On the Robustness of Musical Timbre Perception Models: From Perceptual to Learned Approaches

Barbara Pascal, Mathieu Lagrange

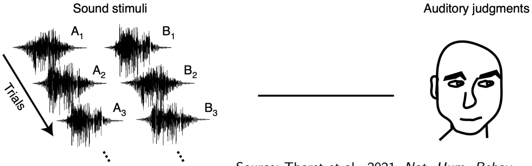
August 27, 2024

Session: Music Information Retrieval, Analysis and Processing

EUSIPCO 2024, Lyon, France

At the frontier of **digital audio processing** & **psychoacoustic**:

How humans make judgments about their environment based on sounds?

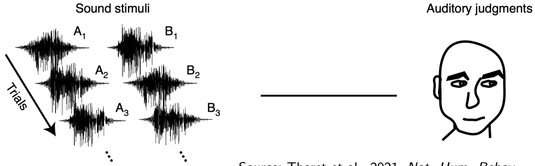


Source: Thoret et al., 2021, *Nat. Hum. Behav.*

Revealing acoustic substrate of human timbre perception

At the frontier of **digital audio processing** & **psychoacoustic**:

How humans make judgments about their environment based on sounds?



Source: Thoret et al., 2021, *Nat. Hum. Behav.*

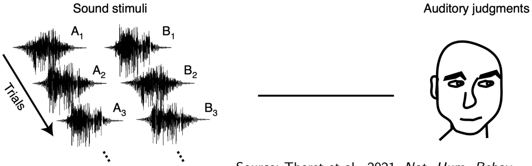
Focus on **timbre**, the "color" of a sound

- perceived sound quality
- emerging from intricate bundle of acoustic cues
- informs about the sound sources and production mechanisms

Revealing acoustic substrate of human timbre perception

At the frontier of **digital audio processing** & **psychoacoustic**:

How humans make judgments about their environment based on sounds?



Source: Thoret et al., 2021, *Nat. Hum. Behav.*

Focus on **timbre**, the "color" of a sound

- perceived sound quality
- emerging from intricate bundle of acoustic cues
- informs about the sound sources and production mechanisms

Yamaha

Important example: the timbre of a musical instrument

▷ modeling of timbre perception remains a burning topic in cognitive neuroscience

Psychoacoustic experiments and resulting datasets

Audio samples $\{a_1, \ldots, a_\ell\}$, ℓ: number of sounds

- recorded and edited natural instruments sounds
- sounds resynthesized with simplifications or systematic modifications
- simulated and hybrid sounds imitating musical instruments

Audio samples $\{a_1, \ldots, a_\ell\}$, ℓ: number of sounds

- recorded and edited natural instruments sounds
- sounds resynthesized with simplifications or systematic modifications
- simulated and hybrid sounds imitating musical instruments

Dissimilarity ratings stored in a vector $\mathbf{s} \in [0,1]^{\ell(\ell-1)/2}$

$$\text{pair of sounds } (a_i, a_j), \quad \text{rating } \mathsf{s}_{\{i,j\}} \in [0,1]$$

- $\mathsf{s}_{\{i,j\}} = 0$: a_i, a_j exactly similar audio samples
- $\mathsf{s}_{\{i,j\}} = 1$: a_i, a_j maximally different audio samples

Ratings are averaged over all participants.

Psychoacoustic experiments and resulting datasets

Audio samples $\{a_1, \ldots, a_\ell\}$, ℓ: number of sounds

- recorded and edited natural instruments sounds
- sounds resynthesized with simplifications or systematic modifications
- simulated and hybrid sounds imitating musical instruments

Dissimilarity ratings stored in a vector $\mathbf{s} \in [0,1]^{\ell(\ell-1)/2}$

$$\text{pair of sounds } (a_i, a_j), \quad \text{rating } s_{\{i,j\}} \in [0,1]$$

- $s_{\{i,j\}} = 0$: a_i, a_j exactly similar audio samples
- $s_{\{i,j\}} = 1$: a_i, a_j maximally different audio samples

Ratings are averaged over all participants.

Datasets from 17 published studies between 1977 and 2016

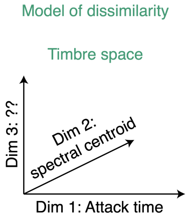
- from $\ell_{\min} = 11$ to $\ell_{\max} = 20$
- diversity of sounds: natural, resynthesized, simulated
- 9 to 34 subjects, from naive listeners to confirmed musicians

From Thoret et al., 2021, *Nat. Hum. Behav.*, github.com/EtienneTho/musical-timbre-studies

Multidimensional Scaling (MDS)

1. collect dissimilarity ratings
2. represent audio samples in a low dimensional space
3. so that distances reflect dissimilarities
4. correlate latent dimensions with acoustic descriptors

$\Longrightarrow$ broad understanding of timbre acoustic correlates

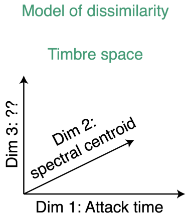


Source: Thoret et al., 2021, *Nat. Hum. Behav.*

Multidimensional Scaling (MDS)

1. collect dissimilarity ratings
2. represent audio samples in a low dimensional space
3. so that distances reflect dissimilarities
4. correlate latent dimensions with acoustic descriptors

$\Longrightarrow$ broad understanding of timbre acoustic correlates



Source: Thoret et al., 2021, *Nat. Hum. Behav.*

Limitations of Multidimensional Scaling (MDS)

- arbitrary choices and ad-hoc parameter tuning impair replicability
- many psychophysic acoustic descriptors: only two correlate with MDS dimensions
- only partial explanation due to low descriptive power of these descriptors

Need alternatives to unveil the intricate mechanisms behind timbre perception

Human dissimilarity ratings based on

- complex perceptual judgments **very hard to model fully**
- intricate high-level audio characteristics

Revealing acoustic substrate of human timbre perception

Human dissimilarity ratings based on

- complex perceptual judgments **very hard to model fully**
- intricate high-level audio characteristics

▷ **Idea:** learn the salient features used by humans to discriminate different timbres
(Thoret et al., 2021, *Nat. Hum. Behav.*)

Revealing acoustic substrate of human timbre perception

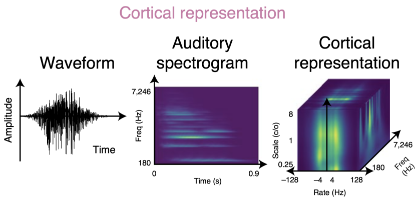
Human dissimilarity ratings based on

- complex perceptual judgments **very hard to model fully**
- intricate high-level audio characteristics

▷ **Idea:** learn the salient features used by humans to discriminate different timbres (Thoret et al., 2021, *Nat. Hum. Behav.*)

Models of the primary auditory cortex: SpectroTemporal Modulations

- auditory spectrum: cochlea representation
  128 *constant-Q asymmetric bandpass filters on log-frequency scale*
- cortical representation: STMF representation
  2D-Fourier of auditory spectrogram with 11 *cycles per octave and* 22 *frequencies*



▷ **metric learning** to extract features relevant from a perceptual point of view

Metric learning framework: design a distance d such that $d(a_i, a_j) \sim s_{i,j}$

Metric learning framework: design a distance d such that $\mathrm{d}(a_i, a_j) \sim \mathrm{s}_{i,j}$

- parametric distance in the space of the representation Ψ (e.g., cochlea, STMF)

$$\mathrm{d}_{\mathbf{w}}^{\Psi}(a_i, a_j)^2 = \sum_{k=1}^{n_\Psi} \frac{1}{\mathrm{w}_k^2} \left(\Psi(a_i)_k - \Psi(a_j)_k\right)^2$$

Metric learning framework: design a distance d such that $\mathrm{d}(a_i, a_j) \sim \mathrm{s}_{i,j}$

- parametric distance in the space of the representation Ψ (e.g., cochlea, STMF)

$$\mathrm{d}_{\mathbf{w}}^{\Psi}(a_i, a_j)^2 = \sum_{k=1}^{n_\Psi} \frac{1}{\mathrm{w}_k^2} \left(\Psi(a_i)_k - \Psi(a_j)_k\right)^2$$

- learn weights by maximizing the reward function

$$\mathbf{w}_\star \in \underset{\mathbf{w} \in \mathbb{R}^{n_\Psi}}{\mathrm{Argmax}}\, \mathcal{P}(\mathrm{d}_{\mathbf{w}}^{\Psi}, \mathbf{s})$$

Pearson correlation (invariant to mean shifts and variance rescalings)

from $\mathcal{P} = -1$: perfect anti-correlation, to $\mathcal{P} = 1$: perfect correlation

▷ the **larger** $\mathcal{P}(\mathrm{d}_{\mathbf{w}_\star}^{\Psi}, \mathbf{s})$ the **better** the fit

Revealing acoustic substrate of human timbre perception

Metric learning framework: design a distance d such that $\mathrm{d}(a_i, a_j) \sim \mathrm{s}_{i,j}$

- parametric distance in the space of the representation Ψ (e.g., cochlea, STMF)

$$\mathrm{d}^{\Psi}_{\mathbf{w}}(a_i, a_j)^2 = \sum_{k=1}^{n_\Psi} \frac{1}{\mathrm{w}_k^2} \left(\Psi(a_i)_k - \Psi(a_j)_k\right)^2$$

- learn weights by maximizing the reward function

$$\mathbf{w}_\star \in \operatorname*{Argmax}_{\mathbf{w} \in \mathbb{R}^{n_\Psi}} \mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}}, \mathbf{s})$$

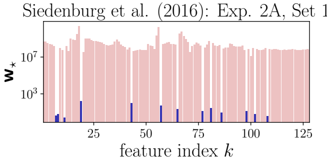
Pearson correlation (invariant to mean shifts and variance rescalings)

from $\mathcal{P} = -1$: perfect anti-correlation, to $\mathcal{P} = 1$: perfect correlation

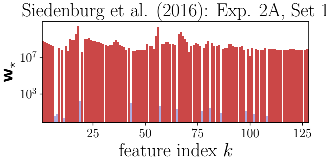
▷ the **larger** $\mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}_\star}, \mathbf{s})$ the **better** the fit

Illustration: for the auditory spectrum Ψ = cochlea representation

✓ **relevant features**



✗ **discarded features**



Metric learning algorithm: influence of initialization

Objective function $\mathbf{w} \mapsto \mathcal{P}(\mathrm{d}_{\mathbf{w}}^{\Psi}, \mathbf{s})$ twice differentiable: **quasi-Newton algorithm**

Limited memory Boyden-Fletcher-Golfarb-Shanno algorithm with box constraints

- descent-step free;
- optimization in large dimension $n_{\Psi} \gtrsim 10^4$;
- quadratic convergence in the neighborhood of local optima

Objective function $\mathbf{w} \mapsto \mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}}, \mathbf{s})$ twice differentiable: **quasi-Newton algorithm**

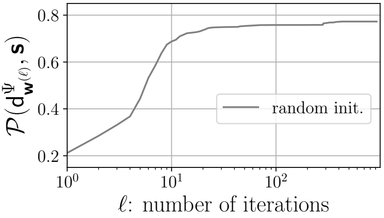
Limited memory Boyden-Fletcher-Golfarb-Shanno algorithm with box constraints

- descent-step free;
- optimization in large dimension $n_\Psi \gtrsim 10^4$;
- quadratic convergence in the neighborhood of local optima

Random initialization (Thoret et al., 2021, *Nat. Hum. Behav.*)

$$\mathbf{w}_k^{[0]} \sim \mathcal{N}(1, 10^{-4})$$

independent identically distributed



Metric learning algorithm: influence of initialization

Objective function $\mathbf{w} \mapsto \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}}, \mathbf{s})$ twice differentiable: **quasi-Newton algorithm**

Limited memory Boyden-Fletcher-Golfarb-Shanno algorithm with box constraints

- descent-step free;
- optimization in large dimension $n_\Psi \gtrsim 10^4$;
- quadratic convergence in the neighborhood of local optima

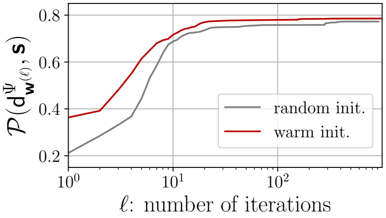
Random initialization (Thoret et al., 2021, *Nat. Hum. Behav.*)

$$\mathbf{w}_k^{[0]} \sim \mathcal{N}(1, 10^{-4})$$

independent identically distributed

Warm start

$$\mathbf{w}^{[0]} \in \operatorname*{Argmin}_{\mathbf{w} \in \mathbb{R}_+^{n_\Psi}} \sum_{\{i,j\}} \left| \mathsf{d}^{\Psi}_{\mathbf{w}}(a_i, a_j)^2 - \mathsf{s}_{\{i,j\}} \right|^2$$



Metric learning in representation spaces: explained variance

Performance criterion: $\mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}_\star}, \mathbf{s})^2 \in [0, 1]$ (Thoret et al., 2021, *Nat. Hum. Behav.*)

▷ squared Pearson correlation between **learned distance** and **dissimilarity ratings**

Metric learning in representation spaces: explained variance

Performance criterion: $\mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}_\star}, \mathbf{s})^2 \in [0, 1]$ (Thoret et al., 2021, *Nat. Hum. Behav.*)

▷ squared Pearson correlation between **learned distance** and **dissimilarity ratings**

17 datasets from studies between 1977 and 2016	**perceptual** cochlea $n_\Psi = 128$	STMF $n_\Psi = 30976$
Grey, 1977	0.48	0.84
Grey et al., 1978	0.11	0.33
Iverson et al., 1993: Whole	0.16	0.87
Iverson et al., 1993: Onset	0.07	0.22
Iverson et al., 1993: Remainder	0.03	0.27
McAdams et al., 1995	0.30	0.77
Lakatos et al., 2000: Harmonic	0.19	0.85
Lakatos et al., 2000: Percussive	0.18	0.27
Lakatos et al., 2000: Combined	0.13	0.33
Barthet et al., 2010	0.74	0.98
Patil et al., 2012: A3	0.62	0.97
Patil et al., 2012: DX4	0.66	0.99
Patil et al., 2012: GD4	0.46	0.95
Siedenburg et al., 2016: Exp 2A, Set 1	0.62	0.95
Siedenburg et al., 2016: Exp 2A, Set 2	0.73	0.99
Siedenburg et al., 2016: Exp 2A, Set 3	0.10	0.53
Siedenburg et al., 2016: Exp 2B, Set 3	0.07	0.46
Median	0.18	0.77
Interquartile range	0.44	0.62

Perceptual representations used by Thoret et al., 2021, *Nat. Hum. Behav.*

- auditory spectrum: cochlea
- cortical representation: STMF

All representations are averaged over time.

	cochlea	STMF
n_Ψ	128	30976

Perceptual representations used by Thoret et al., 2021, *Nat. Hum. Behav.*

- auditory spectrum: cochlea
- cortical representation: STMF

Time–frequency representations

- Short-Time Fourier Transform: STFT
- Joint time–frequency scattering transform: scattering

All representations are averaged over time.

	cochlea	STMF	STFT	scattering
n_Ψ	128	30976	513	2204

Perceptual representations used by Thoret et al., 2021, *Nat. Hum. Behav.*

- auditory spectrum: cochlea
- cortical representation: STMF

Time–frequency representations

- Short-Time Fourier Transform: STFT
- Joint time–frequency scattering transform: scattering

Deep neural network embeddings

- CLAP: trained on general audio for text2speech
- EnCodec: trained on music for compression
- MERT: trained on music for 13 tasks
  - ▷ averaged (MERTAV) or concatened (MERTCAT)

All representations are averaged over time.

	cochlea	STMF	STFT	scattering	CLAP	EnCodec	MERTAV	MERTCAT
n_Ψ	128	30976	513	2204	1024	128	768	9984

Metric learning in representation spaces: explained variance

Performance criterion: $\mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}_\star}, \mathbf{s})^2 \in [0, 1]$ (Thoret et al., 2021, *Nat. Hum. Behav.*)

▷ squared Pearson correlation between **learned distance** and **dissimilarity ratings**

17 datasets from studies between 1977 and 2016	perceptual cochlea $n_\Psi = 128$	 STMF $n_\Psi = 30976$	deep MERTCAT $n_\Psi = 9984$
Grey, 1977	0.48	0.84	**1.00**
Grey et al., 1978	0.11	0.33	**0.77**
Iverson et al., 1993: Whole	0.16	0.87	**0.95**
Iverson et al., 1993: Onset	0.07	0.22	**0.93**
Iverson et al., 1993: Remainder	0.03	0.27	**0.87**
McAdams et al., 1995	0.30	0.77	**0.97**
Lakatos et al., 2000: Harmonic	0.19	0.85	**0.98**
Lakatos et al., 2000: Percussive	0.18	0.27	**0.97**
Lakatos et al., 2000: Combined	0.13	0.33	**0.94**
Barthet et al., 2010	0.74	**0.98**	0.65
Patil et al., 2012: A3	0.62	0.97	**1.00**
Patil et al., 2012: DX4	0.66	0.99	**1.00**
Patil et al., 2012: GD4	0.46	0.95	**1.00**
Siedenburg et al., 2016: Exp 2A, Set 1	0.62	0.95	**1.00**
Siedenburg et al., 2016: Exp 2A, Set 2	0.73	0.99	**1.00**
Siedenburg et al., 2016: Exp 2A, Set 3	0.10	0.53	**1.00**
Siedenburg et al., 2016: Exp 2B, Set 3	0.07	0.46	**1.00**
Median	0.18	0.77	**0.97**
Interquartile range	0.44	0.62	0.06

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\mathrm{subjects}} = 24$ musician participants: musical instruction and playing experience

STFT	cochlea	scattering	STMF	CLAP	EnCodec	MERTAV	MERTCAT
0.40	0.62	0.31	0.95	0.76	0.23	0.11	**1.00**

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience

STFT	cochlea	scattering	STMF	CLAP	EnCodec	MERTAV	MERTCAT
0.40	0.62	0.31	0.95	0.76	0.23	0.11	**1.00**

Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\text{subjects}} = 34$ participants, including 18 musicians

STFT	cochlea	scattering	STMF	CLAP	EnCodec	MERTAV	MERTCAT
0.31	0.19	0.16	0.85	0.74	0.31	0.08	**0.98**

Robustness of the learning procedure against degraded ratings

Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Robustness of the learning procedure against degraded ratings

Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Robustness of the learning procedure against degraded ratings

Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Complement and extend the reported explained variance performance by

i) quantifying robustness of the learning procedure to noisy ratings

ii) comparing robustness for different representations and noise levels

Robustness of the learning procedure against degraded ratings

Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Complement and extend the reported explained variance performance by

i) quantifying robustness of the learning procedure to **noisy ratings**

ii) comparing robustness for different representations and noise levels

Random degradation of ratings

$$y^{(\delta)}_{\{i,j\}} = \min(1, \max(0, s_{\{i,j\}} + \delta \cdot \xi)),$$

$\xi \sim \mathcal{N}(0,1)$ i.i.d. , $\delta > 0$: noise std

$\mathbf{y}^{(\delta)}$: degraded $\mathbf{s}$ at noise level δ

Robustness of the learning procedure against degraded ratings

Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Complement and extend the reported explained variance performance by

i) quantifying robustness of the learning procedure to **noisy ratings**

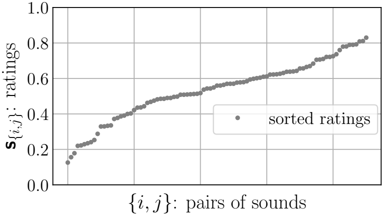
ii) comparing robustness for different representations and noise levels

Random degradation of ratings

$$\mathrm{y}^{(\delta)}_{\{i,j\}} = \min(1, \max(0, \mathrm{s}_{\{i,j\}} + \delta \cdot \xi)),$$

$\xi \sim \mathcal{N}(0,1)$ i.i.d. , $\delta > 0$: noise std

$\mathbf{y}^{(\delta)}$: degraded $\mathbf{s}$ at noise level δ



Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Complement and extend the reported explained variance performance by

i) quantifying robustness of the learning procedure to **noisy ratings**

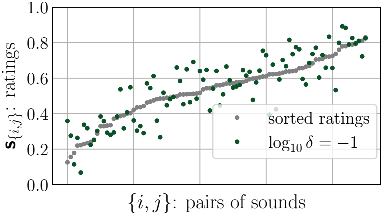
ii) comparing robustness for different representations and noise levels

Random degradation of ratings

$$\mathrm{y}^{(\delta)}_{\{i,j\}} = \min(1, \max(0, \mathrm{s}_{\{i,j\}} + \delta \cdot \xi)),$$

$\xi \sim \mathcal{N}(0,1)$ i.i.d. , $\delta > 0$: noise std

$\mathbf{y}^{(\delta)}$: degraded $\mathbf{s}$ at noise level δ



Averaged dissimilarity ratings

$\Longrightarrow$ no confidence level on explained variance provided

Fluctuations in dissimilarity ratings: very large, both

- between different subjects
- for a subject, between different times and orders of presentation of sound pairs

Complement and extend the reported explained variance performance by

i) quantifying robustness of the learning procedure to **noisy ratings**

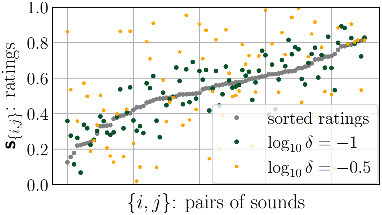
ii) comparing robustness for different representations and noise levels

Random degradation of ratings

$$\mathrm{y}^{(\delta)}_{\{i,j\}} = \min(1, \max(0, \mathrm{s}_{\{i,j\}} + \delta \cdot \xi)),$$

$\xi \sim \mathcal{N}(0,1)$ i.i.d. , $\delta > 0$: noise std

$\mathbf{y}^{(\delta)}$: degraded $\mathbf{s}$ at noise level δ



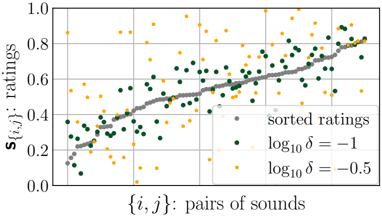
Robustness of the learning procedure against degraded ratings

Random degradation of ratings

$$y^{(\delta)}_{\{i,j\}} = \min(1, \max(0, s_{\{i,j\}} + \delta \cdot \xi)),$$

$\xi \sim \mathcal{N}(0,1)$ i.i.d. , $\delta > 0$: noise std

$\mathbf{y}^{(\delta)}$: degraded $\mathbf{s}$ at noise level δ



Experimental setup to quantify robustness

i) learning on **noisy** dissimilarity ratings

$$\mathbf{w}_\delta \in \underset{\mathbf{w}\in\mathbb{R}^{n_\Psi}}{\mathrm{Argmax}}\, \mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}}, \mathbf{y}^{(\delta)})$$

for 5 realizations of $\mathbf{y}^{(\delta)}$, and 9 values of δ logarithmically spaced in $[0.1, 10]$

ii) explained variance of **averaged ratings** by the learned distance $\mathcal{P}(\mathrm{d}^{\Psi}_{\mathbf{w}_\delta}, \mathbf{s})^2$

Robustness of the learning procedure against degraded ratings

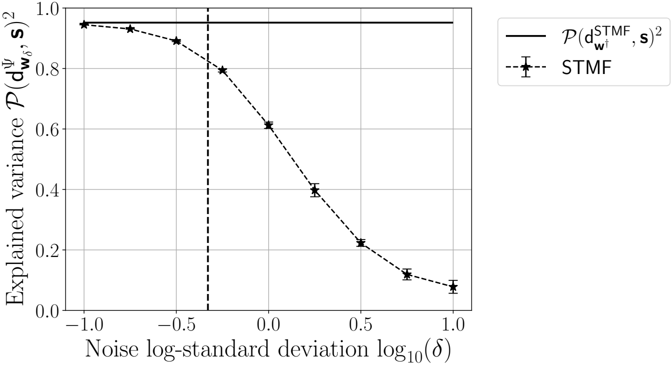
Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience

Robustness of the learning procedure against degraded ratings

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience

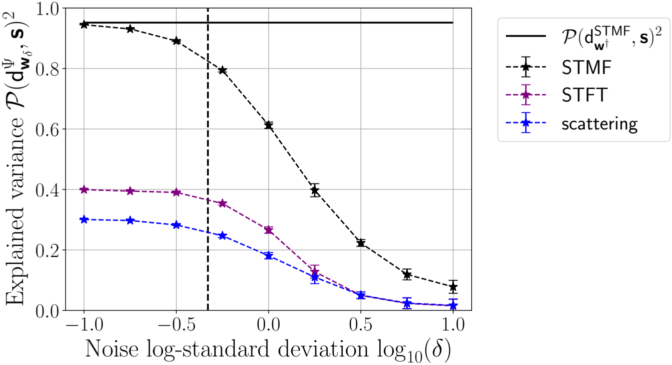


Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience



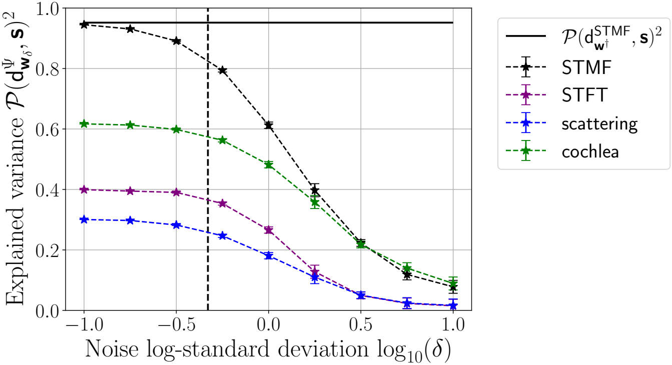
Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Robustness of the learning procedure against degraded ratings

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience



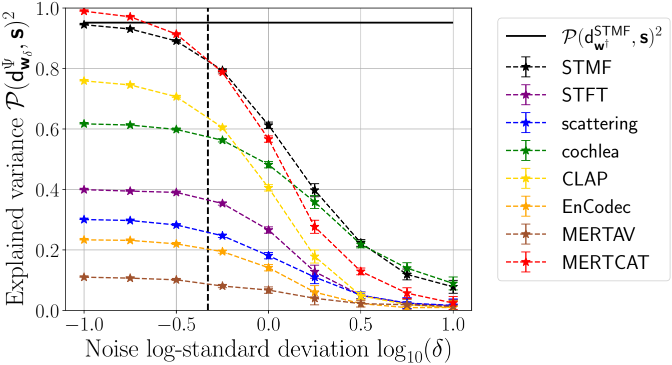
Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Robustness of the learning procedure against degraded ratings

Siedenburg et al., 2016, *Front. Psychol.*: Exp. 2A, Set 1

- 14 acoustic recordings from Vienna Symphonic Library `https://www.vsl.co.at`
- $m_{\text{subjects}} = 24$ musician participants: musical instruction and playing experience



Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

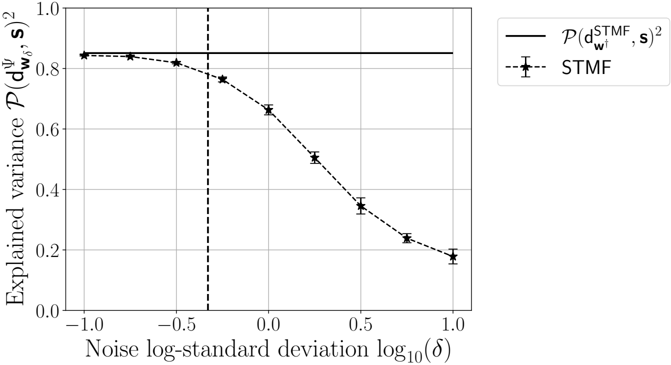
Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\mathrm{subjects}} = 34$ participants, including 18 musicians

Robustness of the learning procedure against degraded ratings

Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\text{subjects}} = 34$ participants, including 18 musicians

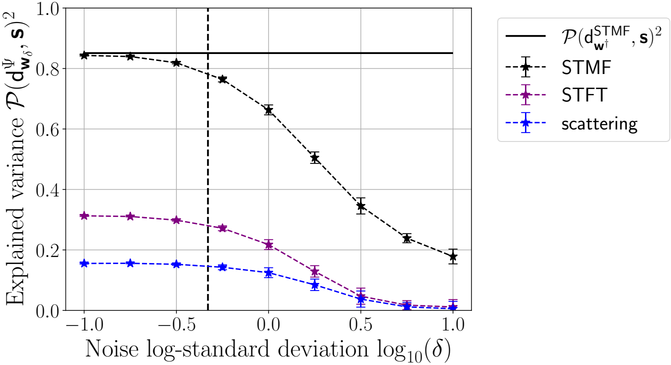


Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\text{subjects}} = 34$ participants, including 18 musicians



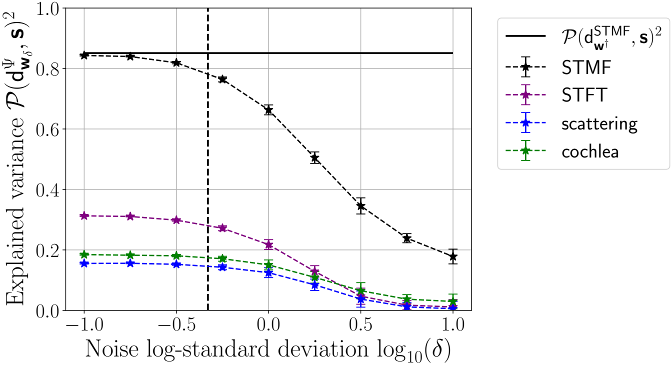
Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Robustness of the learning procedure against degraded ratings

Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\text{subjects}} = 34$ participants, including 18 musicians



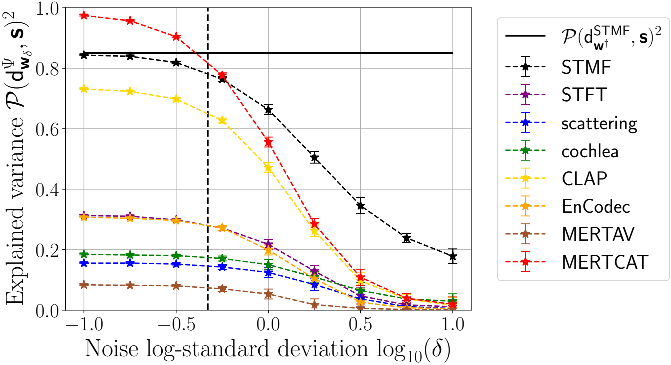
Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Robustness of the learning procedure against degraded ratings

Lakatos et al., 2000, *Percept. Psychophys.*: Harmonic

- 17 recorded sounds
- $m_{\text{subjects}} = 34$ participants, including 18 musicians



Typical standard deviation of human ratings $\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}}$

(P. Aumond et al., 2017,Appl. Sci.)

Robustness against degraded ratings: global results

Compared robustness for the different representations

- if $\delta \leq \bar{\delta}$ best explained variance and robustness for metric learned on MERTCAT
- for $\delta > \bar{\delta}$ explained variance decreases slower for metric learned on STMF
- $\forall \delta$ metrics learned on CLAP: good explained variance and robustness

CLAP	STMF	MERTCAT
$n_\Psi = 1024$	$n_\Psi = 30976$	$n_\Psi = 9984$

▷ quantified by comparison of the areas under the curves $\log_{10} \delta \mapsto \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}_\delta}, \mathbf{s})^2$

See paper and companion toolbox github.com/bpascal-fr/timbre-metric-learning

Meta-analysis on 17 datasets

- deep embeddings vs. classical time-frequency and perceptual representations
- deep neural networks trained on audio: encode substrate of timbre perception
- corrected and augmented metric learning procedure: **explained variance**
- **robustness** against inter- and intra- subject variability in human ratings

github.com/bpascal-fr/timbre-metric-learning

Meta-analysis on 17 datasets

- deep embeddings vs. classical time-frequency and perceptual representations
- deep neural networks trained on audio: encode substrate of timbre perception
- corrected and augmented metric learning procedure: **explained variance**
- **robustness** against inter- and intra- subject variability in human ratings

github.com/bpascal-fr/timbre-metric-learning

Future work: Tackle open questions in auditory cognitive neuroscience

- training with **all** ratings (no averaging over participants): inter-subject variability
- CLAP, MERTCAT: speech, environmental sounds, animal bioacoustics