Explainable Audio Classification of Playing Techniques with Layer-wise Relevance Propagation

Changhong Wang, Vincent Lostanlen, Mathieu Lagrange

EXPLAINABLE AUDIO CLASSIFICATION OF PLAYING TECHNIQUES WITH LAYER-WISE RELEVANCE PROPAGATION

Changhong Wang[1], Vincent Lostanlen[2], and Mathieu Lagrange[2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

ABSTRACT

Deep convolutional networks (convnets) in the time–frequency domain can learn an accurate and fine-grained categorization of sounds. For example, in the context of music signal analysis, this categorization may correspond to a taxonomy of playing techniques: vibrato, tremolo, trill, and so forth. However, convnets lack an explicit connection with the neurophysiological underpinnings of musical timbre perception. In this article, we propose a data-driven approach to explain audio classification in terms of physical attributes in sound production. We borrow from current literature in "explainable AI" (XAI) to study the predictions of a convnet which achieves an almost perfect score on a challenging task: i.e., the classification of five comparable real-world playing techniques from 30 instruments spanning seven octaves. Mapping the signal into the carrier-modulation domain using scattering transform, we decompose the networks' predictions over this domain with layer-wise relevance propagation. We find that regions highly-relevant to the predictions localized around the physical attributes with which the playing techniques are performed.

***Index Terms*—** Layer-wise relevance propagation, scattering transform, playing technique recognition, music signal analysis.

1. INTRODUCTION

Our scientific understanding of sound production has grown considerably since the early years of computer music [1, 2]. For example, the mechanics of a piano can be simulated with enough precision so as to allow faithful synthesis [3]. However, for other instruments, we lack a closed-form description of the interaction between player and instrument [4]. Such is the case, in particular, when this interaction belongs to the "extended" vocabulary of playing techniques: tremolo, vibrato, staccato, and so forth [5]. This is because the gesture of the performer is more difficult to specify for extended techniques than the so-called "ordinary" technique [6].

Meanwhile, the renewed interest for machine learning in audio signal processing has advanced the state of the art in the task of playing technique classification [7]. This task is motivated by the key role of playing techniques in the computational analysis of musical performance [8]. Of particular interest is the subfamily of "periodic modulation techniques" (PMT), in which the musician alternates quickly between two positions: e.g., pressing and releasing a key to produce a trill [9]. In comparison with the ordinary technique, a PMT audio signal modulates periodically in amplitude, in frequency, or both.

For this reason, the scattering transform offers a judicious choice of feature map for PMT classification. Indeed, it represents the audio signal $\boldsymbol{x}$ in terms of a tensor $\mathbf{S}\boldsymbol{x}$ which is indexed by time t, first-order wavelet frequency λ_1, and second-order wavelet frequency λ_2 [10]. Prior research has proven an approximate closed-form expression for an idealized model of PMT, in which both the carrier and the modulator are sinusoids of respective frequencies ω_1 and ω_2 [11]. Under this idealization, the energy in $\mathbf{S}\boldsymbol{x}$ has a local maximum at the scattering path $(\lambda_1, \lambda_2) = (\omega_1, \omega_2)$. Yet, real-world PMT's involve non-sinusoidal carriers and modulators [12]. The corresponding $\mathbf{S}\boldsymbol{x}$ yields several nonzero regions in the scattering transform domain [13].

Passing the tensor $\mathbf{S}\boldsymbol{x}$ as input to a supervised classifier has led to state-of-the-art performance over several datasets for playing technique recognition [14]. It has also allowed to match subjective ratings of auditory similarities between playing techniques across different instruments and mutes [15]. Another strong tendency of recent research is to switch from shallow classifiers (e.g., support vector machines) to deep learning (e.g., convnets) [16, 17]. But despite the growth of data-driven approaches to musical acoustics [18], the perceptual underpinnings of playing technique recognition remains poorly known. What makes the difference, for example, between a vibrato and a trill [19, 20]? Answering this kind of research question requires insight on the spectrotemporal characteristics of the audio signal at hand, and not simply an accurate classification.

In this article, we propose to characterize real-world playing techniques in terms of sparse activations in feature space. A prior publication has tackled this problem with unsupervised dictionary learning [21] over magnitude spectra. The originality of our approach is that it is supervised: rather than decomposing the tensor $\mathbf{S}\boldsymbol{x}$, it decomposes the prediction of a deep neural network $\boldsymbol{f}$ over the domain (λ_1, λ_2). The resulting decomposition does not measure which pairs (ω_1, ω_2) are *present* in $\boldsymbol{x}$; but more specifically, which are *relevant* to the value of $\boldsymbol{f}(\mathbf{S}\boldsymbol{x})$. We decompose the predictions using the *layer-wise relevance propagation* (LRP) method, which explains pre-trained models' predictions by associating each neuron a relevance score. Although LRP has led to many publications in image and text processing [22], it has been rarely applied to speech [23, 24] and never to music. Our main finding is that LRP aligns with current knowledge about sound production and musical gestures.

2. LAYER-WISE RELEVANCE PROPAGATION

2.1. Deep Taylor decomposition

We define a neural network $\boldsymbol{f}$ of depth M recursively over layers:

$$\boldsymbol{f}_m(\mathbf{S}\boldsymbol{x})[j] = \rho\left(\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\boldsymbol{f}_{m-1}(\mathbf{S}\boldsymbol{x})[i] + \boldsymbol{b}_m[j]\right), \quad (1)$$

Supported by an Atlanstic2020 project on Trainable Acoustic Sensors (TrAcS). Companion website: https://github.com/changhongw/examod.

where $\mathbf{W}_m$ and $\boldsymbol{b}_m$ are the matrix of weights and vector of bias in layer m, and ρ is a rectified linear unit (ReLU). i and j index neurons in layer $m-1$ and layer m, respectively. Let $\boldsymbol{y}_m$ denote the layer-wise output of the network, i.e. $\boldsymbol{y}_m = \boldsymbol{f}_m(\mathbf{S}\boldsymbol{x})$. The relevance at the deepest layer, which takes $\boldsymbol{y}_{M-1}$ as input, is defined as the prediction itself: $\mathbf{R}_M(\boldsymbol{y}_{M-1}) = \boldsymbol{f}_M(\mathbf{S}\boldsymbol{x})$. Our goal is to decompose $\mathbf{R}_M$ into shallower layers until reaching $\mathbf{R}_0$ at the level of the input $\boldsymbol{f}_0(\mathbf{S}\boldsymbol{x}) = \mathbf{S}\boldsymbol{x}$. We seek an LRP rule of the form:

$$\mathbf{R}_{m-1}(\boldsymbol{y}_{m-2})[i] = \sum_{j=1}^{N_m} \mathbf{L}_m(\boldsymbol{y}_{m-1})[i,j], \tag{2}$$

in which the link matrix $\mathbf{L}_m$ preserves total relevance:

$$\sum_{i=1}^{N_{m-1}} \mathbf{L}_m(\boldsymbol{y}_{m-1})[i,j] = \mathbf{R}_m(\boldsymbol{y}_{m-1})[j]. \tag{3}$$

Before identifying $\mathbf{L}_m$, we impose that relevance $\mathbf{R}_m$ and activation $\boldsymbol{y}_m$ should be proportional over each node j [25]:

$$\forall \boldsymbol{x},\ \mathbf{R}_m(\boldsymbol{y}_{m-1})[j] = \boldsymbol{c}_m[j]\boldsymbol{y}_m[j], \tag{4}$$

with $\boldsymbol{c}_m[j]$ an unknown proportionality factor. In this case, $\mathbf{R}_m(\boldsymbol{y}_{m-1})[j] = 0$ is equivalent to $\boldsymbol{y}_m[j] = 0$, which defines a plane in dimension $\mathbb{R}^{N_{m-1}}$ according to Eq. (1). We then define a search direction $\boldsymbol{d}_m \in \mathbb{R}^{N_{m-1}}$ and solve for a root point $\tilde{\boldsymbol{y}}_{m-1}$ from:

$$\begin{cases} \boldsymbol{c}_m[j]\rho\left(\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\tilde{\boldsymbol{y}}_{m-1}[i] + \boldsymbol{b}_m[j]\right) = 0, \\ \tilde{\boldsymbol{y}}_{m-1}[i] = \boldsymbol{y}_{m-1}[i] + \alpha_m \boldsymbol{d}_m[i], \end{cases} \tag{5}$$

where α_m is a scalar. We then obtain for every j:

$$\boldsymbol{y}_{m-1}[i] - \tilde{\boldsymbol{y}}_{m-1}[i] = \frac{\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\boldsymbol{y}_{m-1}[i] + \boldsymbol{b}_m[j]}{\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\boldsymbol{d}_m[i]}\boldsymbol{d}_m[i]. \tag{6}$$

To identify $\mathbf{L}_m$, we perform a *deep Taylor decomposition* [26], comprising a series of Taylor decompositions of the relevance recursively at each node. Specifically, we do a Taylor expansion of the function $\mathbf{R}_m$ at the root point $\tilde{\boldsymbol{y}}_{m-1}$:

$$\begin{aligned} \mathbf{R}_m(\boldsymbol{y}_{m-1})[j] = \sum_{i=1}^{N_{m-1}} \left.\frac{\partial \mathbf{R}_m[j]}{\partial \boldsymbol{y}_{m-1}[i]}\right|_{\boldsymbol{y}_{m-1}=\tilde{\boldsymbol{y}}_{m-1}} \left(\boldsymbol{y}_{m-1}[i] - \tilde{\boldsymbol{y}}_{m-1}[i]\right) \\ + O\left(\|\boldsymbol{y}_{m-1}[i] - \tilde{\boldsymbol{y}}_{m-1}[i]\|^2\right). \end{aligned} \tag{7}$$

Neglecting the higher-order terms in Eq.(7) and injecting Eq.(6) into Eq.(7), we obtain the base formula for deriving different LRP rules [26]:

$$\mathbf{R}_{m-1}(\boldsymbol{y}_{m-2})[i] = \sum_{j=1}^{N_m} \frac{\mathbf{W}_m[i,j]\boldsymbol{d}_m[i]}{\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\boldsymbol{d}_m[i]}\mathbf{R}_m(\boldsymbol{y}_{m-1})[j]. \tag{8}$$

2.2. Baseline rule: LRP-0

According to the simplest rule, known as LRP-0, the search direction is defined as $\boldsymbol{d}_m[i] = \boldsymbol{y}_{m-1}[i]$ [26]. Therefore, according to Eq. (8), we obtain the relevance redistribution formula:

$$\mathbf{R}_{m-1}(\boldsymbol{y}_{m-2})[i] = \sum_{j=1}^{N_m} \frac{\mathbf{W}_m[i,j]\boldsymbol{y}_{m-1}[i]}{\sum_{i=1}^{N_{m-1}} \mathbf{W}_m[i,j]\boldsymbol{y}_{m-1}[i]}\mathbf{R}_m(\boldsymbol{y}_{m-1})[j]. \tag{9}$$

To avoid division by zeros, *LRP-ε* rule adds a small positive number ε to the denominator of Eq. (9).

2.3. Advanced rule : LRP-$[\varepsilon, z^+]$

LRP-z^+ rule searches the nearest root point on the segment $(\{\boldsymbol{y}_{m-1}[i]1_{\mathbf{W}_{m[i,j]\leq 0}}\}, \{\boldsymbol{y}_{m-1}[i]\})$ and the name "z^+" originates from [26] which defined $z_{ij}^+ = \mathbf{W}_m^+[i,j]\boldsymbol{y}_{m-1}[i]$. LRP-$z^+$ considers only the contributions of positive weights $\mathbf{W}_m^+[i,j]$:

$$\mathbf{R}_{m-1}(\boldsymbol{y}_{m-2})[i] = \sum_{j=1}^{N_m} \frac{\mathbf{W}_m^+[i,j]\boldsymbol{y}_{m-1}[i]}{\sum_{i=1}^{N_{m-1}} \mathbf{W}_m^+[i,j]\boldsymbol{y}_{m-1}[i]}\mathbf{R}_m(\boldsymbol{y}_{m-1})[j]. \tag{10}$$

LRP-$[\varepsilon, z^+]$ is a composite rule which applies the LRP-ε rule for convolutional layers and the LRP-z^+ rule for fully connected layers. We refer to [25] for a complete list of propagation rules. We implement LRP in Python via the Zennit package[1] .

3. APPLICATION TO PLAYING TECHNIQUE CLASSIFICATION

3.1. Scattering transform

As a biological plausible surrogate for human perceptual judgments of isolated audio events [15], the scattering transform decomposes audio signals using wavelet convolutions, modulus nonlinearities, and average pooling. The first-order scattering transform $\mathbf{S_1}\boldsymbol{x}$ maps the signal $\boldsymbol{x}$ into the time–frequency domain, by convolving it with a wavelet filterbank $\boldsymbol{\psi}_{\lambda_1}$, taking modulus and averaging with a lowpass filter $\boldsymbol{\phi}$:

$$\mathbf{S_1}\boldsymbol{x}(t,\lambda_1) = \left(\left|\boldsymbol{x} * \boldsymbol{\psi}_{\lambda_1}\right| * \phi\right)(t). \tag{11}$$

$\mathbf{S_1}\boldsymbol{x}$ is essentially a constant-Q transform (CQT) which is a commonly-used representation for music signal analysis [16, 17]. Yet, the averaging loses temporal modulations, which are critical to the discrimination of PMTs. To recover this information, we perform a second-order decomposition of the unaveraged $\mathbf{S_1}\boldsymbol{x}$ with a wavelet filterbank $\boldsymbol{\psi}_{\lambda_2}$ [10]:

$$\mathbf{S_2}\boldsymbol{x}(t,\lambda_1,\lambda_2) = \left(\left|\left|\boldsymbol{x} * \boldsymbol{\psi}_{\lambda_1}\right| * \boldsymbol{\psi}_{\lambda_2}\right| * \phi\right)(t). \tag{12}$$

We use $\mathbf{S}\boldsymbol{x} = \mathbf{S_1}\boldsymbol{x} + \mathbf{S_2}\boldsymbol{x}$ as input to a convenet for playing technique classification. Backpropagating the predictions using the LRP rules in Section 2, we obtain the relevance score $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})[t,\lambda_1,\lambda_2]$, which shows the contribution of each element in $\mathbf{S}\boldsymbol{x}$.

3.2. Deep convolutional network

We train a convnet with 3 convolutional layers and one dense layer. Each convolutional layer comprises a one-dimensional convolution unit, a batch norm, a ReLU, and an average pooling. The dense layer is followed by a softmax unit. The input to the convnet is the tensor of scattering coefficients, either $\mathbf{S_1}\boldsymbol{x}$ or $\mathbf{S}\boldsymbol{x}$. The corresponding feature dimensions are 74 and 1200; and the number of trainable parameters are 10.3 K and 2.1 M. For both cases, the network is trained with early stopping and a batch size of 32. We use weighted cross-entropy loss due to the unbalanced classes.

3.3. Studio On Line dataset

We use a subset of the Studio On Line dataset (version 0.9HQ) [7] that includes five types of PMTs: tremolo, flatterzunge, trill, bisbigliando, and vibrato. We call this subset *SOL-PMT*, which contains 2530,

[1] https://github.com/chr5tphr/zennit

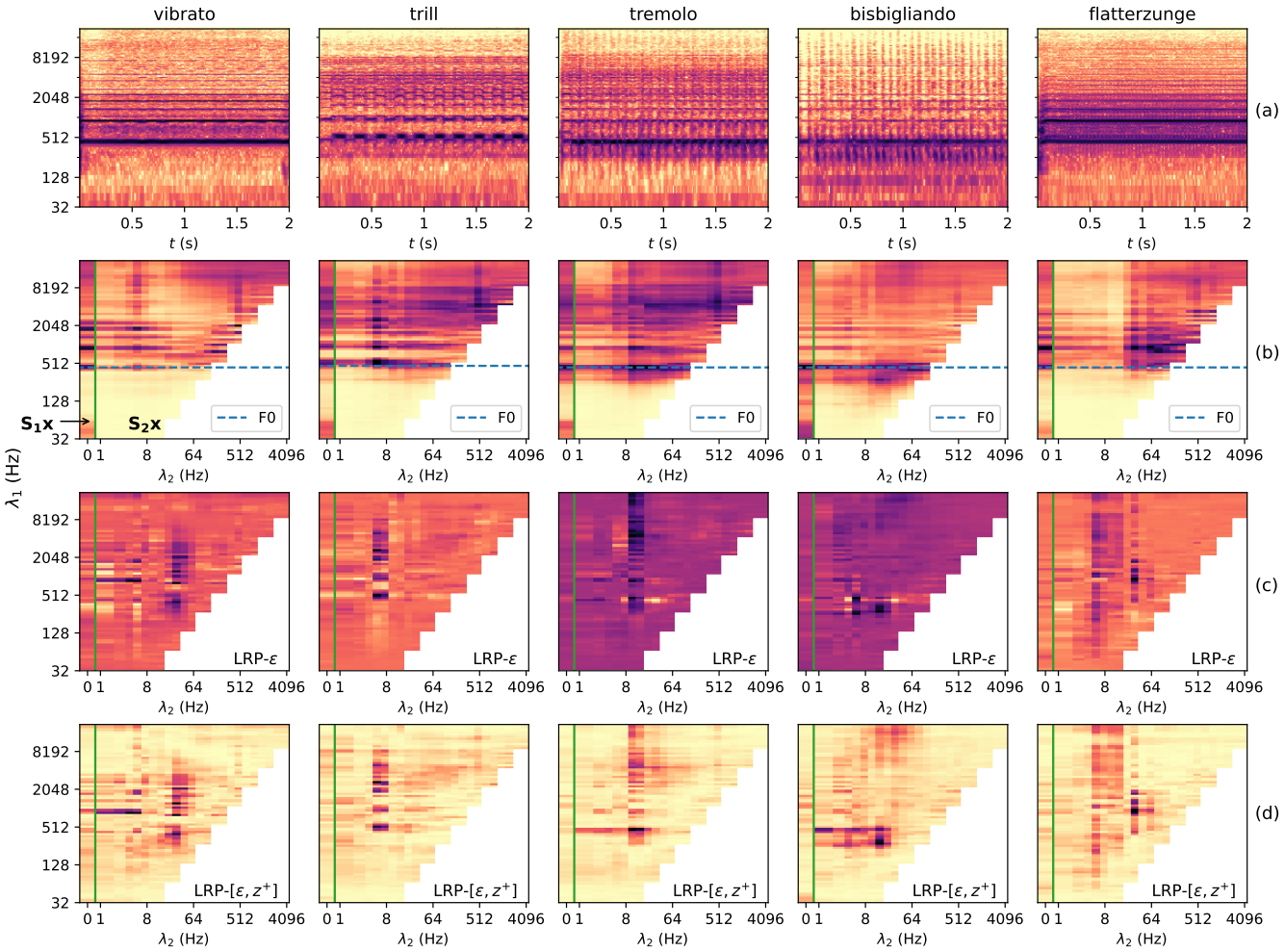

Fig. 1. Local relevance map $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ of five periodic modulation techniques. (a): log-spectrogram; (b): scattering coefficients $\mathbf{S}\boldsymbol{x} = \mathbf{S_1}\boldsymbol{x}+\mathbf{S_2}\boldsymbol{x}$ (F0=fundamental frequency); (c): $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ yielded from LRP-ε rule; and (d) $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ using LRP-$[\varepsilon, z^+]$ rule. λ_1 and λ_2 are the carrier and modulation frequency, respectively.

1523, 1035, 286, and 190 examples, respectively, for the five classes. SOL-PMT offers large intra-class variability for PMTs in terms of instrument (11 instrument families and 30 types of instruments), pitch (C♯1-34.6 Hz to B7-3951.1 Hz), dynamics (pianissimo to fortissimo), and with/without mute. The sampling rate of the dataset is 44.1 kHz.

4. RESULTS AND DISCUSSION

4.1. Evaluation

We extract the scattering features for the SOL-PMT dataset with 8 and 2 filters per octave in the first- and second-order. The averaging scale of the lowpass filter is 2^{13}, resulting in a frame size of $T = 2^{13}$ (186 ms). Coefficients with carrier frequencies below 32 Hz are removed as those coefficients represents spectro-modulations that are inaudible. The disparate length of audio examples are fixed into 2^{18} samples (around 6 seconds) by truncating or zero-padding. The full feature map $\mathbf{S}\boldsymbol{x}$ for each audio example is then sized 1200×32, where 1200 is the feature dimension and 32 is the number of time frames. After randomly shuffling the data, we split each playing technique class into training, validation, and test subsets by a 6:2:2 ratio for each instrument. We provide a full description of the split and the file IDs on the companion website.

Table 1 lists the classification accuracy for each PMT class. The nearly perfect scores demonstrate the effectiveness of $\mathbf{S}\boldsymbol{x}$ for PMT recognition. The considerable performance drop after removing $\mathbf{S_2}\boldsymbol{x}$ from the full feature map verifies the importance of $\mathbf{S_2}\boldsymbol{x}$ for the discrimination of PMTs.

	Vibrato	Trill	Tremolo	Bisbigliando	Flatterzunge
$\mathbf{S_1}\boldsymbol{x}$	72.50	70.33	92.23	88.33	79.61
$\mathbf{S_1}\boldsymbol{x}+\mathbf{S_2}\boldsymbol{x}$	97.50	98.56	99.80	98.33	100.0

Table 1. Classification accuracy (%) for each playing technique class using the first-order ($\mathbf{S_1}\boldsymbol{x}$) and full ($\mathbf{S}\boldsymbol{x} = \mathbf{S_1}\boldsymbol{x} + \mathbf{S_2}\boldsymbol{x}$) scattering coefficients.

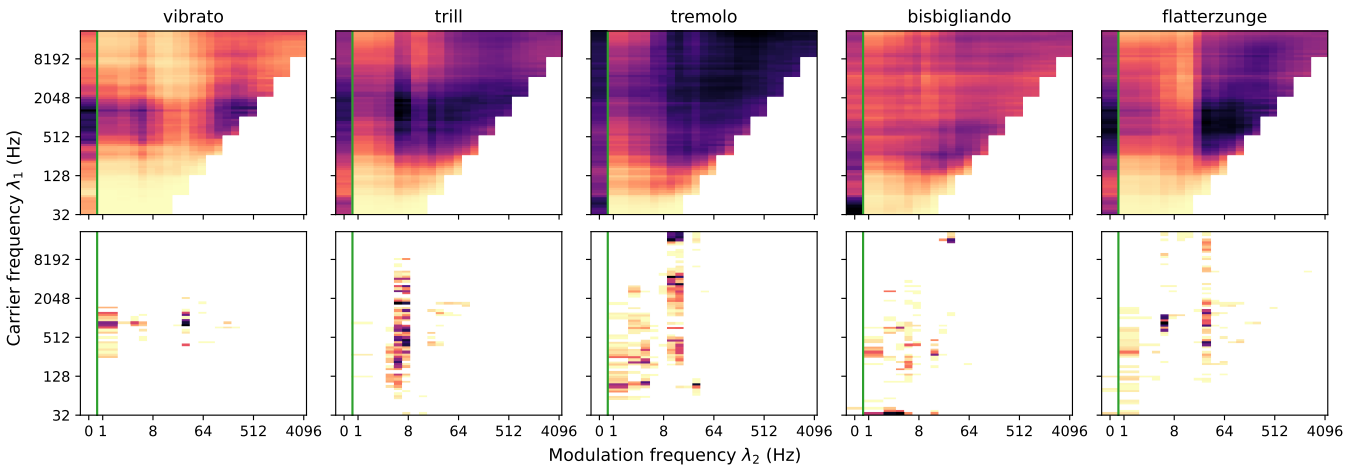

Fig. 2. Class-wise aggregated relevance maps for the five periodic modulation techniques in the test set. Top: averaged scattering coefficients $\mathbf{S}\boldsymbol{x} = \mathbf{S_1}\boldsymbol{x} + \mathbf{S_2}\boldsymbol{x}$; bottom: top-5 argmax aggregated relevance value $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$.

4.2. Local relevance maps

Decomposing the predictions $\boldsymbol{f}(\mathbf{S}\boldsymbol{x})$ to the input $\mathbf{S}\boldsymbol{x}$, we obtain 1200×32 relevance values $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ for each test audio. To obtain example-wise (local) relevance maps, we average the relevance over the 32 time frames and visualize it in terms of carrier-modulation pair, i.e. (λ_1, λ_2). Fig. 1 displays the local explanation maps of five test examples, each from one class at the same pitch A4=440 Hz (except A#4=466 Hz for trill). (a) is the log-spectrogram showing the spectro-temporal characteristics of each technique, followed by their corresponding $\mathbf{S}\boldsymbol{x}$ visualizations in (b). The column ticked by 0 in each subfigure in (b) is $\mathbf{S_1}\boldsymbol{x}$ and the remaining colored region is $\mathbf{S_2}\boldsymbol{x}$, as annotated in the left subfigure. Those regions in (c) and (d) correspond to the relevance values for $\mathbf{S_1}\boldsymbol{x}$ and $\mathbf{S_2}\boldsymbol{x}$, respectively.

Fig. 1 (c) and (d) are the relevance maps $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ for each class by applying the LRP-ε and LRP-$[\varepsilon, z^+]$ rule, respectively. A core observation is that relevance scores are localized around the modulation rate of the playing techniques, e.g. 8 Hz for trill and 32 Hz for flatterzunge. Indeed, these are the physical attributes with which the playing techniques are performed according to our knowledge of music gestures. Additionally, relevance values do not positively correlate with scattering energy. High energy regions in $\mathbf{S}\boldsymbol{x}$, e.g. with $\lambda_2 > 64$ Hz for trill, tremolo, and flatterzunge, do not show strong evidence in the corresponding relevance maps.

Another finding is that $\mathbf{S_1}\boldsymbol{x}$, equivalent to CQT, exhibits low relevance values for all PMTs as compared to $\mathbf{S_2}\boldsymbol{x}$. Although $\mathbf{S_1}\boldsymbol{x}$ is conceptually equivalent to many popular audio front ends in deep learning considered for a wide range of tasks and indeed shows a large amount of energy (see Fig. 1 (b) and Fig. 2 top), our experiments show that it has few impact on the decisions of a classifier that jointly considers $\mathbf{S_1}\boldsymbol{x}$ and $\mathbf{S_2}\boldsymbol{x}$ for PMT classification.

Comparing $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ derived from (c) LRP-ε and (d) LRP-$[\varepsilon, z^+]$ for each playing technique class, we notice that the latter provides more contrasted relevance values. To quantify the effect of LRP rules, we calculate the kurtosis of $\mathbf{R}_0(\mathbf{S}\boldsymbol{x})$ for each class over its test examples. The mean kurtosis obtained from the LRP-ε and LRP-$[\varepsilon, z^+]$ rule are 13.82 and 19.22, respectively; and the corresponding standard deviation are 8.81 and 13.72. The higher values from the LRP-$[\varepsilon, z^+]$ rule verify our observations in Fig. 1 (c) and (d). Therefore we use LRP-$[\varepsilon, z^+]$ rule for class-wise relevance aggregation.

4.3. Class-wise aggregation

We propose to derive class-wise explanations by aggregating local relevance maps in the test set. For a specific class, we first register the locations of the top-n maximal values of each local relevance map $\mathbf{R}_0^k(\mathbf{S}\boldsymbol{x})$:

$$P_k(n) = \{(\lambda_1, \lambda_2)\}_n = \arg\max^n_{\lambda_1,\lambda_2} \mathbf{R}_0^k(\mathbf{S}\boldsymbol{x}), \qquad (13)$$

where $k = 1, ..., K$ indices the test examples of this class. Let $\boldsymbol{I}_k$ be a $\lambda_1 \times \lambda_2$ matrix where $\boldsymbol{I}_{k(\lambda_1,\lambda_2)\in P_k} = 1$ and $\boldsymbol{I}_{k(\lambda_1,\lambda_2)\notin P_k} = 0$. Summing $\boldsymbol{I}_k$ over the test examples derives the class-wise aggregated map: $\mathbf{R}_0(\mathbf{S}\boldsymbol{x}) = \sum_{k=1}^K \boldsymbol{I}_k$.

Fig. 2 bottom shows the top-5 argmax, i.e. $n = 5$ in Eq. (13), aggregated relevance maps for the five PMT classes. To show the corresponding input, we display the averaged scattering coefficients over the test examples for each class (see Fig. 2 top). Similarly to Fig. 1, the left column in each subfigure corresponds to $\mathbf{S_1}\boldsymbol{x}$ and the remaining colored region is $\mathbf{S_2}\boldsymbol{x}$. The class-aggregated relevance maps further support our findings from the local maps in Section 4.2. The relevance values are more localized and globally structured vertically with high values at the modulation rate of the playing techniques across pitch. This means that the convnet successfully enforces the pitch invariance that is needed for the task at hand. $\mathbf{S_1}\boldsymbol{x}$ almost shows no relevance to the prediction as compared to $\mathbf{S_2}\boldsymbol{x}$. For a given class like vibrato, low energy regions in $\mathbf{S_2}\boldsymbol{x}$ exhibit high evidence, probably because they are discriminative to the other PMTs.

5. CONCLUSION

We propose a framework to explicitly connect networks' predictions to the physical attributes of audio signals. This is achieved by mapping the signal into a carrier-modulation domain using scattering transform, a surrogate of auditory perception. We then decompose the predictions of a convnet trained for playing technique classification to this domain using the layer-wise relevance propagation method. Our findings show that highly relevant regions are localized around the modulation rates of playing techniques, regardless of pitch. This explicit connection between networks' predictions and physical attributes of audio signals, fully data-driven, opens new avenues for sound production and music gesture analysis.

6. REFERENCES

[1] Max V Mathews, Joan E Miller, John R Pierce, and James Tenney, "Computer study of violin tones," *The Journal of the Acoustical Society of America*, vol. 38, no. 5, pp. 912–913, 1965.

[2] Jean-Claude Risset, "Computer study of trumpet tones," *The Journal of the Acoustical Society of America*, vol. 38, no. 5, pp. 912–912, 1965.

[3] Juliette Chabassier, Antoine Chaigne, and Patrick Joly, "Modeling and simulation of a grand piano," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 648–665, 2013.

[4] Nicholas Giordano and Vasileios Chatziioannou, "Status and future of modeling of musical instruments: Introduction to the JASA special issue," *The Journal of the Acoustical Society of America*, vol. 150, no. 3, pp. 2294–2301, 2021.

[5] Stefan Kostka and Matthew Santa, *Materials and techniques of post-tonal music*, Routledge, 2018.

[6] Rolf Inge Godøy and Marc Leman, *Musical gestures: Sound, movement, and meaning*, Routledge, 2010.

[7] Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange, "Extended playing techniques: The next milestone in musical instrument recognition," in *Proceedings of the International Conference on Digital Libraries for Musicology*, 2018, pp. 1–10.

[8] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani, "An interdisciplinary review of music performance analysis," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 221–246, 2020.

[9] Hiroshi Kinoshita and Satoshi Obata, "Left hand finger force in violin playing: Tempo, loudness, and finger differences," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 388–395, 2009.

[10] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[11] Vincent Lostanlen, Alice Cohen-Hadria, and Juan Pablo Bello, "One or two frequencies? The scattering transform answers," in *Proceedings of the IEEE European Signal Processing Conference (EUSIPCO)*, 2021, pp. 2205–2209.

[12] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew, "Adaptive time-frequency scaterring for periodic modulation recognition in music signals," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[13] Joakim Andén and Stéphane Mallat, "Scattering representation of modulated sounds," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2012, vol. 9, pp. 17–21.

[14] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew, "Adaptive scattering transforms for playing technique recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1407–1421, 2022.

[15] Vincent Lostanlen, Christian El-Hajj, Mathias Rossignol, Grégoire Lafay, Joakim Andén, and Mathieu Lagrange, "Time–frequency scattering accurately models auditory similarities between instrumental playing techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–21, 2021.

[16] Jean-Francois Ducher and Philippe Esling, "Folded CQT RCNN for real-time recognition of instrument playing techniques," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, 2019.

[17] Yu-Fen Huang, Jeng-I Liang, I-Chieh Wei, and Li Su, "Joint analysis of mode and playing technique in guqin performance with machine learning," in *Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference*, 2020, pp. 85–92.

[18] Marco Olivieri, Raffaele Malvermi, Mirco Pezzoli, Massimiliano Zanoni, Sebastian Gonzalez, Fabio Antonacci, and Augusto Sarti, "Audio information retrieval and musical acoustics," *IEEE Instrumentation & Measurement Magazine*, vol. 24, no. 7, pp. 10–20, 2021.

[19] Jean Hakes, Thomas Shipp, and E Thomas Doherty, "Acoustic characteristics of vocal oscillations: vibrato, exaggerated vibrato, trill, and trillo," *Journal of Voice*, vol. 1, no. 4, pp. 326–331, 1988.

[20] M Castellengo, "Fusion or separation: From vibrato to vocal trill," in *Proceedings of the Stockholm Music Acoustics Conference*, 1993.

[21] Li Su, Hsin-Ming Lin, and Yi-Hsuan Yang, "Sparse modeling of magnitude and phase-derived spectra for playing technique classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2122–2132, 2014.

[22] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[23] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *arXiv preprint arXiv:1807.03418*, 2018.

[24] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[25] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[26] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.