Per-Channel Energy Normalization: an Asymptotic Analysis

Vincent Lostanlen and others

***Abstract*—**

***Index Terms*—Acoustic signal detection, acoustic noise, multilayer neural network, supervised learning, ecosystems.**

I. Introduction

THE human auditory system exhibits a remarkable ability to identify cues from distant sources, despite the presence of environmental absorption, reverberation, and noise [**?**]. At the level of the cochlea, two of the predominant 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 [**?**]. Yet, the time-frequency representations commonly involved in computational auditory scene analysis (CASA) – such as the logarithm of the mel-frequency spectrogram (logmelspec) – imitate the former of these properties, but not the latter. This shortcoming hinders the robustness of spectrogram-based pattern recognition systems for automatic speech recognition (ASR) and acoustic event detection (AED). Indeed, most available datasets for training these systems were recorded under well-controlled experimental conditions: a high signal-to-noise ratios (SNR above 30 dB) and a microphone placed near the source, i.e. 1 m or less. Consequently,

Let $\mathbf{E}(t,f)$ be the mel-frequency spectrogram of a monophonic signal, that is, the matrix of its acoustic energies around times t within frequency bands f tuned to the perceptual mel scale – or, indeed, any "channel" variable other than time. The discrete-time implementation of PCEN begins by smoothing E with a first-order infinite impulse response (IIR) filter

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

where $0 < s < 1$ is the weight of the associated autoregressive (AR(1)) filter and h is the discretization time step ("hop size") of the filter. The smoothed mel-frequency spectrogram $\mathbf{M}(t,f)$ adjusts the gain level of the short-term energy $\mathbf{E}(t,f)$ by means of the renormalization

$$\mathbf{G}(t,f) = \frac{\mathbf{E}(t,f)}{\delta \times (\mathbf{M}(t,f)+\varepsilon)^{\alpha}}, \tag{2}$$

where $0 < \alpha < 1$ is the AGC exponent and $\delta > 1$ is the AGC factor.

$$\mathbf{PCEN}(t,f) = \frac{(\mathbf{G}(t,f)+1)^r - 1}{r} \tag{3}$$

$$= \frac{1}{\delta^r r}\left(\left(\frac{\mathbf{E}(t,f)}{(\mathbf{M}(t,f)+\varepsilon)^{\alpha}} + \delta^r\right) - \delta^r\right) \tag{4}$$

V. Lostanlen, A. Farnsworth, and S. Kelling are with the Cornell Lab of Ornithology. J. Salamon and J. P. Bello are with New York University. R. A. Saurous and R. F. Lyon are with Google LLC.

For the sake of conciseness, we drop the explicit dependency in (t,f) for scalar values in $\mathbf{E}$, $\mathbf{M}$, and $\mathbf{G}$.

II. Root compression

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

Proof. First, applying Taylor's theorem to $x \mapsto (1+x)^r$ gives $|\mathbf{PCEN} - \mathbf{G}| \leq \frac{1-r}{2}\mathbf{G}^2$, 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 $(1-r)$ and dividing the result by 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. In both cases, the relative error is null if and only if $r = 1$, i.e. if there is no root compression. ■

III. Renormalization

Proposition III.1. $\mathbf{G}$ *is asymptotically equivalent to: (i)* $\mathbf{E}/(\delta\varepsilon^{\alpha})$ *if* $\mathbf{M} \ll \varepsilon$ *and to (ii)* $\mathbf{E}/(\delta\mathbf{M}^{\alpha})$ *if* $\mathbf{M} \gg \varepsilon$.

Proof. 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 $\alpha\frac{\varepsilon}{\mathbf{M}}$. ■

IV. Transient and stationary regimes

Proposition IV.1. *Let* $T = \frac{1-s}{s}h$. *If* $s \ll 1$, *then* $\mathbf{E}$ *is asymptotically equivalent to: (i)* $\mathbf{M}$ *if* $T\left|\frac{\partial \log \mathbf{M}}{\mathrm{d}t}\right| \ll 1$ *and to (ii)* $T\frac{\partial \mathbf{M}}{\partial t}$ *if* $T\frac{\partial \log \mathbf{M}}{\partial t} \gg 1$.

Proof. Applying Taylor's theorem to $t \mapsto \mathbf{M}(t,f)$ near t_0 yields

$$\mathbf{M}(t_0-h,f) = \mathbf{M}(t_0,f) - h\frac{\partial \mathbf{M}}{\partial t}(t_0,f) + \frac{h^2}{2}\mathbf{R}(t_0,h,f) \tag{5}$$

where the residual term $\mathbf{R}(t_0,h,f)$ satisfies

$$|\mathbf{R}(t_0,h,f)| \leq \sup_{u\in[t_0-h,t_0]}\left|\frac{\partial^2 \mathbf{M}}{\partial t^2}(u,f)\right|. \tag{6}$$

Equation 1

$$\mathbf{M}(t_0,f) + \frac{1-s}{s}h\frac{\partial \mathbf{M}}{\partial t}(t_0,f) = \mathbf{E}(t_0,f) + \frac{1-s}{s}\frac{h^2}{2}\mathbf{R}(t_0,h,f). \tag{7}$$

We define $T = \frac{1-s}{s}h$ and let h approach zero while decreasing $s = \frac{h}{h+T}$ accordingly, so that T remains constant. Once in continuous time, we obtain the following differential equation:

$$\mathbf{M}(t_0,f) + T\frac{\partial \mathbf{M}}{\partial t}(t_0,f) = \mathbf{E}(t_0,f) + \frac{hT}{2}\mathbf{R}(t_0,h,f). \tag{8}$$

■

V. Conclusion

Acknowledgment

This work is partially supported by NSF awards 1633259 and 1633206, Leon Levy Foundation, and a Google faculty award.