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

Barbara Pascal, Mathieu Lagrange

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

1st Barbara Pascal
Nantes Université, École Centrale Nantes, CNRS
LS2N, UMR 6004, F-44000 Nantes, France
barbara.pascal@cnrs.fr

2nd Mathieu Lagrange
Nantes Université, École Centrale Nantes, CNRS
LS2N, UMR 6004, F-44000 Nantes, France
mathieu.lagrange@ls2n.fr

***Abstract*—Timbre, encompassing an intricate set of acoustic cues, is key to identify sound sources, and especially to discriminate musical instruments and playing styles. Psychoacoustic studies focusing on timbre deploy massive efforts to explain human timbre perception. To uncover the acoustic substrates of timbre perceived dissimilarity, a recent work leveraged metric learning strategies on different perceptual representations and performed a meta-analysis of seventeen dissimilarity rated musical audio datasets. By learning salient patterns in very high-dimensional representations, metric learning accounts for a reasonably large part of the variance in human ratings. The present work shows that combining the most recent deep audio embeddings with a metric learning approach makes it possible to explain almost all the variance in human dissimilarity ratings. Furthermore, the robustness of the learning procedure against simulated human rating variability is thoroughly investigated. Intensive numerical experiments support the explanatory power and robustness against degraded dissimilarity ratings of the learning metric strategy using deep embeddings.**

***Index Terms*—Audio timbre perception, distance metric learning, time/frequency analysis, scattering transform, spectrotemporal modulations, deep neural networks, deep embeddings, robustness analysis.**

I. Introduction

Context. Understanding the way humans extract information and make judgments about their environment based on sounds still trigger much interdisciplinary research at the frontier between digital audio processing and psychoacoustics [1], [2]. In particular, the notion of *timbre*, related to the perceived sound quality, emerges from an intricate bundle of acoustic cues and provides important information about the sources and mechanisms which produced the sounds. It is hence key to recognizing complex sound sources, encountered, e.g., in music. However, modeling timbre human perception is still a burning question in cognitive neurosciences [3]–[7].

Related work. The historical approach to reveal the acoustic explanatory features of timbre perception uses *multidimensional scaling* [3], [4], [8]–[12]. It relies on *dissimilarity ratings*, and consists in representing audio samples in a low dimensional space, called the *timbre* space, such that the distance between a pair of sounds reflects their dissimilarity. Then, the uncovered latent dimensions of the timbre space are correlated with psychophysic acoustic descriptors, such as logarithm attack time and spectral centroid [13]. Despite providing a broad understanding of the timbre acoustic correlates, multidimensional scaling requires arbitrary choices and some ad-hoc parameter-tuning, impairing its replicability [14]. Moreover, due to their low descriptive power, standard acoustic features only partly explains timbre perception [5].

Instead of correlating dimensions of timbre space with subjective a priori crafted acoustic features, it has thus been proposed in [14] to model human dissimilarity ratings with weighted distances, computed on audio representations mimicking the primary auditory cortex. The weights are learned by maximizing the correlation of the distance with human ratings, so that they should fire on salient patterns based on which the human dissimilarity judgment is made. One major advantage of this metric learning strategy is that the hidden acoustic features involved in timbre perception are inferred in a fully data-driven manner through the learning process; it is thus more objective than multidimensional scaling. Furthermore, it is more flexible, and able to extract abstract acoustic features hidden in high-dimensional representations, which could not be devised from scratch. Though, while the high inter- and intra-subject variability in dissimilarity rating tasks is documented [15], the metric learning procedure of [14] is performed solely on averaged dissimilarity ratings, impairing uncertainty quantification in the explained variances. Furthermore, most learning procedures are highly sensitive to input noise, mostly because of the optimization of a nonconvex criterion, with numerous local minima [16]. The conclusions drawn from the metrics learned according to [14] could hence change drastically if applied to dissimilarity ratings collected at a different time or from different participants.

Going a step further in complexity and abstraction, deep embeddings have recently shown impressive performance in standard audio processing tasks [17]–[19]. These successes motivated the search for acoustic features explaining human listening encoded in the structure of trained deep neural networks, exploiting parallels with vision [2]. Recently, [1] demonstrated that the layer hierarchy of neural networks trained on audio data, encodes psychoacoustic features of increasing complexity involved in human listening.

Contributions and outline. First, the metric learning procedure developed in [14] is revisited to correct and robustify the learning scheme. Then, classical time–frequency and perceptually motivated audio representations are compared to the most recent deep embeddings in terms of variance of the human ratings explained by the learned metric. Finally, the robustness of the learning procedure depending on the representation is assessed through intensive numerical simulations. All audio representations considered in the paper are presented in Sec. II, and the timbre datasets in Sec. III. Sec. IV describes the metric learning framework and compares the explained variance reached with the different representations. Sec. V proposes a robustness assessment procedure and compares the different embeddings through exhaustive numerical experiments.

II. Models of Human Audio Timbre Perception

A. Time–Frequency representations

Fourier based representations have long been considered to model musical sounds. The Short-Time Fourier Transform (STFT) [20], is thus considered as a baseline. Window and hop sizes of respectively 1024 and 512 bins are considered, leading to a feature dimension of $n_{\mathsf{STFT}} = 513$.

Depth of representation have been shown to improve modeling performance [21], hence the joint time–frequency scattering [22]

is also considered. It consists of a two-step cascade of wavelets and modulus operators. Quality factors for the wavelet operators are determined according to the knowledge of time–frequency properties of musical sounds. For the first order, the quality factor is 8, and, for the second, the scale and rate quality factors are both equal to 2 leading to a larger feature dimension of $n_{\mathsf{scattering}} = 2204$.

B. Perceptual representations

Both the lowest and the highest-dimensional perceptually motivated representations studied in [14] are considered.

The cochlea feature is computed with perceptually motived post-processing of 128 constant-Q asymmetric bandpass filters equally spaced on a logarithmic frequency–scale, leading to a feature dimension of $n_{\mathsf{cochlea}} = 128$.

The SpectroTemporal Modulation Frequency (STMF) representation can be seen as a series of spectrograms filtered according to different rates and scales by the application of a two-dimensional Fourier transform to the cochlear spectrogram. The second stage of filtering aims at modeling the evidence of rate–scale sensitive population of neurons in the early auditory cortex [23]. It results in a two-dimensional array, also called the modulation power spectrum, whose dimensions are *i)* spectral modulation (scale, for 11 cycles per octave), and *ii)* temporal modulation (rate, for 22 frequencies). This leads to a much higher-dimensional feature with $n_{\mathsf{STMF}} = 128 \times 11 \times 22 = 30976$. It worth noting that the scattering representation can be considered as an idealized STMF representation.

C. Learned representations

Many learned embeddings are now available for representing audio. The use of VGGish embeddings [24] are effective for general audio processing [17] and still of widespread use. However recent empirical evidence tends to demonstrate that deep embeddings learned on domain-specific audio data lead to significant performance improvement of the modeling quality [25].

Aiming at modeling *musical* timbre perception, this study thus focuses on state-of-the-art embeddings largely trained on musical data:

- EnCodec [26] with quantization removed, leading to a feature EnCodec with dimension of $n_{\mathsf{EnCodec}} = 128$,
- CLAP [27], a feature CLAP with dimension of $n_{\mathsf{CLAP}} = 1024$,
- MERT [28].

The latter exposes thirteen different embeddings with contrasted performance for various downstream tasks, concatenated into a MERTCAT feature with $n_{\mathsf{MERTCAT}} = 9984$, and a learned weighted average, yielding a MERTAV embedding with $n_{\mathsf{MERTAV}} = 768$.

All features are considered as averaged over time.

III. Datasets

The seventeen datasets reanalyzed by [14], listed in Tab. I, are considered. Each is composed of audio samples, denoted $\{a_1, \dots, a_\ell\}$. The number of sounds ℓ is comprised between eleven and twenty for the datasets considered in this study. The datasets from [4], [5], [9], [11] contain recorded instruments sounds, while the datasets of [3], [8], [10], [12] are composed of resynthesized and simulated sounds.

Given a collection of audio samples, standard experiments in psychoacoustics consists in asking the subjects to attribute to each pair (a_i, a_j) a *dissimilarity* rating $\mathsf{s}_{\{i,j\}} \in [0, 1]$, where $\mathsf{s}_{\{i,j\}} = 0$ accounts for exactly similar audio samples a_i, a_j, while $\mathsf{s}_{\{i,j\}} = 1$ corresponds to maximally different samples. For each dataset, the dissimilarity ratings of all the unordered pairs $\{i, j\}$ of distinct elements are averaged over all participants and stored in a vector $\mathbf{s}$. As this annotation task requires significant cognitive efforts, it is limited to small datasets of not more than twenty sounds [3]–[5], [7].

IV. Metric learning in representation spaces

Human dissimilarity ratings condense complex perceptual judgments relying on intricate high-level audio characteristics. Therefore, the decision-making process can hardy be fully modeled. Instead, to capture the main features of timbre perception, human ratings are fitted to a parametric distance through a learning procedure [14]. Then, provided the obtained fit is sufficiently good, the salient patterns on which the learned weights fire can be considered as relevant explanatory features for timbre perception.

Parametric distance in representation space. In the present paper, following the general principles of metric learning [29], the dissimilarity ratings are fitted using a *parametric* distances of the form

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

where Ψ is a n_Ψ-dimensional audio *representation*, e.g., one of those described in Sec. III, and $\mathbf{w}$ is an n_Ψ-dimensional vector of weights.

Reward function. Metric learning consists in selecting the weights such that the distance fits the human dissimilarity ratings. The quality of the fit is measured through the *Pearson correlation*

$$\mathcal{P}(\mathsf{d}_{\mathbf{w}}^{\Psi}, \mathbf{s}) = \sum_{\{i,j\}} \frac{\left(\mathsf{d}_{\mathbf{w}}^{\Psi}(a_i, a_j)^2 - \mu_{\mathbf{w}}\right)\left(\mathsf{s}_{\{i,j\}} - \mu_{\mathbf{s}}\right)}{\sigma_{\mathbf{w}}\sigma_{\mathbf{s}}}, \tag{2}$$

where the sum runs over all unordered pairs of distincts sounds; $\mu_{\mathbf{w}}$ (resp. $\mu_{\mathbf{s}}$) denotes the empirical mean of $\mathsf{d}_{\mathbf{w}}^{\Psi}(a_i, a_j)^2$ (resp. $\mathsf{s}_{\{i,j\}}$) over all pairs of sounds; and $\sigma_{\mathbf{w}}$ (resp. $\sigma_{\mathbf{s}}$) refers to the empirical standard deviation. Using the *squared* distance in the Pearson correlation instead of the distance itself simplifies the computations, but does not impact the interpretation.

The Pearson correlation ranges from -1, corresponding to perfect linear anticorrelation, to 1, in case of perfect linear correlation between the parametric distance and human ratings. Its main advantage is that it is invariant under mean shifts and variance rescalings.

Learning framework. The learning task consists in maximizing the reward function (2) over the training audio dataset, that is to find

$$\mathbf{w}_\star \in \operatorname*{Argmax}_{\mathbf{w} \in \mathbb{R}^{n_\Psi}} \mathcal{P}(\mathsf{d}_{\mathbf{w}}^{\Psi}, \mathbf{s}). \tag{3}$$

Although the learned weights depend on the representation, Ψ is omitted in $\mathbf{w}_\star$ for the sake of readability of the learned distance $\mathsf{d}_{\mathbf{w}_\star}^{\Psi}$. Performance are then quantified through the *explained variance*, defined as the squared Pearson correlation. Large explained variance corresponds to accurate modeling of the human ratings by the learned metric, while values close to zero indicate poor fit. Achieving a good fit ensures that audio samples with similar timbre are closed in terms of the learned distance, and reciprocally, audio samples with different timbres are far apart.

Datasets in psychoacoustics are mostly of very limited size: ℓ is smaller than twenty for the datasets considered in the present study. Hence, the reward function (7) is maximized over the *entire* dataset of pairs of sounds. To rule out the risk of *overfitting*, [14, Methods & Supplementary] performed an exhaustive leave-one-sound-out cross validation confirming that the learning procedure is consistent.

Optimization. The Pearson correlation maximized in (7) being differential in the learned weights, optimization is performed using the limited memory Boyden-Fletcher-Golfarb-Shanno quasi-Newton algorithm with box constraints [30]–[32]. This algorithm has three major advantages : first, it is descent-step free, second, it is capable

Dimension n_Ψ	STFT 513	cochlea 128	scattering 2204	STMF 30976	CLAP 1024	EnCodec 128	MERTAV 768	MERTCAT 9984	m_{subjects}
Grey1977 [8]	0.29	0.48	0.25	0.84	0.73	0.23	0.02	**1.00**	22
Grey1978 [3]	0.18	0.11	0.21	0.33	0.36	0.16	0.08	**0.77**	22
Iverson1993_Whole [9]	0.46	0.16	0.25	0.87	0.59	0.30	0.21	**0.95**	10
Iverson1993_Onset [9]	0.18	0.07	0.12	0.22	0.42	0.06	0.11	**0.93**	9
Iverson1993_Remainder [9]	0.16	0.03	0.02	0.27	0.39	0.16	0.07	**0.87**	9
McAdams1995 [10]	0.14	0.30	0.06	0.77	0.31	0.14	0.05	**0.97**	24
Lakatos2000_Harm [11]	0.31	0.19	0.16	0.85	0.74	0.31	0.08	**0.98**	34
Lakatos2000_Perc [11]	0.10	0.18	0.07	0.27	0.37	0.08	0.13	**0.97**	34
Lakatos2000_Comb [11]	0.16	0.13	0.19	0.33	0.50	0.07	0.13	**0.94**	34
Barthet2010 [12]	0.11	0.74	0.18	**0.98**	0.29	0.08	0.21	0.65	16
Patil2012_A3 [4]	0.79	0.62	0.75	0.97	0.97	0.55	0.46	**1.00**	20
Patil2012_DX4 [4]	0.85	0.66	0.81	0.99	0.94	0.72	0.18	**1.00**	20
Patil2012_GD4 [4]	0.85	0.46	0.76	0.95	0.98	0.65	0.74	**1.00**	20
Siedenburg2016_e2set1 [5]	0.40	0.62	0.31	0.95	0.76	0.23	0.11	**1.00**	24
Siedenburg2016_e2set2 [5]	0.60	0.73	0.35	0.99	0.59	0.46	0.07	**1.00**	24
Siedenburg2016_e2set3 [5]	0.33	0.10	0.40	0.53	0.92	0.23	0.35	**1.00**	24
Siedenburg2016_e3 [5]	0.26	0.07	0.32	0.46	0.90	0.18	0.18	**1.00**	24
Median	0.26	0.18	0.21	0.77	0.59	0.18	0.11	**0.97**	22
Interquartile range	0.27	0.44	0.19	0.62	0.45	0.19	0.12	0.06	4

TABLE I: **Squared Pearson correlation between collected dissimilarity scores and learned metrics in different representation spaces.** Pearson correlation close to one (resp. zero) indicates very good (resp. very poor) fit. Best fits are emphasized in bold.

to optimize over variables of large dimension, and, finally, its convergence is *quadratic*, and hence very fast, provided the initial point is close enough to the optimum. The overall learning procedure has been reimplemented end-to-end in order to augment and correct the procedure of [14].[2] First, in contrast to [14], in which the initialization of the learning algorithm is random, the present work proposes a deterministic *warm* initialization

$$\mathbf{w}^{[0]} \in \underset{\mathbf{w}\in\mathbb{R}^{n_\Psi}}{\operatorname{Argmin}} \sum_{\{i,j\}} \left| \mathsf{d}^{\Psi}_{\mathbf{w}}(a_i, a_j)^2 - \mathsf{s}_{\{i,j\}} \right|^2 \tag{4}$$

in order to robustify the learning process, to converge faster, and toward a better local minimum. The initialization (4) is computed using a nonnegative least squares solver [33]. Additionally, in the codes[1] accompanying [14], the terms corresponding to the derivatives of the mean $\mu_{\mathbf{w}}$ with respect to the weights are missing in the computation of the gradient of the Pearson correlation reward function (2). This omission has been corrected in the proposed reimplementation. In practice, the maximum number of iterations is set to 10^4, and maximum number of line search steps at each iteration to 50. Weights are forced to belong to lie between 1 and 10^{15}. Tolerance on both the reward function increments and projected gradient is set to 10^{-36} to ensure high accuracy in the computation of the optimizer.

Numerical results. Following [14], the explained variance of human ratings by the learned metrics is quantified through the squared Pearson correlation, reported for the seventeen datasets and the eight audio representations in Table I. First, the performance obtained using the cochlea representation (second column) and, more significantly using the STMF representation (fourth column), are larger by up to 6% of explained variance compared to those reported in [14, Table 1, columns 4 and 6], demonstrating that the learning algorithm converges toward a better optimum when using the proposed warm initialization (4), especially when the optimization is performed in high dimension. Second, the MERTCAT representation-based learned distance almost systematically achieves the largest explained variance compared to all other representations, and often reaches one, corresponding to a *perfect* fit. Quantitatively, on median across the seventeen datasets, the learned distance using the MERTCAT representation explains 97% of the variance in human rating,s against 77% for the STMF representation-based distance. It worth noting that the MERTCAT representation, although concatenating thirteen embeddings, is still three times smaller compared to the STMF representation: $n_{\mathsf{MERTCAT}} = 9984$ while $n_{\mathsf{STMF}} = 30976$. This results in a significantly decrease of the complexity of the optimization algorithm solving (7), which is reduced by a factor of almost ten. It opens the way to learning over larger datasets, e.g., taking into account all subject ratings instead of learning on ratings averaged over the participants of the experiment. Data and codes to reproduce the results of Tab. I have been made publicly available.[2]

V. Robustness analysis

Dissimilarity ratings are prone to large fluctuations, both between different subjects and for a given subject at different time and depending on the order in which the audio sample pairs are presented. The analysis in [14] is performed on averaged dissimilarity ratings over all the participants to the experiments. Consequently, the reported performance are not accompanied with grounded confidence levels. The present section proposes a framework to complete and extend the study of [14] in two ways: *i)* by using randomly generated degraded dissimilarity ratings it enables to quantify the robustness of the learning procedure, and hence to support the consistency of the explained variances obtained in Sec. IV on averaged human ratings, *ii)* by comparing the robustness of the learned metrics depending on the audio representation and on the degradation level.

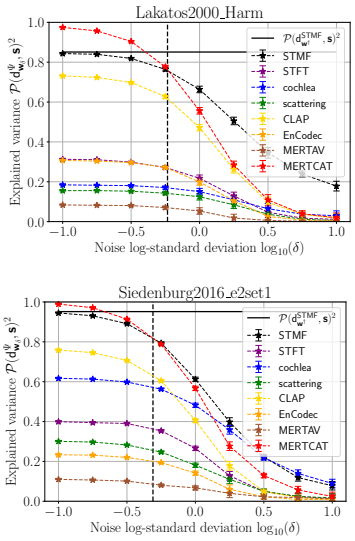
Random degradation of human ratings. To assess the robustness against variability of the dissimilarity ratings of the learning procedure described in Sec. IV, degraded versions of the collected ratings are considered. In order to mimic the large uncertainty surrounding the annotation process, resulting in both inter- and intra-subject variability, *noisy* ratings vectors are generated under the form

$$\mathsf{y}^{(\delta)}_{\{i,j\}} = \min(1, \max(0, \mathsf{s}_{\{i,j\}} + \delta \cdot \xi)), \tag{5}$$

where $\xi \sim \mathcal{N}(0,1)$ are i.i.d. random perturbations; δ encodes the level of degradation; and the $\min$ and $\max$ perform clipping, ensuring that degraded ratings lie in $[0, 1]$.

[1] https://github.com/EtienneTho/musical-timbre-studies

[2] https://github.com/bpascal-fr/timbre-metric-learning/



(a) Curves $\log_{10}\delta \mapsto \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}_\delta}, \mathbf{s})^2$.

	Area (a.u.)	Relative area $-10\log_{10}(A_{\mathsf{STMF}}/A_{\mathsf{r}})$ (dB)						
	STMF	STFT	cochlea	scattering	CLAP	EnCodec	MERTAV	MERTCAT
Grey1977 [8]	*1.09*	-5.75	-1.44	-6.27	-1.06	-6.20	-15.48	**-0.23**
Grey1978 [3]	*0.34*	-2.69	-3.55	-2.17	0.20	-2.92	-5.99	**3.66**
Iverson1993_Whole [9]	*1.12*	-2.97	-7.24	-5.79	-2.45	-4.79	-6.58	**-0.54**
Iverson1993_Onset [9]	*0.21*	-0.99	-4.56	-2.51	2.89	-6.20	-2.18	**6.28**
Iverson1993_Remainder [9]	*0.28*	-2.21	-10.21	-10.61	1.69	-1.92	-5.84	**4.70**
McAdams1995 [10]	*0.99*	-8.01	-3.12	-12.26	-5.05	-8.15	-12.50	**0.06**
Lakatos2000_Harm [11]	*1.17*	-5.09	-6.71	-7.76	-1.54	-5.36	-11.26	**-0.57**
Lakatos2000_Perc [11]	*0.34*	-5.24	-1.28	-6.72	0.54	-5.15	-3.39	**4.52**
Lakatos2000_Comb [11]	*0.41*	-3.35	-3.51	-2.24	0.96	-6.94	-3.89	**3.85**
Barthet2010 [12]	*1.30*	-10.38	**-0.10**	-8.04	-6.14	-11.67	-7.66	-2.98
Patil2012_A3 [4]	*1.15*	-1.35	-1.63	-1.19	-0.82	-2.71	-3.45	**-0.46**
Patil2012_DX4 [4]	*1.08*	-0.49	-0.83	-1.10	-0.30	-0.81	-7.50	**-0.26**
Patil2012_GD4 [4]	*0.97*	-0.23	-2.23	-0.58	0.32	-1.07	-0.42	**0.35**
Siedenburg2016_e2set1 [5]	*1.12*	-3.96	-1.29	-5.21	-1.61	-6.55	-9.60	**-0.28**
Siedenburg2016_e2set2 [5]	*1.13*	-2.30	**-0.48**	-4.55	-3.12	-2.91	-11.49	-0.61
Siedenburg2016_e2set3 [5]	*0.54*	-1.60	-6.80	-0.76	2.90	-3.39	-1.09	**2.99**
Siedenburg2016_e3 [5]	*0.48*	-2.33	-8.74	-1.22	2.73	-4.51	-3.46	**3.16**
Mean	*0.72*	-3.10	-3.35	-4.16	-0.52	-4.28	-5.88	**1.24**
Standard deviation	*0.40*	2.61	2.98	3.47	2.39	2.75	4.21	2.39

(b) (Relative) Area under the curve $\log_{10}\delta \mapsto \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}_\delta}, \mathbf{s})^2$, with weights $\mathbf{w}_\delta$ learned according to Eq. (7).

Fig. 1: **Robustness of the learning procedure against degraded ratings.** (a) Explained variance of the human ratings by the metric learned on synthetic degraded ratings for several levels of noise with error bars computed on 5 realizations of the degraded ratings. The vertical dashed line indicates the typical standard deviation of human ratings $\overline{\delta}$ given in Eq. (6). (b) First column: area under the explained variance curve for a learned metric using the STMF representation, to be used as a reference; second to fifth columns: ratio, expressed in decibels, between the reference area of first column and the area under the explained variance curve for each of the seven remaining representations. The higher the ratio, the more robust the learning procedure using this representation. All quantities are averaged over 5 realizations of the noisy ratings. Relative area of 0 dB corresponds to exactly the same robustness as STMF.

Experimental setup. The exact same setup is applied to each of the seventeen datasets. Nine levels of degradation are considered, logarithmically distributed between 0.1 and 10. For each level of degradation, five independent realizations of the noisy ratings are generated. For reference, the vertical dashed line in Fig. 1a indicates the typical standard deviation of human ratings reported in [15]

$$\overline{\delta} = 0.1 \times \sqrt{m_{\text{subjects}}} \tag{6}$$

where m_{subjects} denotes the number of participants to the study, provided in Tab. I, rightmost column. Then, for each of the eight representations, each degradation level, and each realization, the learning framework described in Sec. IV is applied to learn weights from *noisy* ratings

$$\mathbf{w}_\delta \in \operatorname*{Argmax}_{\mathbf{w}\in\mathbb{R}^{n_\Psi}} \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}}, \mathbf{y}^{(\delta)}). \tag{7}$$

Performance metrics. For very low noise level, the metric learned from degraded ratings $\mathsf{d}^{\Psi}_{\mathbf{w}_\delta}$ is expected to be very close to the metric learned from original ratings $\mathsf{d}^{\Psi}_{\mathbf{w}_\star}$, and hence to achieve similar explained variance of human ratings. As the noise level increases, the training ratings gets more and more degraded, and the explained variance should decrease. To illustrate the performance drop off induced by training on noisy ratings, the curves $\log_{10}\delta \mapsto \mathcal{P}(\mathsf{d}^{\Psi}_{\mathbf{w}_\delta}, \mathbf{s})^2$ are displayed in Fig. 1a for two datasets. The quality of the learned metric and the robustness to learning from degraded ratings are jointly evaluated through the area under the explained variance curve. This area is comprise between 0 and 2, and its maximal value is reached if and only if the explained variance saturates at one whatever the noise level, corresponding to perfect fit of the learned metric to human ratings and perfect robustness. On the contrary, it equals zero if and only if the learned metrics explains absolutely nothing about neither human nor degraded ratings. The area for the STMF is used as a reference. To emphasize improvements and degradations compared to this state-of-the-art representation [14], the robustness of the learning procedure relying on the other representations is quantified relatively to this reference though the log relative area expressed in decibels.

Numerical results. Fig 1a shows that, for the two example datasets, the explained variance curves, corresponding to the eight representations, are all monotonously decreasing as the noise level increases. For both Lakatos2000_Harm [11] and Siedenburg2016_e2set1 [5], as soon as the degradation level remains smaller than the expected human variability (6), the metric learned on the MERTCAT representation yields better explained variance and is more robust compared to metric learning on the STMF representation. On the contrary, for high degradation levels, the explained variance is decreasing slower for STMF-based than for MERTCAT-based metrics. Similar conclusions are consistently observed for all other datasets.[3]

Positive log relative areas in Tab. 1b indicates higher robustness to degraded ratings compared to the reference STMF. On average, the MERTCAT deep embedding leads to higher robustness than all other representations. Interestingly, while being far smaller dimensional, CLAP achieves the third best robustness on average, just after the very high-dimensional reference STMF, corresponding to zero log relative area by definition.

VI. Conclusion and Perspectives

Leveraging the metric learning framework designed in [14], the present meta-analysis demonstrates, through an exhaustive comparison of the most recent deep embeddings with classical time–frequency representations and perceptually motivated audio representations informed by the physiology of the human auditory cortex, the impressive ability of deep neural networks trained on music

[3] https://github.com/bpascal-fr/timbre-metric-learning/figures

datasets to encode the acoustic substrates of timbre perception. To uncover the salient patterns in audio representations explaining timbre dissimilarity, a corrected and augmented version of the metric learning procedure from [14] has been devised and implemented,[4] leading to a quantitative comparison of explained variances across eights audio representations and seventeen historical datasets. As a second major and original contribution, a framework to assess the robustness of the metric learning procedure against inter- and intra-subject variability in the human ratings dataset is proposed. This framework permits to quantify the superiority of deep embeddings in terms of both explained variance and robustness to degraded ratings.

Deep learning representations, especially CLAP and MERTCAT, are very promising for addressing a wealth of open questions in auditory cognitive neuroscience, beyond timbre perception. The understanding of speech [34], environmental sounds [35] listening, and even animal bioacoustics [36], [37], draws very appealing lines of research leveraging the high explanatory power and robustness of deep representations-based metric learning.

ACKNOWLEDGMENT

This project started with the engineer project research of Q. Banet and A. Berthier who thoroughly studied the codes[1] from [14] and pointed out the missing term in the reward function gradient.

REFERENCES

[1] N. Huang, M. Slaney, and M. Elhilali, "Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals," *Front. Neurosci.*, vol. 12, p. 532, 2018.

[2] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[3] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *J. Acoustical Soc. Am.*, vol. 63, no. 5, pp. 1493–1500, 1978.

[4] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, p. e1002759, 2012.

[5] K. Siedenburg, K. Jones-Mollerup, and S. McAdams, "Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds," *Front. Psychol.*, vol. 6, p. 1977, 2016.

[6] K. Siedenburg and S. McAdams, "Four distinctions for the auditory "wastebasket" of timbre," *Front. Psychol.*, vol. 8, p. 1747, 2017.

[7] M. Ogg and L. R. Slevc, "Acoustic correlates of auditory object and event perception: Speakers, musical timbres, and environmental sounds," *Front. Psychol.*, vol. 10, p. 1594, 2019.

[8] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoustical Soc. Am.*, vol. 61, no. 5, pp. 1270–1277, 1977.

[9] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbrea," *J. Acoustical Soc. Am.*, vol. 94, no. 5, pp. 2595–2603, 1993.

[10] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Res.*, vol. 58, pp. 177–192, 1995.

[11] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Percep. Psychophys.*, vol. 62, no. 7, pp. 1426–1439, 2000.

[12] M. Barthet, P. Guillemain, R. Kronland-Martinet, and S. Ystad, "From clarinet control to timbre perception," *Acta Acust. U. Acust.*, vol. 96, no. 4, pp. 678–689, 2010.

[13] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J. Acoustical Soc. Am.*, vol. 130, no. 5, pp. 2902–2916, 2011.

[14] E. Thoret, B. Caramiaux, P. Depalle, and S. Mcadams, "Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre," *Nat. Hum. Behav.*, vol. 5, no. 3, pp. 369–377, 2021.

[4]https://github.com/bpascal-fr/timbre-metric-learning/

[15] P. Aumond, A. Can, B. De Coensel, C. Ribeiro, D. Botteldooren, and C. Lavandier, "Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments," *Appl. Sci.*, vol. 7, no. 2, p. 144, 2017.

[16] T. Pethick, W. Xie, and V. Cevher, "Stable Nonconvex-Nonconcave Training via Linear Lnterpolation," *Adv. Neural Inf. Process.*, vol. 36, 2024.

[17] 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 *Proc. Int. Conf. Acoust., Speech Signal Process.* IEEE, 2017, pp. 131–135.

[18] A. Jansen, J. F. Gemmeke, D. P. Ellis, X. Liu, W. Lawrence, and D. Freedman, "Large-scale audio event discovery in one million youtube videos," in *Proc. Int. Conf. Acoust., Speech Signal Process.* IEEE, 2017, pp. 786–790.

[19] H. Han, V. Lostanlen, and M. Lagrange, "Perceptual–Neural–Physical Sound Matching," in *Proc. Int. Conf. Acoust., Speech Signal Process.* IEEE, 2023, pp. 1–5.

[20] P. Flandrin, *Time–frequency/time–scale analysis*. Academic press, 1998.

[21] S. Mallat, "Understanding deep convolutional networks," *Philos. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150203, 2016.

[22] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3704–3718, 2019.

[23] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoustical Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Preprint arXiv:1409.1556*, 2014.

[25] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," *Preprint arXiv:2311.01616*, 2023.

[26] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Preprint arXiv:2210.13438*, 2022.

[27] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2023.

[28] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training," *ICLR*, 2024.

[29] A. Bellet, A. Habrard, and M. Sebban, *Metric learning*. Springer Nature, 2022.

[30] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[31] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[32] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[33] R. Bro and S. De Jong, "A fast non-negativity-constrained least squares algorithm," *J. Chemom.*, vol. 11, no. 5, pp. 393–401, 1997.

[34] P. Albouy, L. Benjamin, B. Morillon, and R. J. Zatorre, "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," *Science*, vol. 367, no. 6481, pp. 1043–1047, 2020.

[35] I. Nelken and A. De Cheveigné, "An ear for statistics," *Nat. Neurosci.*, vol. 16, no. 4, pp. 381–382, 2013.

[36] M. R. Bregman, A. D. Patel, and T. Q. Gentner, "Songbirds use spectral shape, not pitch, for sound pattern recognition," *Proc. Natl Acad. Sci. USA*, vol. 113, no. 6, pp. 1666–1671, 2016.

[37] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin *et al.*, "Learning to detect an animal sound from five examples," *Ecol. Inform.*, vol. 77, p. 102258, 2023.