Renormaliser le spectrogramme : pourquoi ? comment ?

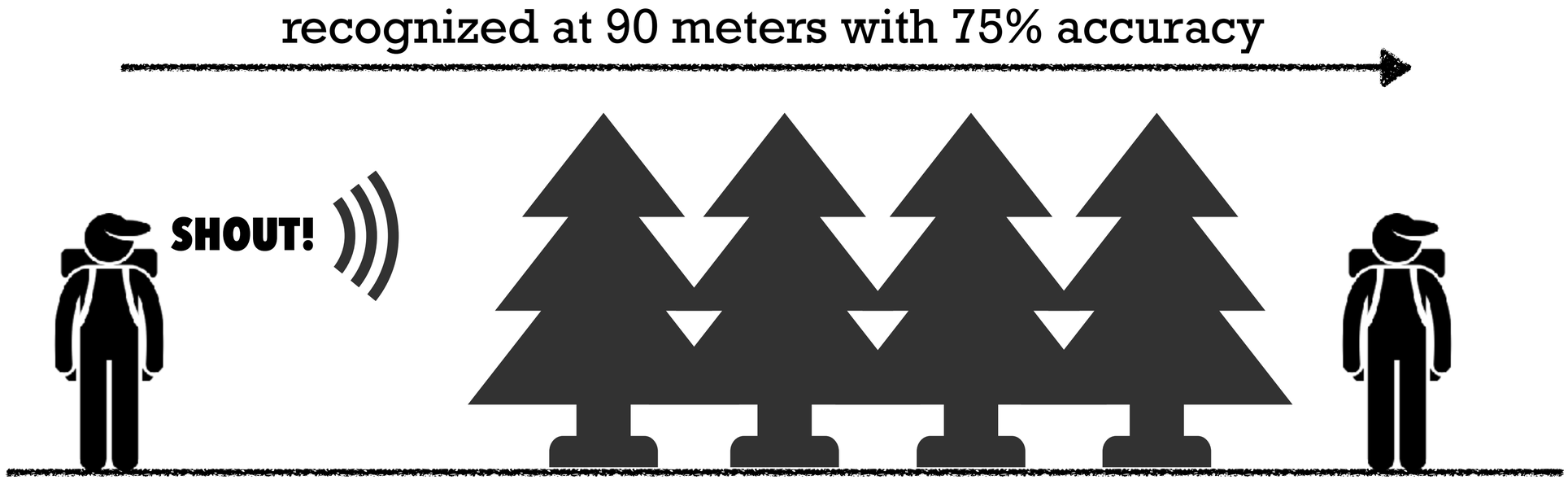
Vincent Lostanlen

LS2N, Centrale Nantes, CNRS

Paris, FR, March 2022. My research is funded by the French government.

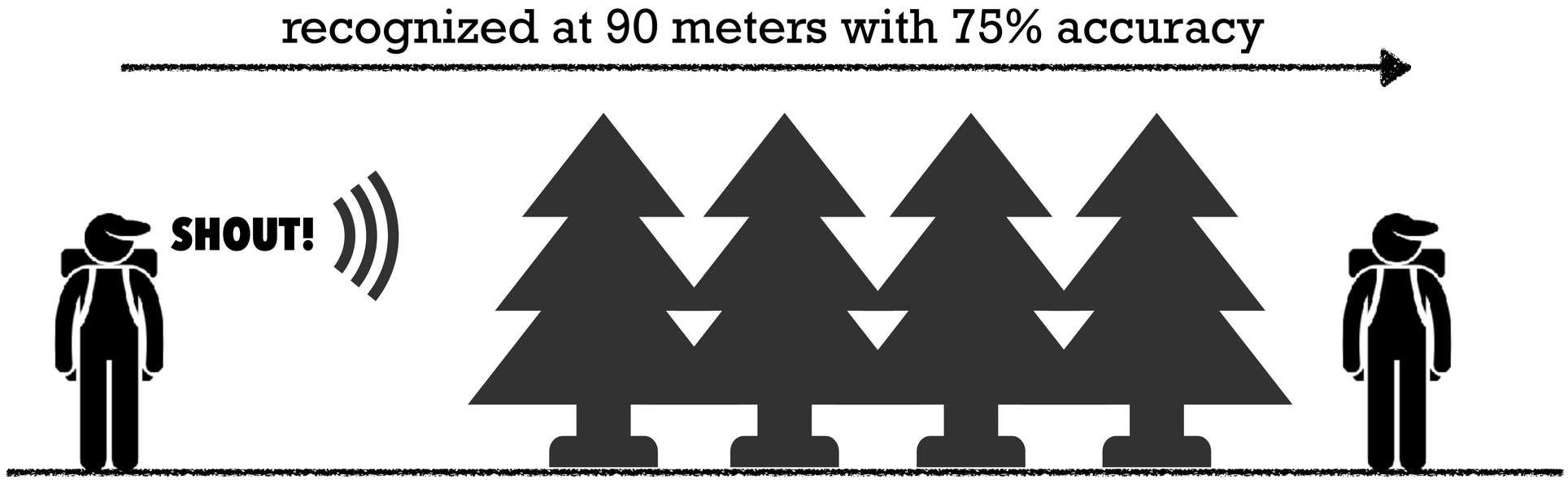
Hearing at a distance

[Meyer et al. 2018]



Hearing at a distance

[Meyer et al. 2018]



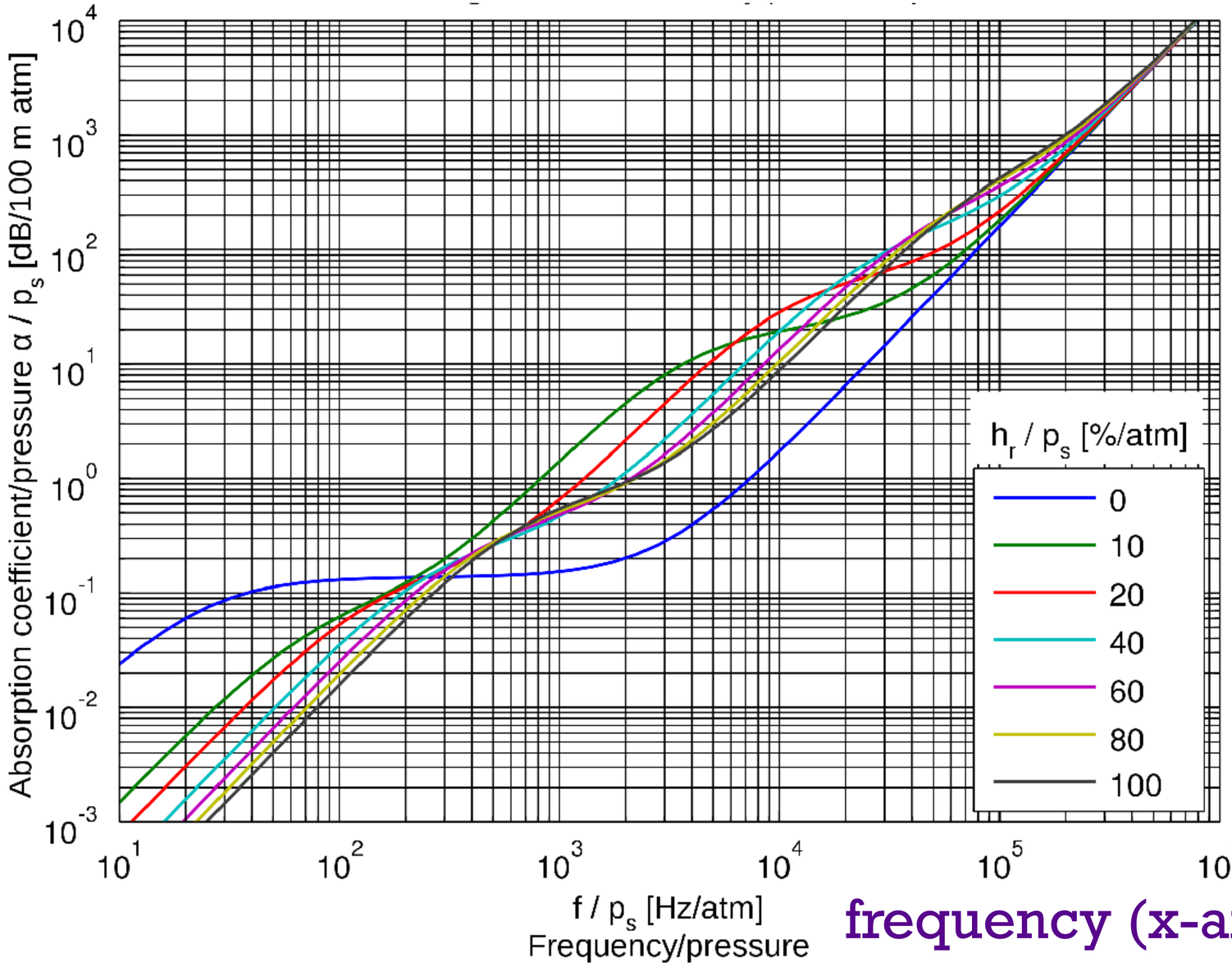
Yet, long-distance recognition is difficult for machines.

Why?

Atmospheric absorption

[Bass et al. 1984]

absorption (y-axis)

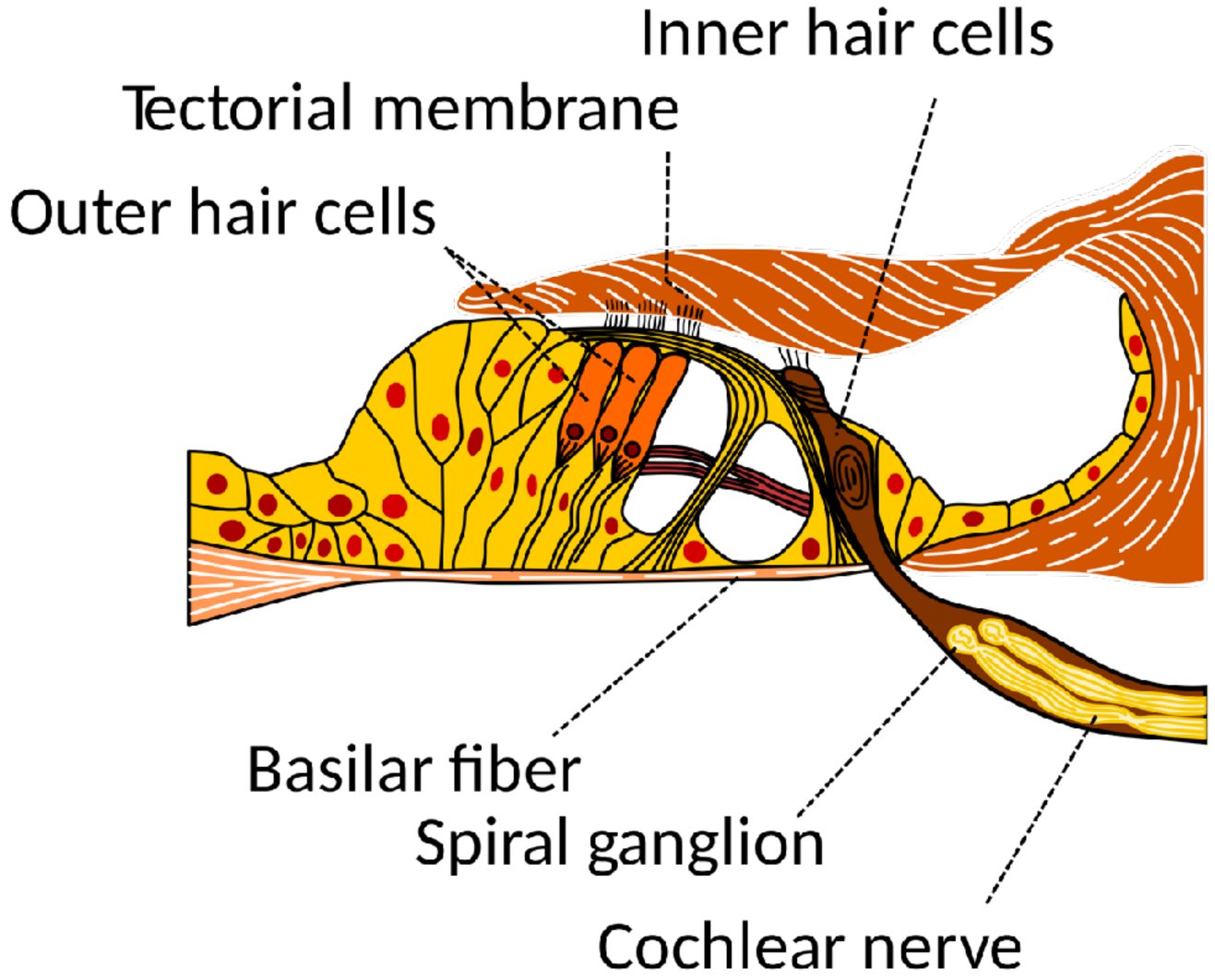


hygrometry (color)

frequency (x-axis)

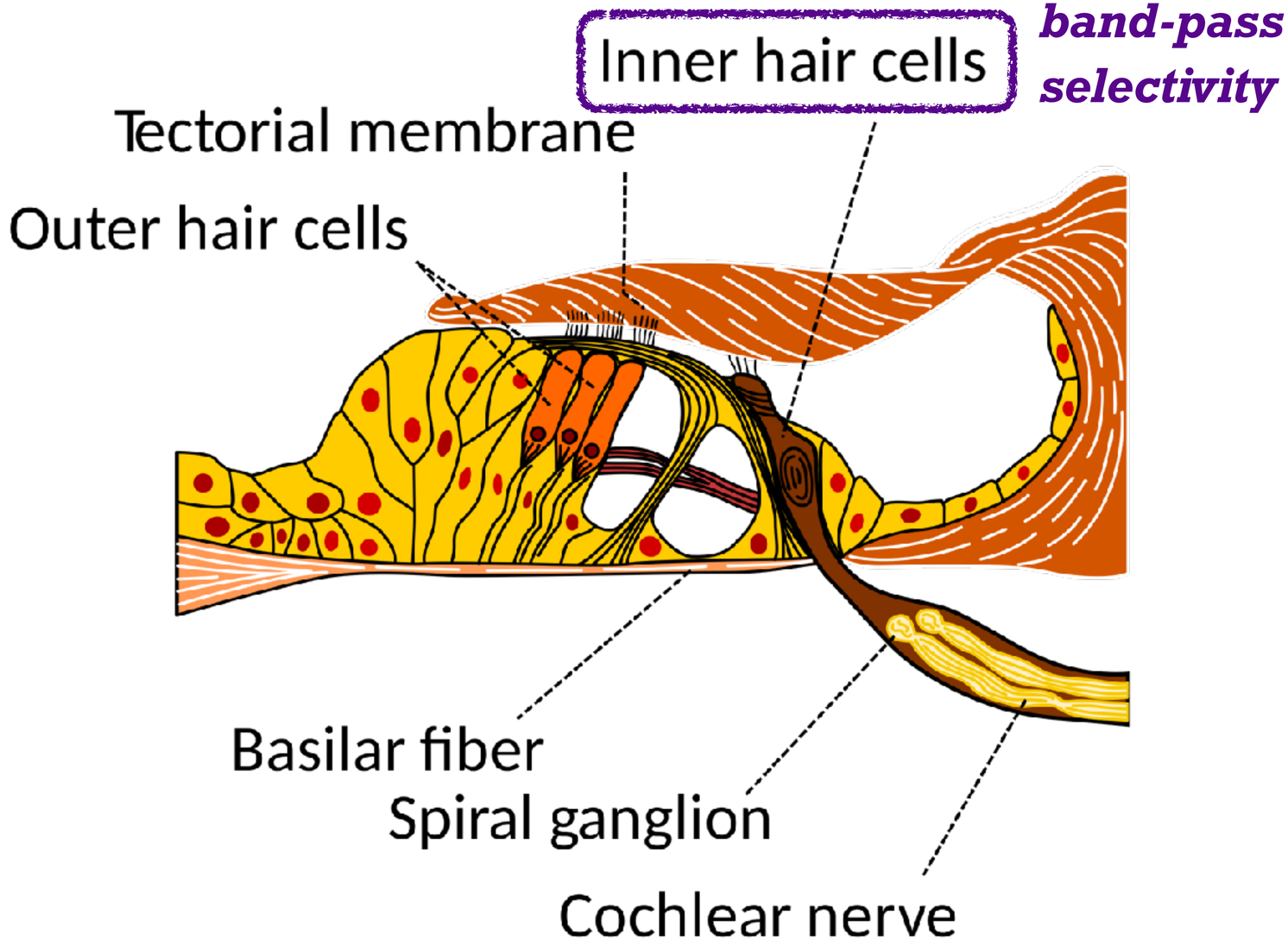
Auditory neurophysiology 101

[Mann and Kelley 2011, Robles et al. 2001]



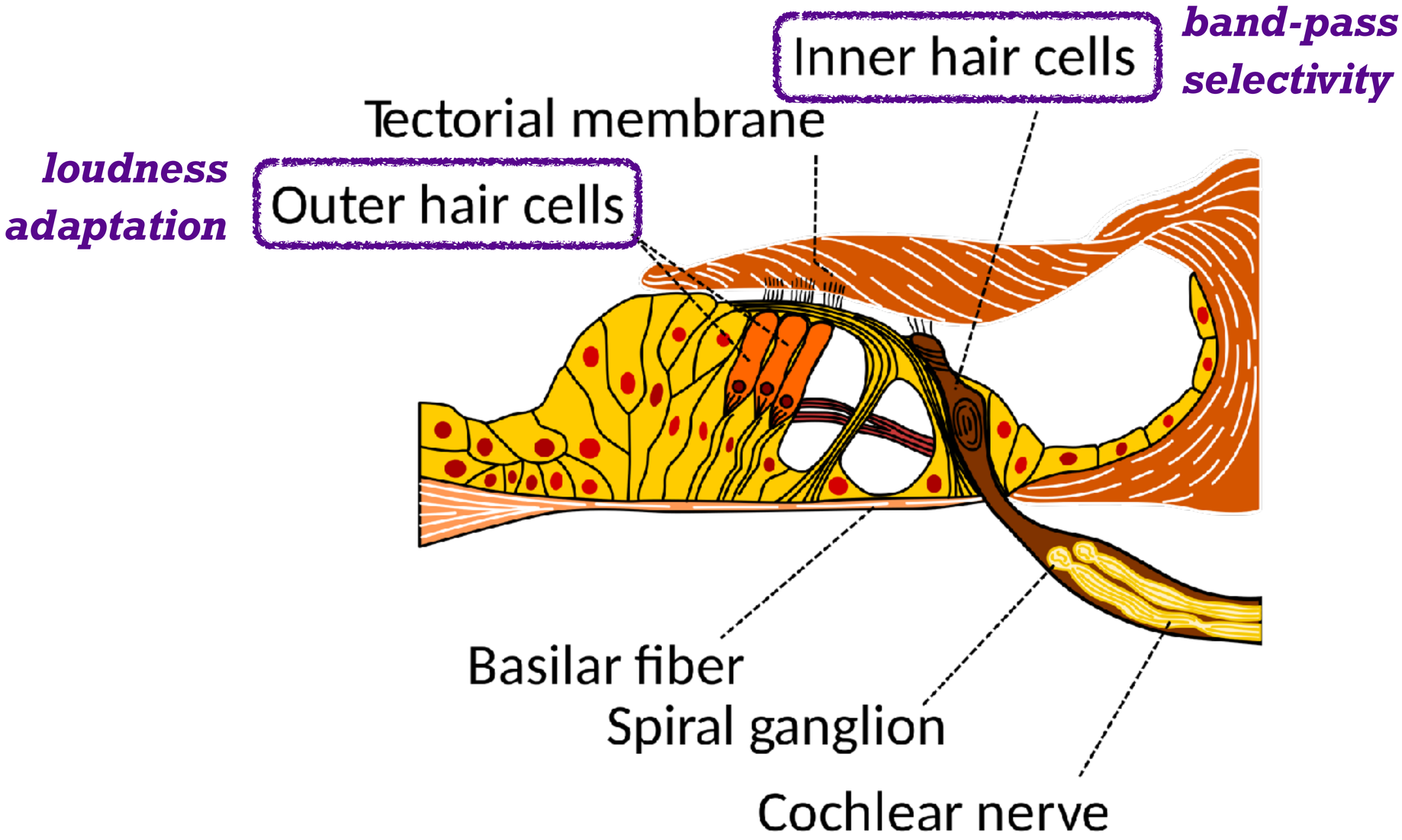
Auditory neurophysiology 101

[Mann and Kelley 2011, Robles et al. 2001]

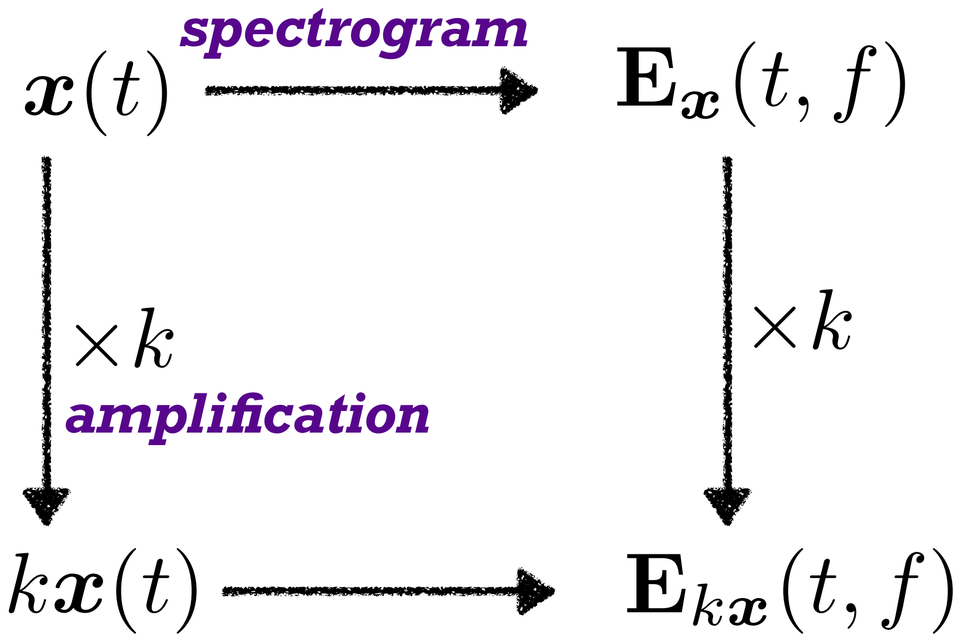


Auditory neurophysiology 101

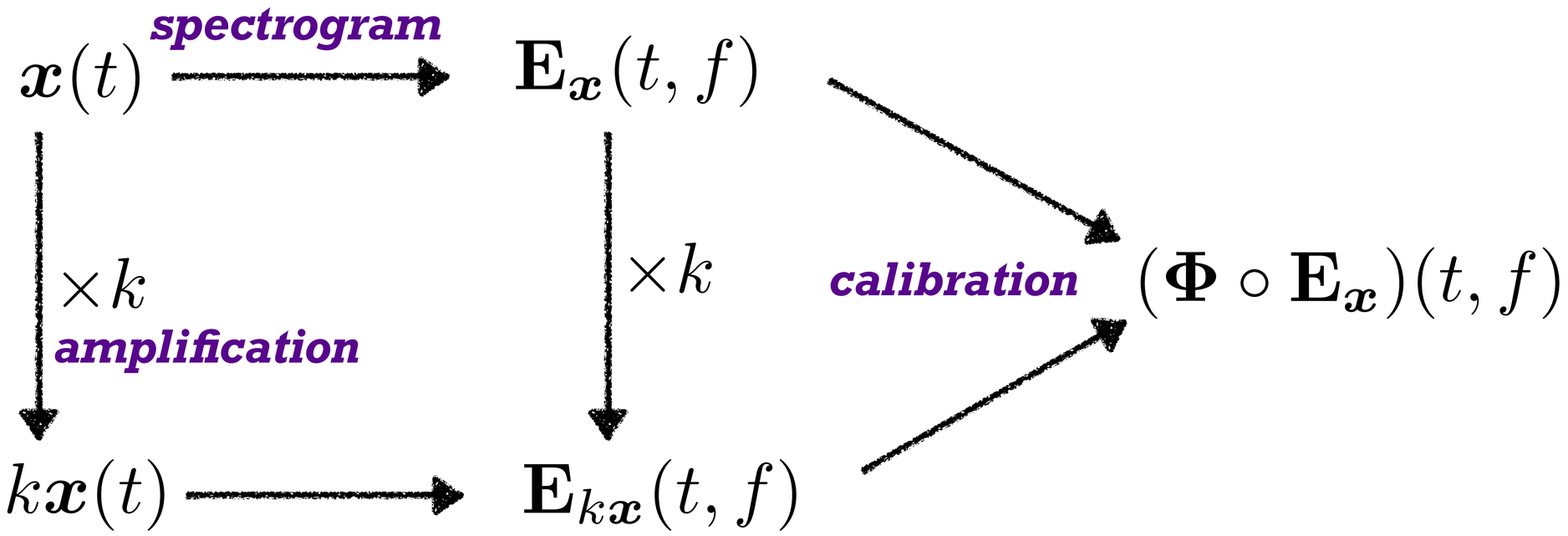
[Mann and Kelley 2011, Robles et al. 2001]



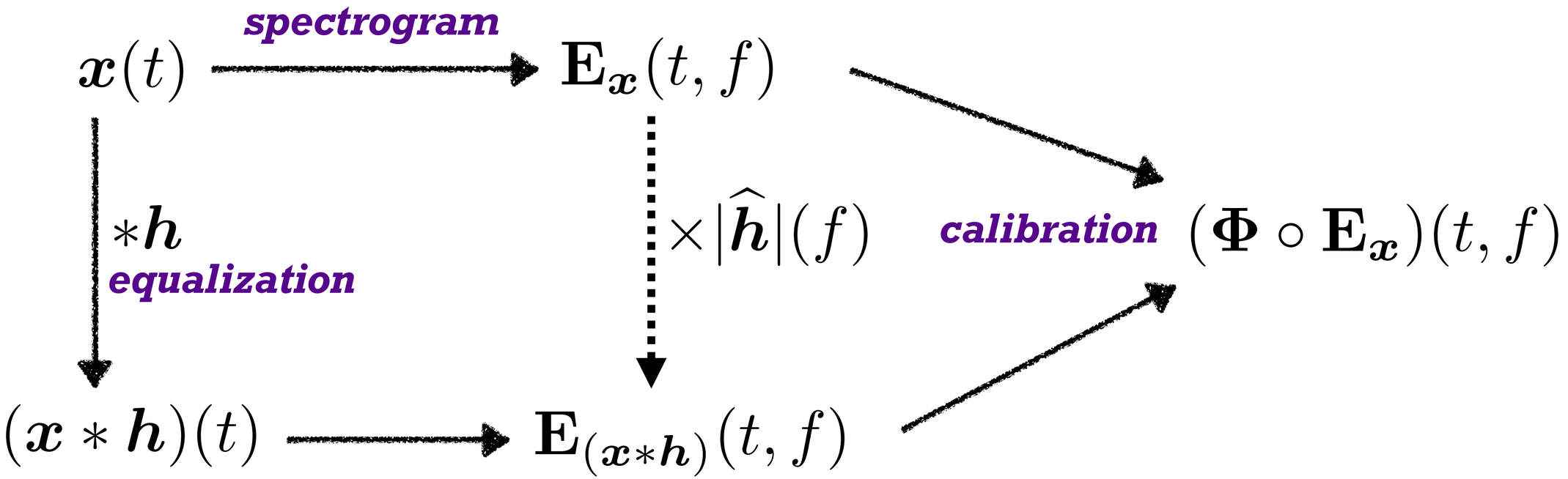
Invariance to amplitude



Invariance to amplitude



Invariance to equalization



Per-Channel Energy Normalization

[Wang et al. 2017, Lostanlen et al. 2019]

$$(\Phi \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \left(\delta + \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\left(\varepsilon + (\mathbf{E}_{\boldsymbol{x}} \overset{t}{*} \boldsymbol{\phi}_T)(t, f)\right)^{\alpha}} \right)^{r} - \delta^{r}$$

Per-Channel Energy Normalization

[Wang et al. 2017, Lostanlen et al. 2019]

$$(\Phi \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \left(\delta + \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\left(\varepsilon + (\mathbf{E}_{\boldsymbol{x}} \overset{t}{*} \boldsymbol{\phi}_T)(t, f)\right)^{\alpha}} \right)^{r} - \delta^r$$

hyperparameters
(predefined or learned)

Per-Channel Energy Normalization

[Wang et al. 2017, Lostanlen et al. 2019]

$$(\Phi \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \left(\delta + \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\left(\varepsilon + (\mathbf{E}_{\boldsymbol{x}} \overset{t}{*} \boldsymbol{\phi}_T)(t, f)\right)^{\alpha}} \right)^{r} - \delta^{r}$$

adaptive gain control

dynamic range compression

Per-Channel Energy Normalization

[Wang et al. 2017, Lostanlen et al. 2019]

$$\phi_T(t) = \exp(-t/T)\mathbf{1}(t > 0)$$

$$(\Phi \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \left(\delta + \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\left(\varepsilon + (\mathbf{E}_{\boldsymbol{x}} \overset{t}{*} \phi_T)(t, f)\right)^{\alpha}} \right)^{r} - \delta^{r}$$

low-pass filter

adaptive gain control

dynamic range compression

Typically:

- **r ~ 0.33**
- **delta ~ 0.5**
- **epsilon ~ 0**
- **alpha ~ 1**
- **T ~ 400 ms**

Simplified PCEN

Under the simplest parametrization, PCEN boils down to

$$(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\mathbf{M}_{\boldsymbol{x}}(t, f)}$$

where $\mathbf{M}_{\boldsymbol{x}}(t, f)$ is a moving average of $\mathbf{E}_{\boldsymbol{x}}(t, f)$.

Simplified PCEN

Under the simplest parametrization, PCEN boils down to

$$(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\mathbf{M}_{\boldsymbol{x}}(t, f)}$$

where $\mathbf{M}_{\boldsymbol{x}}(t, f)$ is a moving average of $\mathbf{E}_{\boldsymbol{x}}(t, f)$.

- If $\mathbf{E}_{\boldsymbol{x}}(t, f)$ is a **"noise"** region: $\mathbf{E}_{\boldsymbol{x}}(t, f) \approx \mathbf{M}_{\boldsymbol{x}}(t, f)$

  and thus $(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) \approx 1$ on average

Simplified PCEN

Under the simplest parametrization, PCEN boils down to

$$(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\mathbf{M}_{\boldsymbol{x}}(t, f)}$$

where $\mathbf{M}_{\boldsymbol{x}}(t, f)$ is a moving average of $\mathbf{E}_{\boldsymbol{x}}(t, f)$.

- If $\mathbf{E}_{\boldsymbol{x}}(t, f)$ is a **"noise"** region: $\mathbf{E}_{\boldsymbol{x}}(t, f) \approx \mathbf{M}_{\boldsymbol{x}}(t, f)$

  and thus $(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) \approx 1$

- If $\mathbf{E}_{\boldsymbol{x}}(t, f)$ is a **"transient"** region: $\mathbf{E}_{\boldsymbol{x}}(t, f) \gg \mathbf{M}_{\boldsymbol{x}}(t, f)$

  and thus $(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) \gg 1$

Simplified PCEN

Under the simplest parametrization, PCEN boils down to

$$(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) = \frac{\mathbf{E}_{\boldsymbol{x}}(t, f)}{\mathbf{M}_{\boldsymbol{x}}(t, f)}$$

where $\mathbf{M}_{\boldsymbol{x}}(t, f)$ is a moving average of $\mathbf{E}_{\boldsymbol{x}}(t, f)$.

- If $\mathbf{E}_{\boldsymbol{x}}(t, f)$ is a **"noise"** region: $\mathbf{E}_{\boldsymbol{x}}(t, f) \approx \mathbf{M}_{\boldsymbol{x}}(t, f)$

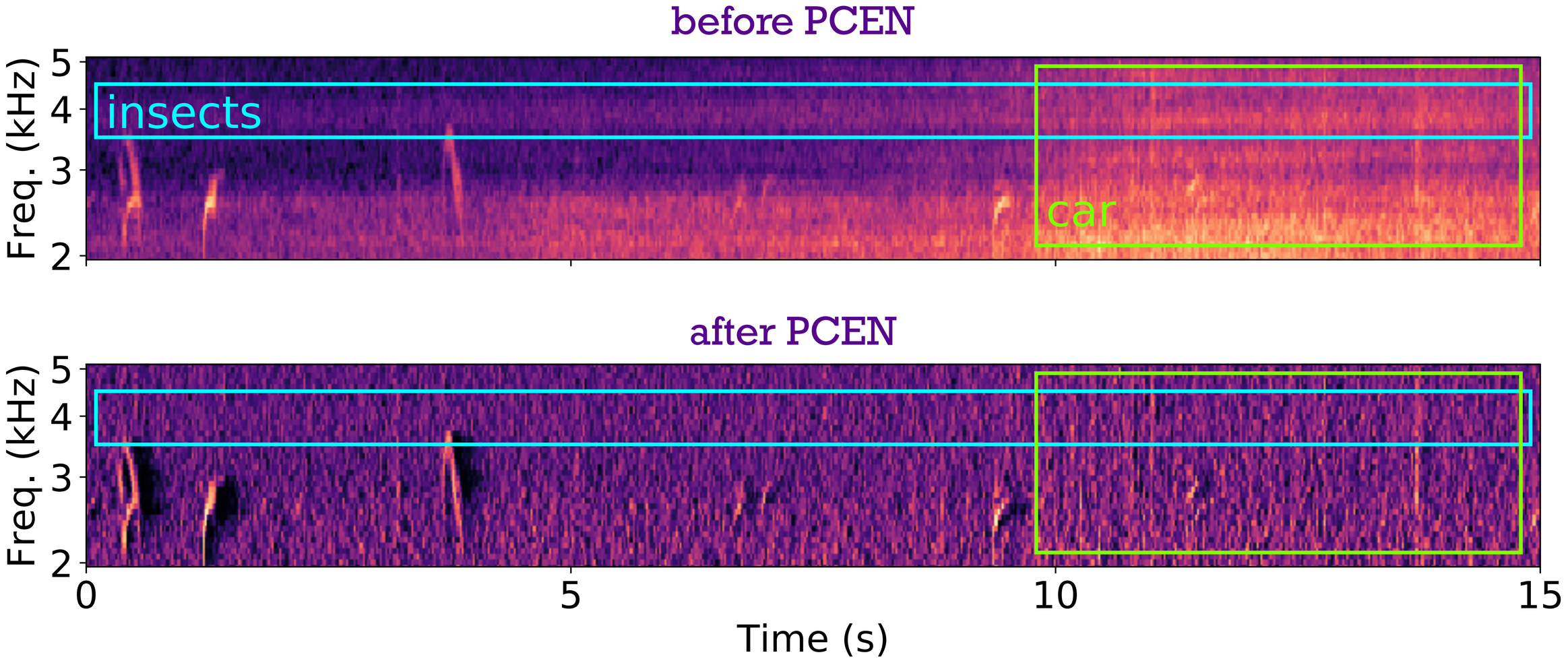
  and thus $(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) \approx 1$

- If $\mathbf{E}_{\boldsymbol{x}}(t, f)$ is a **"transient"** region: $\mathbf{E}_{\boldsymbol{x}}(t, f) \gg \mathbf{M}_{\boldsymbol{x}}(t, f)$

  and thus $(\mathbf{\Phi} \circ \mathbf{E}_{\boldsymbol{x}})(t, f) \gg 1$

PCEN in practice

PCEN enhances the visual contrast between events and background.

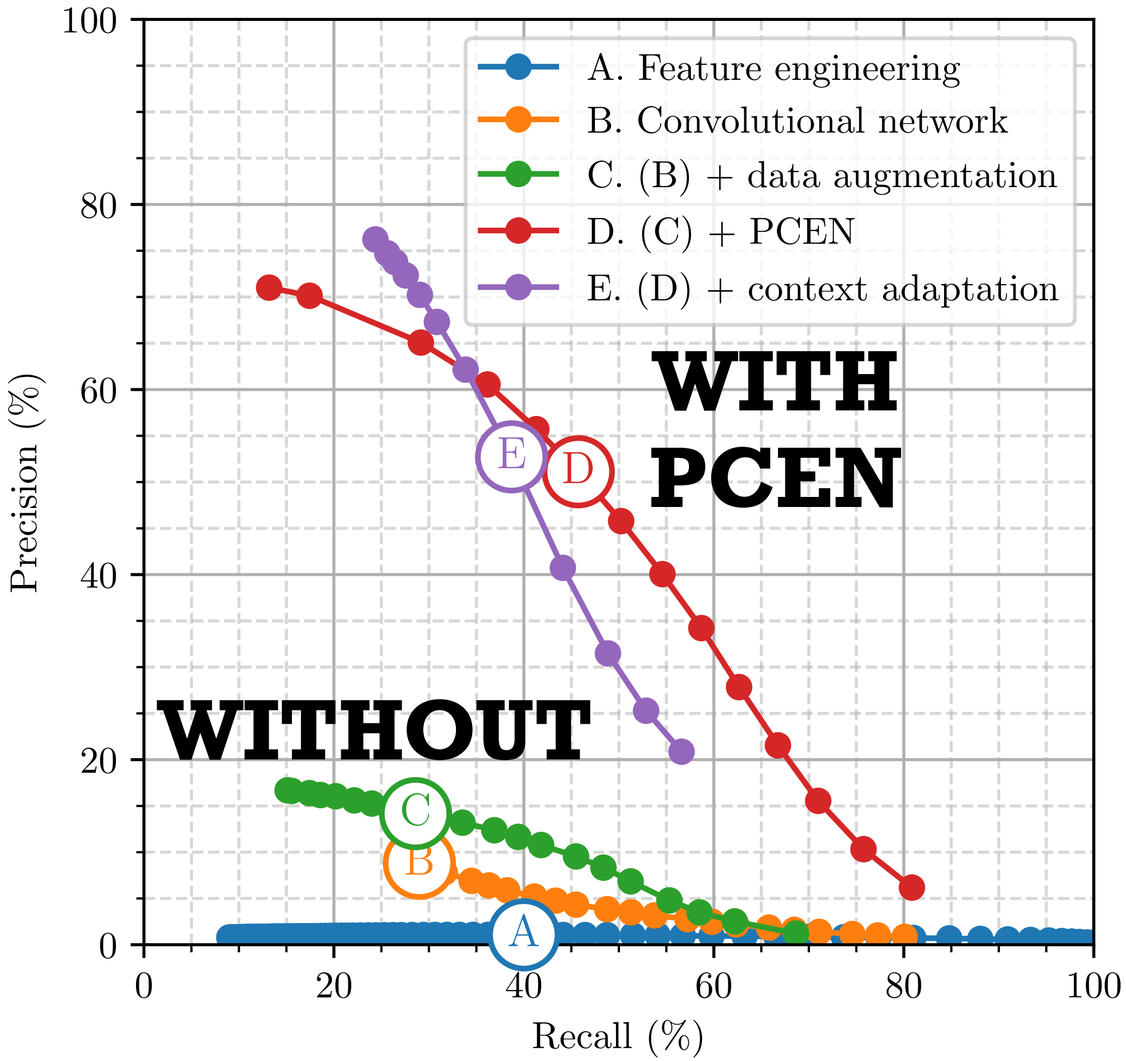


Data: BirdVox project

PCEN + deep learning

[Lostanlen et al. 2022]

Evaluated on 300h hours of annotated audio from a sensor network.



PCEN plays a crucial role in statistical generalization.

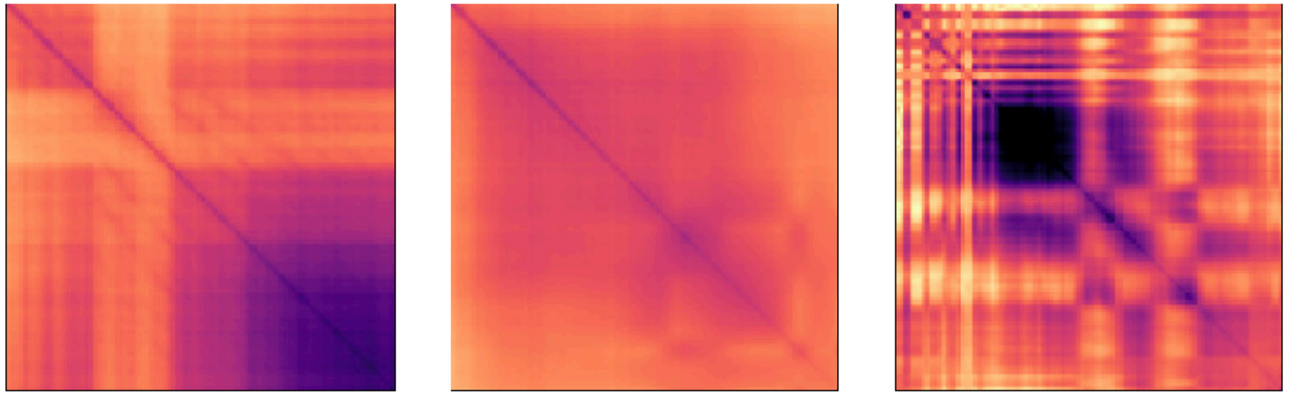
Models A, B, C: without PCEN
Models D, E: with PCEN

how to justify this?

PCEN decorrelates subbands

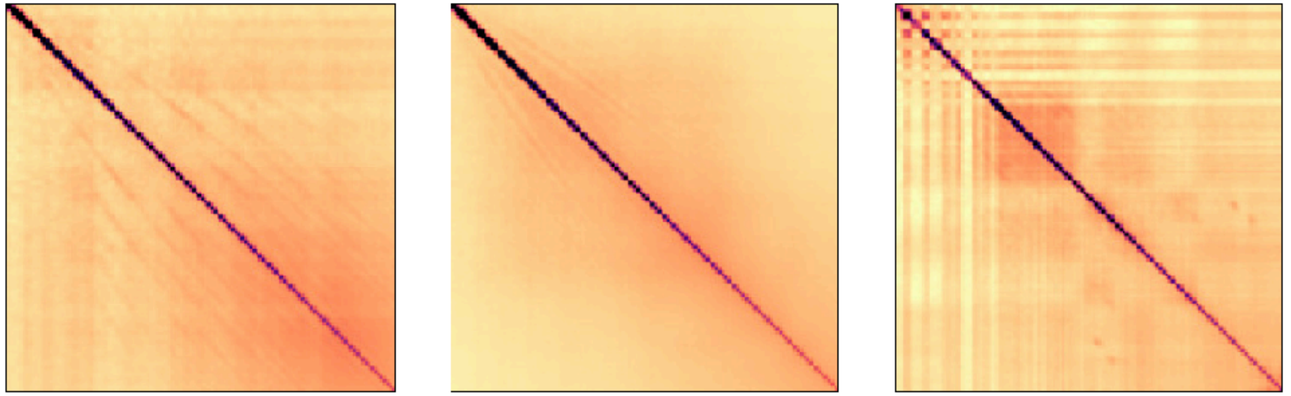
[Lostanlen et al. 2019a]

SONYC | DCASE 2013 SC | BirdVox



(a) Logarithmic transformation.

(urban noise) | *(indoor noise)* | *(rural noise)*

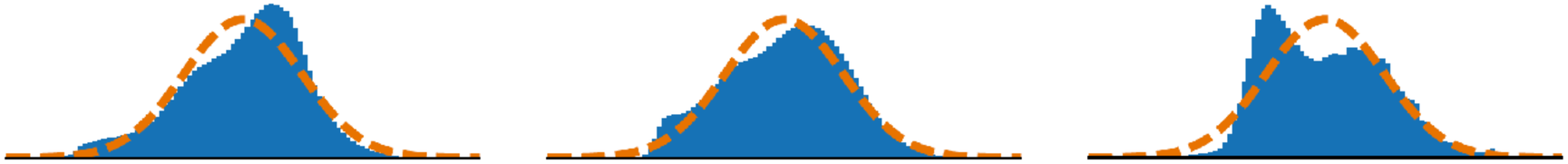


(b) Per-channel energy normalization (PCEN).

PCEN Gaussianizes magnitudes

[Lostanlen et al. 2019a]

(urban noise) SONYC | *(indoor noise)* DCASE 2013 SC | *(rural noise)* BirdVox

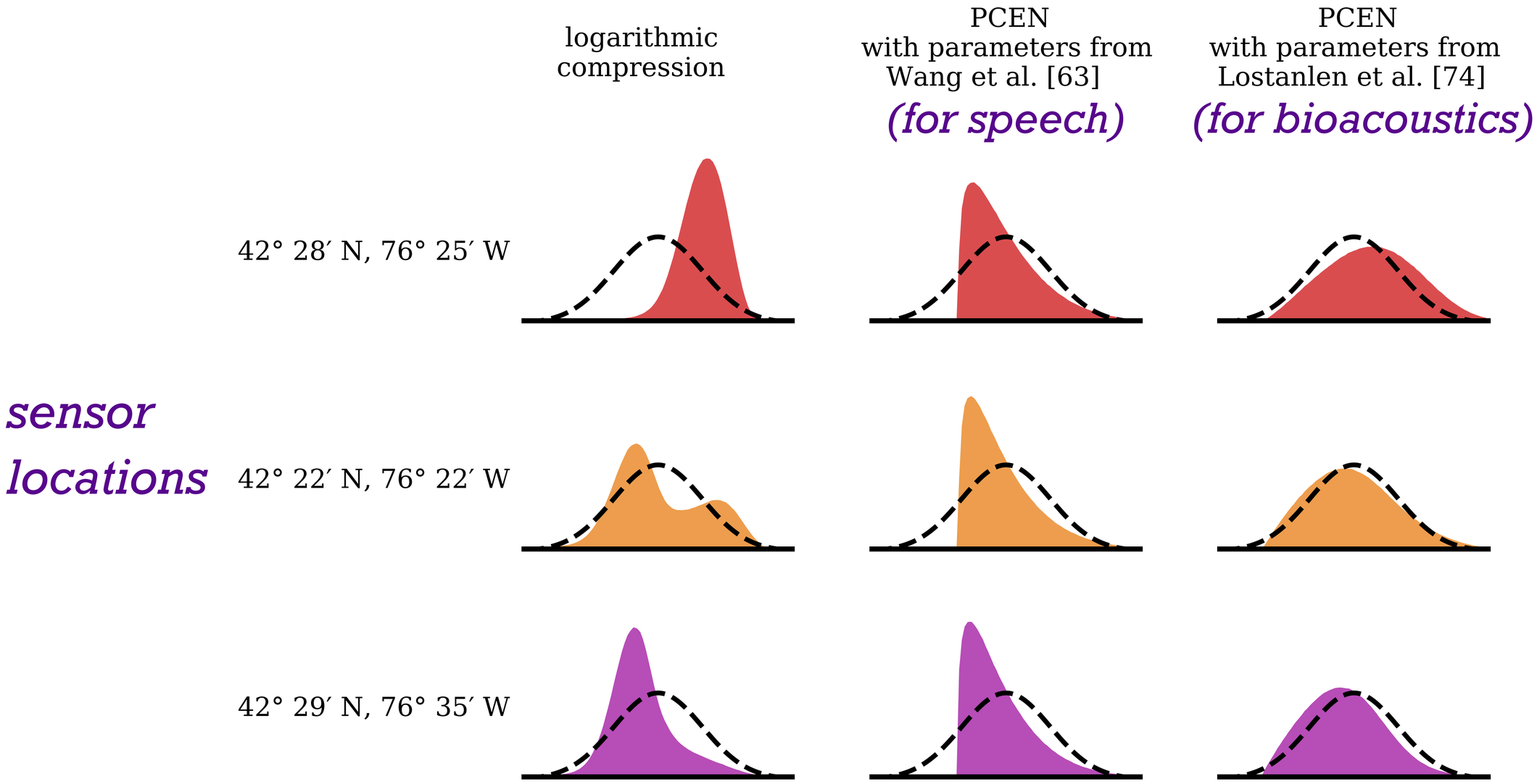


(a) Logarithmic transformation.

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

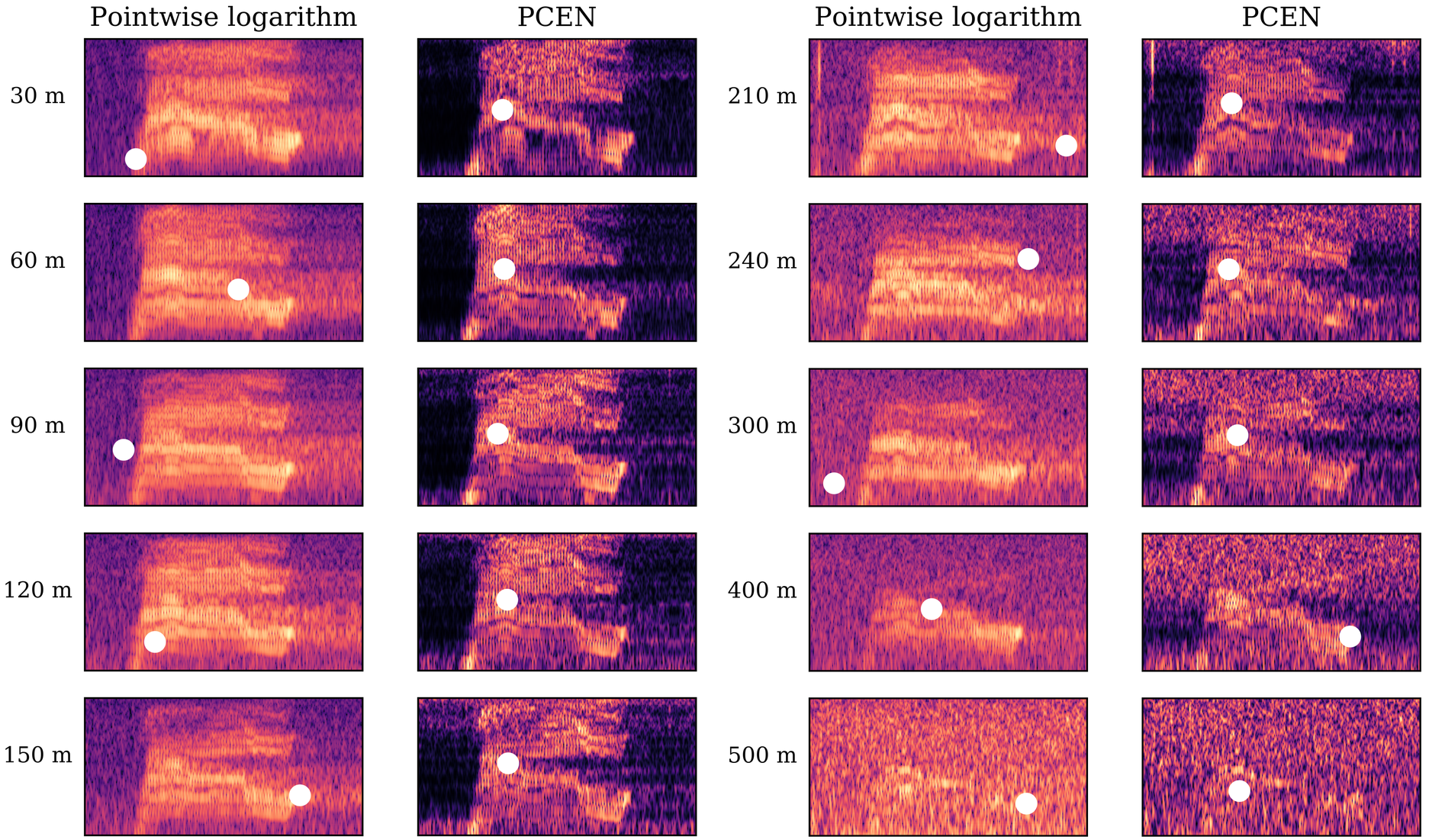
PCEN generalizes (if well-tuned)

[Lostanlen et al. 2019b]



PCEN improves detection range

[Lostanlen et al. 2019c. Audio data from Elly Knight]



End-to-end PCEN learning

[Wang et al. 2017, Zinemanas et al. 2018, Zeghidour et al. 2021, many others]

LEAF (Learnable Audio Frontend) from Google Research.

State-of-the-art performance on many tasks:
speech commands, **birdsong** detection, **musical** instruments …

Table 2: Impact of the compression layer on the performance (single task, average accuracy in %).

	mel-filterbank			LEAF- filterbank		
	log	PCEN	sPCEN	log	PCEN	sPCEN
Average	73.9 ± 0.8	76.4 ± 0.8	76.0 ± 0.7	74.6 ± 0.7	76.4 ± 0.8	$\mathbf{76.9} \pm 0.8$

Table from Zeghidour et al. 2021

Conferences > 2019 IEEE Workshop on Applica...

Tricycle: Audio Representation Learning from Sensor Network Data Using Self-Supervision

Publisher: IEEE | Cite This | PDF

Mark Cartwright ; Jason Cramer ; Justin Salamon ; Juan Pablo Bello All Authors

Conferences > ICASSP 2021 - 2021 IEEE Inter...

Sound Event Detection in Urban Audio with Single and Multi-Rate Pcen

Publisher: IEEE | Cite This | PDF

Christopher Ick ; Brian McFee All Authors

Conferences > 2019 IEEE International Smart...

Activating accessible pedestrian signals by voice using keyword spotting systems

Publisher: IEEE | Cite This | PDF

Mirzodaler Muhsinzoda ; Carlos Cruz Corona ; David A. Pelta ; Jose Luis Verdegay All Authors

Conferences > ICASSP 2020 - 2020 IEEE Inter...

Audio Feature Extraction for Vehicle Engine Noise Classification

Publisher: IEEE | Cite This | PDF

Luca Becker ; Alexandru Nelus ; Johannes Gauer ; Lars Rudolph ; Rainer Martin All Authors

[Submitted on 7 Oct 2021]

Improving Bird Classification with Unsupervised Sound Separation

Tom Denton, Scott Wisdom, John R. Hershey

Conferences > 2019 24th Conference of Open ...

End-to-end Convolutional Neural Networks for Sound Event Detection in Urban Environments

Publisher: IEEE | Cite This | PDF

Pablo Zinemanas ; Pablo Cancela ; Martín Rocamora All Authors

[Submitted on 24 Sep 2021]

Parameterized Channel Normalization for Far-field Deep Speaker Verification

Xuechen Liu, Md Sahidullah, Tomi Kinnunen

Conferences > 2019 IEEE Workshop on Applica...

Tricycle: Audio Representation Learning from Sensor Network Data Using Self-Supervision

Publisher: IEEE Cite This PDF

Mark Car ... right ; Jason Cramer ; Justin Salamon ; Juan Pablo Bello All Authors

Conferences > ICASSP 2021 - 2021 IEEE Inter...

Sound Event Detection in Urban Audio with Single and Multi-Rate Pcen

Publisher: IEEE Cite This PDF

Christopher Ick ; Brian McFee All Authors

conda install -c conda-forge librosa

Conferenc ... > 2019 IEEE International Smart ...

Activating accessible pedestrian signals by voice using keyword spotting systems

Publisher: IEEE Cite This PDF

Mirzodaler Muhsinzoda ; Carlos Cruz Corona ; David A. Pelta ; Jose Luis Verdegay All Authors

Conferences > ICASSP 2020 - 2020 IEEE Inter...

Audio Feature Extraction for Vehicle Engine Noise Classification

Publisher: IEEE Cite This PDF

Luca Becker ; Alexandru Nelus ; Johannes Gauer ; Lars Rudolph ; Rainer Martin All Authors

[Submitted on 7 Oct 2021]

Improving Bird Classification with Unsupervised Sound Separation

Tom Denton, Scott Wisdom, John R. Hershey

Conferences > 2019 24th Conference of Open ...

End-to-end Convolutional Neural Networks for Sound Event Detection in Urban Environments

Publisher: IEEE Cite This PDF

Pablo Zinemanas ; Pablo Cancela ; Martín Rocamora All Authors

[Submitted on 24 Sep 2021]

Parameterized Channel Normalization for Far-field Deep Speaker Verification

Xuechen Liu, Md Sahidullah, Tomi Kinnunen