each of the three datasets, logmelspec magnitudes exhibit a skewed distribution, either left (BirdVox) or right (SONYC, DCASE 2013 SC). Nevertheless, PCEN successfully brings the distribution of magnitudes closer to a Gaussian distribution.

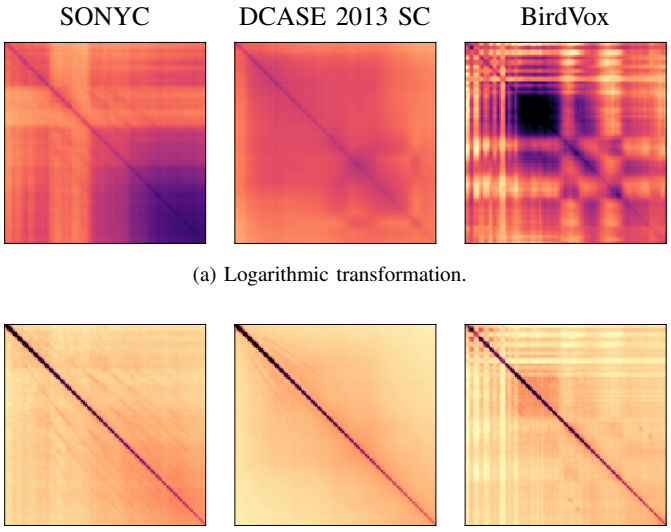
The Shapiro-Wilk test of normality indicates statistically significant evidence to reject the claim that the logarithmic transformation Gaussianizes the distribution of spectrogram magnitudes ($p < 0.005$ on all three datasets). At the same time, the same test fails to reject the null hypthesis of normality in the distribution of PCEN magnitudes.



(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

Fig. 2. Distributions of magnitudes in the mel-frequency spectrogram after logmelspec (a), and PCEN (c), as estimated on three datasets of acoustic scenes: SONYC (left); DCASE 2013 SC (middle); and BirdVox (right). Each distribution is scaled to null mean and unit variance, and discretized with 500 histogram bins ranging between -4 and 4. For comparison, the dashed line indicates the standard normal distribution. See Subsection II-B for details.



(a) Logarithmic transformation.

(b) Per-channel energy normalization (PCEN).

Fig. 3. Covariance matrices of frequency channels after logarithmic transformation (a) and PCEN (b), as estimated on three datasets of acoustic scenes: SONYC (left); DCASE 2013 (middle); and BirdVox (right). Darker shades indicate larger covariances in absolute value. See Subsection II-C for details.

C. Decorrelation of frequency bands

Figure 3 displays the covariance matrices of mel-spectrogram coefficients across frequency channels. While the logarithmic transformation suffers from strong cross-correlations between non-adjacent bands, the covariance matrix of PCEN is close to identity, thus suggesting that noise is "whitened".

III. Asymptotic analysis

A. Temporal integration

Filtering each subband f in $\mathbf{E}(t,f)$ with ϕ_T aims at estimating the intensity of background noise at f while remaining invariant to the intensity of foreground events. Under the assumption that the amplitude modulations (AM) of the foreground at f are faster than those of the background, T should be chosen to be above typical periods of foreground AM and below those of background AM. The same can be said

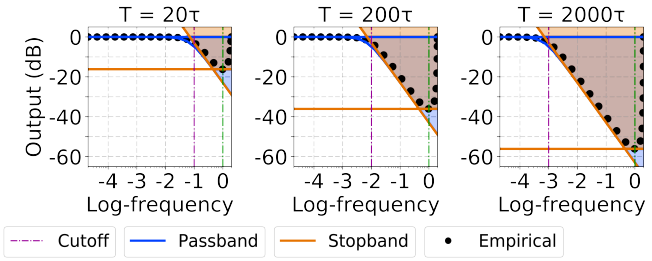


Fig. 4. Bode plot of the filter $|\widehat{\phi_T}(\omega)|^2$, measured in relative magnitude (dB) as a function of the ratio $\frac{\omega}{\omega_c}$ between frequency and cutoff frequency $\omega_c = \frac{2\pi\tau}{T}$. The time scale T is alternatively set to $10^1 2\tau$ (left), $10^2 2\tau$ (middle), and $10^3 2\tau$ (right). We observe a sidelobe falloff of 10 dB per decade in the stopband. Solid lines and shaded areas respectively denote asymptotic bounds and their corresponding error margins, as proved in Proposition III.1. The dashed purple (resp. green) vertical line denotes the cutoff (resp. Nyquist) frequency.

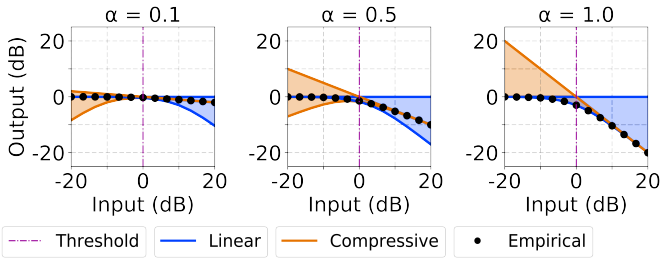


Fig. 5. Static compression characteristic of gain $\mathbf{M} \mapsto (\varepsilon + \mathbf{M})^{-\alpha}$, as a function of the ratio $\frac{\mathbf{M}}{\varepsilon}$ between input magnitude $\mathbf{M}$ and soft threshold ε. The exponent α is alternatively set to 0.1 (left), 0.5 (middle), and 1.0 (right). Solid lines and shaded areas respectively denote asymptotic bounds and their corresponding error margins, as proved in Proposition III.2. The dashed purple vertical line denotes the transition $\mathbf{M} = \varepsilon$.

of frequency modulation (FM): PCEN enhances chirped events in the mel-frequency spectrogram that move from one subband f to the next in less time than T while attenuating slower FM. Thus, T is the transition threshold between a stationary regime of background and a transient regime of foreground.

The original implementation of PCEN [14] defines $\phi_T(t)$ as a first-order IIR filter whose response to $\mathbf{E}(t,f)$ is

$$\mathbf{M}(t,f) = (\mathbf{E} * \phi_T)(t,f) = s\mathbf{E}(t,f) + (1-s)\mathbf{M}(t-\tau,f), \quad (2)$$

where $0 < s < 1$ is the weight of the associated autoregressive process (AR(1)) and τ is the discretization time step ("hop size" in seconds). Figure 4 illustrates the frequency response of ϕ_T for different values of T.

Proposition III.1. *The autoregressive filter ϕ_T defined in Equation 2 is a low-pass filter of gain* 0 dB*, cutoff frequency $\omega_c = \frac{2\pi\tau}{T} = \arccos(1 - \frac{s^2}{2(1-s)})$ at* 3 dB*, and sidelobe falloff of* 10 dB *per decade near ω_c.*

B. Adaptive gain control (AGC)

The smoothed mel-frequency spectrogram $\mathbf{M}(t,f)$ estimates the level of stationary background noise level in each frequency band f (where background is defined as slower AM than T), and serves to adapt the gain level in the denominator of the following equation:

$$\mathbf{G}(t,f) = \frac{\mathbf{E}(t,f)}{(\mathbf{M}(t,f) + \varepsilon)^\alpha}, \quad (3)$$

where $0 < \alpha < 1$ (resp. $\varepsilon > 0$) is the exponent (resp. soft threshold) of AGC. This stage resembles mean-variance renormalization [22], relative spectra (RASTA) [23], and cepstral mean normalization [24].

The parameter ε distinguishes two regimes: silent ($\mathbf{M}(t,f) \ll \varepsilon$) and active ($\mathbf{M}(t,f) \gg \varepsilon$). Multiplying $\mathbf{E}(t,f)$ by some constant C leads to $\mathbf{G}(t,f)$ being multiplied by approximately C in the silent regime and by $C^{1-\alpha}$ in the active regime. For ε of the order of unit roundoff and α close to 1, the following proposition proves that AGC is nonexpansive in quasi-silent frequency bands and strongly compressive in active frequency bands.

Proposition III.2. *$\mathbf{G}(t,f)$ is asymptotically equivalent to: (i) $\mathbf{E}(t,f)/\varepsilon^\alpha$ if $\mathbf{M}(t,f) \ll \varepsilon$ and to (ii) $\mathbf{E}(t,f)/\mathbf{M}(t,f)^\alpha$ if $\mathbf{M}(t,f) \gg \varepsilon$.*

Figure 5 illustrates the empirical fit of the characteristic $\mathbf{M} \mapsto (\mathbf{M} + \varepsilon)^{-\alpha}$ to the asymptotic regimes described in Proposition III.2. In the active regime, bringing α closer to 1 (resp. to 0) leads to more (resp. less) cancellation of background noise.

In the limit case $\varepsilon = 0$ and $\alpha = 1$, the proposition below proves that spectral equalization does not affect $\mathbf{G}$, because its effect on the numerator $\mathbf{E}$ is compensated by AGC with $\mathbf{M}$.

Proposition III.3. *Let $\mathbf{h}(t)$ be the impulse response of some acoustic environment or recording device. If $|\widehat{\mathbf{h}}|(f) = 0$ for $f < \frac{1}{T}$ and $|\widehat{\mathbf{h}}|(f) > 0$ for every f in the audible range, $\mathbf{G}$ is invariant to the filtering of the underlying waveform by $\mathbf{h}$.*

This result makes PCEN suitable for remote sensing applications, where acoustic models need to be robust to variations in the absorption properties of the environment, as well as in sensor technology [5], [25].

C. Dynamic range compression (DRC)

The last stage of PCEN is the addition of a positive bias δ to $\mathbf{G}(t,f)$, followed by pointwise exponentiation of the sum:

$$\mathbf{PCEN}(t,f) = (\mathbf{G}(t,f) + \delta)^r - \delta^r, \quad (4)$$

where $0 < r < 1$ (resp. $\delta > 1$) is the exponent (resp. soft threshold) of dynamic range compression. By nonnegativity of $\mathbf{E}$, $\mathbf{PCEN}$ is nonnegative and the existence of some pair (t,f) such that $\mathbf{PCEN}(t,f) = 0$ implies $\mathbf{E}(t',f) = 0$ for any $t' < t$.

The parameter δ distinguishes two regimes: quiet ($\mathbf{G} \ll \delta$) and loud ($\mathbf{G} \gg \delta$) after AGC. For $\mathbf{M}(t,f) \gg \varepsilon$, multiplying $\mathbf{E}(t,f)$ by some constant C leads to $\mathbf{G}(t,f)$ being multiplied by $C^{1-\alpha}$ in the background regime, and by $C^{r(1-\alpha)}$ in the foreground regime. Therefore, DRC is stronger for smaller values of r.

Proposition III.4. **PCEN** *is asymptotically equivalent to: (i) $r\delta^{(r-1)}\mathbf{G}$ for $\mathbf{G} \ll \delta$ and to (ii) $\mathbf{G}^r$ for $\mathbf{G} \gg \delta$.*

DRC resembles a spectral substraction in the context of speech restoration [26]. Figure 6 illustrates the empirical fit of the characteristic $\mathbf{G} \mapsto (\mathbf{G} + \delta)^r$ to the asymptotic regimes described in Proposition III.4.

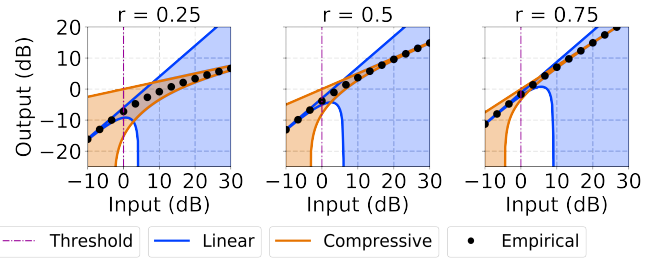


Fig. 6. Static compression characteristic of dynamic range compressions $\mathbf{G} \mapsto (\mathbf{G}+\delta)^r - \delta$, as a function of the ratio $\frac{\mathbf{G}}{\delta}$ between input magnitude $\mathbf{G}$ and soft threshold δ, for different values of r: 0.25 (left), 0.5 (right), and 0.75 (right). Solid lines and shaded areas respectively denote asymptotic bounds and their corresponding error margins, as proved in Proposition III.4.

IV. Practical recommendations

A. Setting parameters T and s

As discussed in Subsection III-A, the time constant T (directly linked to the dimensionless parameter s) should be longer than the time taken by a frequency-modulated foreground event to move from one subband f to another, adjacent subband. For a mel-frequency spectrogram of N bands ranging between $\mathrm{mel}(f_{\min})$ and $\mathrm{mel}(f_{\max})$, a rule of thumb for PCEN in bioacoustic species detection is

$$\frac{T \times c \times N}{\mathrm{mel}(f_{\max}) - \mathrm{mel}(f_{\min})} = K, \tag{5}$$

where c is the typical chirp rate of the species of interest, measured in mels per second; and K is some constant, depending on the reverberation properties of the environment. If the mel-frequency spectrogram is replaced by a constant-Q transform, the rule of thumb simply becomes $T \times c \times Q = K$, where c (resp. Q) is measured in octaves (resp. octaves per second). K is of the order of 1 in dry environments and above 10 in highly reverberant environment, e.g. bioacoustic event detection [27], [28].

In Equation 5, the optimal value of T do not solely depend on the physical phenomenon of interest (through the chirp rate c and reverberation constant K), but also on the choice of parametrization of the mel-frequency spectrogram (through N, $f_{\min}$, and $f_{\max}$). Therefore, in the context of hyperparameter optimization, any change in the resolution of the time-frequency representation should be reflected in an update of T, which in turn updates s through the following formula.

Proposition IV.1. *At a discrete rate τ^{-1}, the weight s of the autoregressive filter ϕ_T defined in Equation 2 is*

$$s = \sqrt{1 - \cos\frac{2\pi\tau}{T}} \left(\sqrt{3 - \cos\frac{2\pi\tau}{T}} - \sqrt{1 - \cos\frac{2\pi\tau}{T}} \right). \tag{6}$$

B. Setting parameters ε and α

In accordance with [14], we found empirically that T and α were the most important parameters. Although $\alpha = 1$ leads to an optimal cancellation of stationary background (see Prop. III.3), it may skew the distribution of magnitudes towards the right. Setting α below 1 reduces skewness and brings the background closer to AWGN. However, we have found ε to have no effect as long as it is set below unit roundoff.

C. Setting parameters δ and r

The effects of δ and r are more noticeable on the foreground time-frequency regions than on the background. The DRC threshold $\delta > 1$ sets a tradeoff between improving average foreground-to-background ratio ($\delta \to +\infty$ in highly noisy applications) and reducing variance in the loudness of foreground events ($\delta \to 1$). Moreover, if the foreground source is transient with respect to the time scale T and at distance d from the sensor, the energy in $\mathbf{E}(t,f)$ decays like $\frac{1}{d^2}$: therefore, under a fixed background noise level $\mathbf{M}(t,f)$, one has $\mathbf{G} \sim \frac{1}{d^2}$ and $\mathbf{PCEN} \sim \frac{1}{d^{2r}}$. We recommend $r = \frac{1}{2}$ for indoor applications ($d \sim 10\,\mathrm{m}$) and $r = \frac{1}{4}$ for outdoor applications ($d \sim 100\,\mathrm{m}$).

D. Open source implementation of PCEN in librosa

We release an open source implementation of PCEN in librosa v0.6.1 [29], whose default parameters are identical to [14]: $T = 400\,\mathrm{ms}$ (i.e. $s \approx 0.025$ with $\tau = 23\,\mathrm{ms}$), $\varepsilon = 10^{-6}$, $\alpha = 0.98$, $\delta = 2$, and $r = \frac{1}{2}$. Whereas these defaults are best suited to indoor applications (e.g. ASR in the smart home), bioacoustic event detection distinguishes itself by faster modulations of foreground (lower T), higher skewness of background magnitudes (lower α), a louder background (higher δ), and more distant sources (lower r). BirdVox adopts the following settings: $T = 60\,\mathrm{ms}$ with $Q = 50$ and $\tau = 1.5\,\mathrm{ms}$, $\varepsilon = 10^{-6}$, $\alpha = 0.8$, $\delta = 10$, and $r = 0.25$.

V. Conclusion

The human auditory system exhibits a remarkable ability to identify cues from distant sources, despite the presence of environmental absorption, reverberation, and noise [30]. At the level of the cochlea, two factors explaining such ability are: the band-pass frequential selectivity of inner hair cell stereocilia, known as tonotopy; and the loudness adaptation of outer hair cells, known as electromotility [31]. While the mel scale simulates tonotopy [32], PCEN simulates electromotility: as we have found empirically, it Gaussianizes real-world background noise, in real time, at a modest computational cost.

Unlike data-driven decorrelation procedures such as principal component analysis (PCA), PCEN can be distributed across sensors [18] and preserves the locality structure of harmonic patterns along the mel-frequency axis [33]. Furthermore, although it depends on five parameters (T, α, ε, r, and δ) that are possibly frequency-dependent, we have shown that each of these parameters has an interpretable purpose, and given asymptotic approximations of the PCEN equations in ideal regimes: silent vs. active (ε), stationary vs. transient (T), and quiet vs. loud (δ). In the context of deep learning for ASR and AED, our results could yield well-adapted initial values for the trainable version of PCEN [14], as well as a post hoc interpretation of all learned parameters.

Acknowledgment

The authors wish to thank Richard F. Lyon for helpful discussions. This work is partially supported by NSF awards 1633259 (BIRDCAST) and 1633206 (BIRDVOX), the Leon Levy Foundation, and a Google faculty award.

REFERENCES

[1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions Audio Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[2] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio Speech Music Process.*, 2015.

[3] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proc. IEEE ICASSP*, 2017.

[4] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale," in *Proc. IEEE ICASSP*, 1999.

[5] V. Lostanlen, "Convolutional operators in the time-frequency domain," Ph.D. dissertation, École normale supérieure, 2017.

[6] J. Andén and V. Lostanlen, "Time-frequency scattering for audio classification," in *Proc. IEEE MLSP*, 2015.

[7] R. Kondor and S. Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," in *Proc. ICML*, 2018.

[8] S. Mallat, "Understanding deep convolutional networks," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, p. 2015.0203, 2016.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[10] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Sig. Proc. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

[11] R. Badeau, "Preservation of whiteness in spectral and time-frequency transforms of second-order processes," Institut Mines Télécom, Tech. Rep., 2016.

[12] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[13] J.-Y. Franceschi, A. Fawzi, and O. Fawzi, "Robustness of classifiers to uniform ℓ^p and gaussian noise," in *Proc. AISTATS*, 2018.

[14] 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.

[15] E. Battenberg, R. Child, A. Coates, C. Fougner, Y. Gaur, J. Huang, H. Jun, A. Kannan, M. Kliegl, A. Kumar *et al.*, "Reducing bias in production speech models," *arXiv preprint arXiv:1705.04400*, 2017.

[16] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 335–371.

[17] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end models for small-footprint keyword spotting," *arXiv preprint arXiv:1803.10916*, 2018.

[18] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for the monitoring, analysis and mitigation of urban noise pollution," *Comm. Am. Comput. Mach.*, vol. in press, 2018.

[19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[20] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[21] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: a dataset and benchmark for avian flight call detection," in *Proc. IEEE ICASSP*, 2017.

[22] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[23] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[24] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[25] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos, and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing," in *Proc. IEEE ICASSP*. IEEE, 2014.

[26] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. ICASSP*. IEEE, 1984.

[27] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLoS One*, vol. 11, no. 11, 2016.

[28] J. Shonfield and E. Bayne, "Autonomous recording units in avian ecological research: current use and future applications," *Av. Cons. Ecol.*, vol. 12, no. 1, 2017.

[29] B. McFee, M. McVicar, S. Balke, C. Thom'e, C. Raffel, D. Lee, O. Nieto, E. Battenberg, D. Ellis, R. Yamamoto, J. Moore, R. Bittner, K. Choi, P. Friesch, F.-R. Stöter, V. Lostanlen, S. Kumar, 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: 0.6.1," May 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1252297

[30] J. Meyer, L. Dentel, and F. Meunier, "Speech recognition in natural background noise," *PLOS ONE*, vol. 8, no. 11, pp. 1–14, 11 2013.

[31] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiol. Rev.*, vol. 81, no. 3, pp. 1305–1352, 2001.

[32] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160101, 2017.

[33] V. Lostanlen and C. E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," in *Proc. ISMIR*, 2016.

[34] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Sig. Proc.*, vol. 62, no. 16, pp. 4114–4128, 2014.

VI. Supplementary material

A. Proof of Proposition III.1 on temporal integration

At a discrete rate τ^{-1}, the Fourier transform of ϕ_T is

$$\widehat{\phi_T}(\omega) = \mathscr{Z}\{\phi_T\}(\mathrm{e}^{\mathrm{j}\omega}) = \frac{s\mathrm{e}^{\mathrm{j}\omega}}{\mathrm{e}^{\mathrm{j}\omega} - (1-s)}. \tag{7}$$

The squared magnitude of the denominator simplifies as

$$\begin{aligned} |\mathrm{e}^{\mathrm{j}\omega} - (1-s)|^2 &= (\cos\omega - (1-s))^2 + (\sin\omega)^2 \\ &= 1 + (1-s)^2 - 2(1-s)\cos\omega \\ &= (1-(1-s))^2 + 2(1-s)(1-\cos\omega) \\ &= s^2 + 2(1-s)(1-\cos\omega). \end{aligned} \tag{8}$$

Let $\omega_0 = \frac{s}{\sqrt{1-s}}$. The equivalent $(1-\cos\omega) \sim \frac{\omega^2}{2}$ at the limit $\omega \to 0$ indicates a quadratic decay in the lowest frequencies, i.e. a sidelobe falloff of 10 dB per decade. On one hand, the double inequality $-1 \leq \cos\omega \leq 1$ leads to

$$\frac{\omega_0^2}{\omega_0^2+2} \leq |\widehat{\phi_T}|^2(\omega) \leq 1. \tag{9}$$

The upper (resp. lower) bound is attained if and only if $\omega = 0$ (resp. $\omega = \pi$). Consequently, ϕ_T has a DC gain of 0 dB at 0 Hz, and a gain of $10\log_{10}\frac{\omega_0^2}{2+\omega_0^2}$ (in dB) at the Nyquist frequency. On the other hand, the cutoff frequency at 3 dB ω_c is defined by $|\widehat{\phi_T}|^2(\omega_c) = \frac{1}{2}$, which can be reformulated as

$$1 - \cos\omega_\mathrm{c} = \frac{s^2}{2(1-s)} = \frac{\omega_0^2}{2}. \tag{10}$$

Solving Equation 10 with $\omega_\mathrm{c} > 0$ completes the proof. ■

Remark. For large T, we can approximate ω_c by ω_0, which is polynomial in s. Indeed, applying Taylor's theorem to the cosine function yields the inequality $|1 - \cos\omega - \frac{\omega^2}{2}| \leq \frac{\omega^4}{24}$, of which we deduce

$$\omega_0 \leq \omega_\mathrm{c} \leq \frac{\omega_0}{\sqrt{1 - \frac{\omega_\mathrm{c}^2}{12}}} \leq \omega_0 + \frac{\omega_0^2}{12} + \frac{\omega_0^3}{12}. \tag{11}$$

B. Proof of Proposition III.2 on adaptive gain control

First, applying Taylor's theorem to $x \mapsto (1+x)^{-\alpha}$ gives $|(\mathbf{M}+\varepsilon)^{-\alpha} - \varepsilon^{-\alpha}| \leq \alpha\frac{\mathbf{M}}{\varepsilon^{1+\alpha}}$, from which we deduce (i) with a relative error of at most $\alpha\frac{\mathbf{M}}{\varepsilon}$. Symmetrically, we have $|(\mathbf{M}+\varepsilon)^{-\alpha} - \mathbf{M}^{-\alpha}| \leq \alpha\frac{\varepsilon}{\mathbf{M}^{1+\alpha}}$ from which we deduce (ii) with a relative error of at most $\alpha\frac{\varepsilon}{\mathbf{M}}$. ■

C. Proof of Proposition III.4 on dynamic range compression

First, applying Taylor's theorem to $x \mapsto (1+x)^r$ gives $|\mathbf{PCEN} - r\delta^{(r-1)}\mathbf{G}| \leq \frac{1-r}{2}r\delta^{(r-1)}\mathbf{G}^2$, of which we deduce (i) with a relative error of at most $\frac{1-r}{2}\mathbf{G}$. Secondly, we have $\left|\left(1+\frac{1}{\mathbf{G}^r}\right)^r - \left(r + \frac{r}{\mathbf{G}^r}\right)\right| \leq \frac{r(1-r)}{2\mathbf{G}^{2r}}$, which, after adding the constant $(1-r)$, leads to (ii) with a relative error of at most $\frac{1-r}{\mathbf{G}^r}(1+\frac{r}{2\mathbf{G}^r})$ by application of the triangular inequality. ■

Remark. In both cases, the relative error is null if and only if $r = 1$, i.e. if there is no root compression.

D. Proof of Proposition IV.1 on the link between T and s

Let $\omega_0^2 = 2\left(1 - \cos\frac{2\pi\tau}{T}\right)$. Proposition III.1 yields $\omega_0^2 = \frac{s^2}{(1-s)}$, and thus the quadratic equation $s^2 + \omega_0^2 s - \omega_0^2 = 0$. Its discriminant is $\Delta = \omega_0^4 + 4 \times \omega_0^2 = \omega_0^2(\omega_0^2+4) > 0$, and its two real-valued roots are

$$s = \frac{-\omega_0^2 \pm \omega_0\sqrt{\omega_0^2+4}}{2} = \frac{\omega_0}{2}\left(\pm\sqrt{\omega_0^2+4} - \omega_0\right) \tag{12}$$

We retain the positive root $s > 0$. Replacing ω_0 by its definition in the above completes the proof. ■

E. Proof of Proposition III.3 on invariance to impedance curve

Proposition. *Let* $\mathbf{s}(t)$ *a realization of AWGN with null mean and unit variance. Let* $\mathbf{a}(t) > 0$ *a deterministic amplitude envelope* $\mathbf{h}(t)$ *a filter. Let* $\mathbf{E}_\mathbf{x}(t,f)$ *the mel-frequency spectrogram associated to the source-filter model* $\mathbf{x}(t) = \mathbf{a}(t) \times (\mathbf{s} * \mathbf{h})(t)$. *If*

1) $\forall t_0,\ \frac{\mathrm{d}\log a}{\mathrm{d}t}(t_0) \ll \frac{1}{T}$,
2) $\forall t_0,\ \int_{t_0}^{t_0+\tau}\mathbf{h}(t)\,\mathrm{d}t \ll \frac{1}{\tau}$, *and*
3) $\forall f_0,\ \frac{\mathrm{d}\log|\widehat{\mathbf{h}}|}{\mathrm{d}f}(f_0) \ll \frac{1}{\Delta f_0}$,

where Δf_0 *is the frequency interval between* f_0 *and its adjacent subbands on the mel scale, then* $\mathbf{PCEN}_\mathbf{x}(t,f) \approx \mathbf{PCEN}_\mathbf{s}(t,f)$.

We adapt a result from [34]. The mel-frequency spectrogram $\mathbf{E}_\mathbf{x}(t,f)$ can be defined as the complex modulus of the convolution between the signal $\mathbf{x}(t)$ and a filterbank $\psi_f(t)$ of N whose center frequencies f are tuned to the mel scale, ranging betweeen f_min and f_max. For a given f, the first hypothesis allows to factorize the amplitude envelope $\mathbf{a}(t)$ out of the convolution $(\mathbf{s} * \mathbf{h} * \psi_f)(t)$ Furthermore, one has:

$$\begin{aligned} \mathbf{E}_{(\mathbf{s}*\mathbf{h})}(t,f) &= |\mathbf{s} * \mathbf{h} * \psi_f|(t) \\ &= \frac{1}{2\pi}\left|\int_\mathbb{R} \widehat{\mathbf{s}}(\omega)\widehat{\mathbf{h}}(\omega)\widehat{\psi_f}(\omega)\exp(2\pi\mathrm{i}\omega t)\,\mathrm{d}t\right|. \end{aligned} \tag{13}$$

The second and third hypotheses allow to approximate $\widehat{\mathbf{h}}(\omega)$ by $\widehat{\mathbf{h}}(f)$ in the equation above. This approximation leads to $\mathbf{E}_{(\mathbf{s}*\mathbf{h})}(t,f) \approx |\widehat{\mathbf{h}}|(f)\mathbf{E}_\mathbf{s}(t,f)$. Combining the factorization of the amplitude term $\mathbf{a}(t)$ with the factorization of the filter $|\mathbf{h}|(\omega)$ leads to an approximation of the form:

$$|\mathbf{E}_\mathbf{x}(t,f) - \mathbf{a}(t)|\widehat{\mathbf{h}}|(f)\mathbf{E}_\mathbf{s}(t,f)| = \eta(t,f). \tag{14}$$

An upper bound on the expectation of the stochastic residual η is given in [34, Appendix]. Likewise, the first hypothesis allows to approximate $\mathbf{M}_\mathbf{x}(t,f)$ by $\mathbf{a}(t)|\widehat{\mathbf{h}}(f)|\mathbf{M}_s(t,f)$ with a stochastic residual term $\nu(t,f)$ of bounded expectation. In the active regime ($\mathbf{M} \gg \varepsilon$), AGC cancels $\mathbf{a}(t)$ and $|\widehat{h}(f)|$ in Equation 3. Thus, Proposition III.2 leads to the inequality

$$|\mathbf{G}_\mathrm{x} - \mathbf{G}_\mathrm{s}|(t,f) \leq \frac{\eta(t,f)}{(\varepsilon + \mathbf{M}_\mathbf{s}(t,f))^\alpha} + \frac{\alpha\nu(t,f)}{(\varepsilon + \mathbf{M}_\mathbf{s}(t,f))^{1+\alpha}}. \tag{15}$$

Because $\delta > 1$ and $r \geq 1$, DRC is nonexpansive (see Proposition III.4). Therefore, it maps each member of the approximate equality $\mathbf{G}_\mathbf{x}(t,f) \approx \mathbf{G}_\mathbf{s}(t,f)$ to $\mathbf{PCEN}_\mathbf{x}(t,f) \approx \mathbf{PCEN}_\mathbf{s}(t,f)$ while reducing the expectation of the residual (right-hand side of Equation 15), which completes the proof. ■