Classification with Joint Time-Frequency Scattering

Joakim Andén, Vincent Lostanlen, and Stéphane Mallat

***Abstract*—In time series classification, signals are typically mapped into some intermediate representation which is used to construct models. We introduce the joint time-frequency scattering transform, a locally time-shift invariant representation which characterizes the multiscale energy distribution of a signal in time and frequency. It is computed through wavelet convolutions and modulus non-linearities and may therefore be implemented as a deep convolutional neural network whose filters are not learned but calculated from wavelets. We consider the progression from mel-spectrograms to time scattering and joint time-frequency scattering transforms, illustrating the relationship between increased discriminability and refinements of convolutional network architectures. The suitability of the joint time-frequency scattering transform for characterizing time series is demonstrated through applications to chirp signals and audio synthesis experiments. The proposed transform also obtains state-of-the-art results on several audio classification tasks, outperforming time scattering transforms and achieving accuracies comparable to those of fully learned networks.**

***Index Terms*—Acoustic signal processing, continuous wavelet transform, convolutional neural networks, supervised learning.**

I. Introduction

Signals are typically classified by first mapping them into a lower-dimensional representation space where models may be generated. The suitability of these representations depends on their ability to capture signal structure relevant to the classification task. For time series, this often includes the signal's time-frequency geometry. Figure 1 shows a time-frequency decomposition, the wavelet transform, applied to two audio recordings. Both are recordings of a person laughing, so their time-frequency structure is similar, but they also exhibit significant variability. A good representation is invariant to such variability while capturing discriminative time-frequency geometry. This work studies the construction of such invariants.

Initial work on audio classification computed representations from time-frequency decompositions, such windowed Fourier transforms. These include mel-spectrograms, mel-frequency cepstral coefficients (MFCCs) [1], modulation spectrograms [2], [3] and correlograms [4], [5]. More recent work employs deep convolutional neural networks—cascades of filter banks alternated with nonlinearities [6], [7]. Filters are learned from data, so each network is adapted to a given task, often resulting in excellent performance [8]. However, learning typically requires large training sets and extensive computational resources.

This work is supported by the ERC InvariantClass 320959.

J. Andén is with the Flatiron Institute, New York, NY, USA (e-mail: janden@flatironinstitute.org).

V. Lostanlen is with the Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA and the Music and Audio Research Laboratory, New York University, New York, NY, USA (e-mail: vincent.lostanlen@nyu.edu).

S. Mallat is with the Département dInformatique, Ecole Normale Supérieure, Paris, France (e-mail: mallat@di.ens.fr).

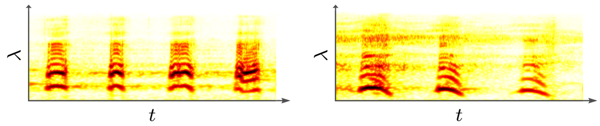


Fig. 1: The wavelet transform amplitudes, or scalograms, of two recordings as a function of time t and log-frequency λ. Both recordings are of one person laughing.

This work provides a bridge between traditional time-frequency representations and deep convolutional neural networks. In particular, we implement the mel-spectrogram as a convolutional network and increase its discriminability by adding certain filters to that network. These filters are not learned but fixed according to the invariance and discriminability needs of the task. This simplifies analysis and interpretation of the network. Fixed filters also reduces the computational burden of the representation as no training is necessary.

A convolutional network cascades convolutions, subsampling operators, and pointwise nonlinearities (such as rectifiers) [9], [10]. Its convolution kernels, or filters, are optimized over a training set. Section II-A describes how the wavelet transform is computed by a similar cascade of convolutions, but with fixed filters. A wavelet transform is thus a convolutional network with filters specified by certain time-frequency topology.

Many time series classification tasks are invariant to small time shifts and time-warping deformations. Indeed, class membership is typically unaffected by these transformations. A good representation is therefore invariant to such changes. In Section II-B, the wavelet transform is rendered invariant by computing its modulus, known as the scalogram, and averaging in time. This yields a variant of the popular mel-spectrogram.

Although powerful, mel-spectrograms do not capture large-scale temporal structure, such as amplitude modulation. In Section II-C, the time scattering transform extends the mel-spectrogram through multiscale modulation coefficients [11], [12]. Instead of averaging the scalogram, it applies a second wavelet transform in time, takes the modulus, and averages. This representation is more discriminative and performs well for several classification tasks [12], [13], [14], [15]. Extending the wavelet transform network now lets us implement both mel-spectrograms and time scattering as convolutional networks.

A significant limitation of the time scattering transform is its restriction to convolutions along the time axis. In other words, its convolutional network is actually a tree, with each node having only a single parent. A consequence is that time scattering cannot separate signals subjected to time shifts which vary in frequency, which is shown in Section III-A. To remedy this, we must capture time and frequency structure jointly.

With this goal in mind, we introduce the joint time-frequency scattering transform. As described in Section III-B, it replaces the one-dimensional, channel-by-channel wavelet decomposition of the scalogram by a two-dimensional wavelet transform. Its construction is inspired by the cortical transform of Shamma et al. [16], [17], which provides neurophysiological models of auditory processing in the mammalian brain. The corresponding joint scattering network introduces additional filters into time scattering network, breaking its tree structure and increasing its discriminative power. To illustrate this, Section III-C shows how the joint scattering transform captures the chirp rate of frequency-modulated excitations.

The representative ability of the proposed transform is further demonstrated in Section IV through synthesis experiments. Here, a signal is synthesized from a target scattering transform by minimizing the distance of its transform to that target. The resulting synthesized signals show how certain structures which are not captured by the mel-spectrogram and time scattering are better characterized by the joint scattering transform.

Section V concludes by evaluating the joint time-frequency scattering transform on several audio classification tasks. These include classification of phone segments, musical instruments, and acoustic scenes. The joint transform outperforms the mel-spectrogram and time scattering while achieving results comparable to, or better than, state-of-the-art convolutional networks. All figures and tables may be reproduced using software available at http://www.di.ens.fr/data/software/.

II. Time-Shift Invariant Representations

Section II-A defines the wavelet transform, a representation well suited for time series with multiscale structure. The modulus of the wavelet transform, known as the scalogram, is averaged in time to yield the time-shift invariant mel-spectrogram, as described in Section II-B. Section II-C introduces the time scattering transform, which extends the invariant mel-spectrogram. Instead of just averaging the scalogram, it also applies a second wavelet transform, demodulates, and averages the result in time. These representations are cascades of convolutions and non-linearities and may thus be implemented as deep convolutional networks with fixed filters.

A. Wavelet Transform Filter Bank

The wavelet transform of a signal is obtained by convolving it with a set of dilated bandpass filters known as wavelets. It captures both short, transient structures and long-range oscillations in a localized manner. In the frequency domain, the ratio between center frequency and bandwidth, the Q factor, is the same for all filters. These transforms are therefore constant-Q transforms [18]. Wavelet filter banks provide good models for cochlear function in mammals [19], [16], [17] and form the basis for many audio representations [20]. The transform may be computed using a multirate filter bank [20].

Let $\boldsymbol{x}(t)$ be a continuous signal for $t \in \mathbb{R}$. Its Fourier transform is given by $\hat{\boldsymbol{x}}(\omega) = \int_{\mathbb{R}} \boldsymbol{x}(t) \mathrm{e}^{-2\pi \mathrm{i} \omega t}\, \mathrm{d}t$ for $\omega \in \mathbb{R}$. We consider a complex analytic wavelet $\boldsymbol{\psi}(t)$ whose Fourier transform $\widehat{\boldsymbol{\psi}}(\omega)$ is concentrated in the frequency interval $[2^{-1/Q}, 1]$ for some $Q \geq 1$. Dilating $\boldsymbol{\psi}(t)$ by factors $2^{-\lambda}$ now yields the wavelet filter bank

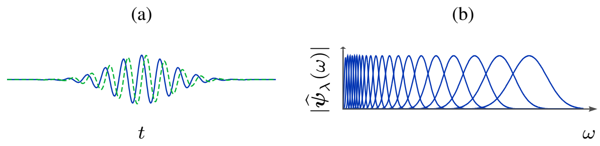


Fig. 2: (a) Real and imaginary parts of a Morlet mother wavelet with $Q = 4$. (b) The resulting wavelet filter bank in Fourier.

$$\boldsymbol{\psi}_\lambda(t) = 2^\lambda \boldsymbol{\psi}(2^\lambda t) \quad \Longleftrightarrow \quad \widehat{\boldsymbol{\psi}}_\lambda(\omega) = \widehat{\boldsymbol{\psi}}(2^{-\lambda}\omega)\ , \tag{1}$$

for $\lambda \in \mathbb{R}$. Consequently, $\widehat{\boldsymbol{\psi}}_\lambda(\omega)$ is concentrated in $[2^{\lambda-1/Q}, 2^\lambda]$. This interval has approximate center 2^λ and bandwidth $2^\lambda/Q$. We therefore need Q filters to cover an octave, independent of frequency. Since $\widehat{\boldsymbol{\psi}}_\lambda(\omega)$ is concentrated around 2^λ, we refer to λ as the wavelet's log-frequency index.

We are typically interested only in structures shorter than some fixed time scale T. In time, $\boldsymbol{\psi}_\lambda(t)$ has approximate duration $2^{-\lambda}Q$. We therefore require λ to satisfy $2^{-\lambda}Q \leq T$. Unfortunately, certain low frequencies are then not covered by any wavelet. For audio signals, these frequencies typically contain a small amount of energy and may be safely ignored. In the following, we instead add a set of constant-bandwidth filters covering these frequencies (see Andén & Mallat [12]).

In numerical experiments, we use the Morlet wavelet due to its near-optimal time-frequency localization [20], [12]. Figure 2 shows a sample Morlet wavelet and its wavelet filter bank.

We now define the continuous wavelet transform of $\boldsymbol{x}(t)$ as

$$\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t) \tag{2}$$

for λ such that $2^{-\lambda}Q \leq T$. It captures the local oscillations of $\boldsymbol{x}$ at time t and frequency 2^λ with resolution $2^{-\lambda}Q$ and $2^\lambda/Q$ in time and frequency, respectively. In audio applications, we typically set $Q \approx 8$ to better resolve oscillatory components.

Now let $\boldsymbol{x}[n]$ be a discrete signal for $n \in \mathbb{Z}$. Its discrete-time Fourier transform is $\hat{\boldsymbol{x}}(\omega) = \sum_{n \in \mathbb{Z}} \boldsymbol{x}[n]\, \mathrm{e}^{-2\pi \mathrm{i} t \omega}$ for $\omega \in [-1/2, 1/2]$. We now define a discrete analog of the continuous wavelet transform (2), implemented as a multirate filter bank.

To achieve this, we consider the multiresolution pyramid obtained by averaging $\boldsymbol{x}[n]$ at different scales 2^j. We initialize the finest scale to $\boldsymbol{a}_0[n] = \boldsymbol{x}[n]$. For $j > 0$, $\boldsymbol{a}_j[n]$ is obtained from $\boldsymbol{a}_{j-1}[n]$ through convolution by a lowpass filter $\boldsymbol{h}[n]$ whose transfer function $\hat{\boldsymbol{h}}(\omega)$ is concentrated in $[-1/4, 1/4]$. We then subsample by 2 to obtain

$$\boldsymbol{a}_j[n] = \boldsymbol{a}_{j-1} * \boldsymbol{h}[2n]\ . \tag{3}$$

Note that $\boldsymbol{a}_j[n] = \boldsymbol{x} * \boldsymbol{h}_j[2^j n]$ for some filter $\boldsymbol{h}_j[n]$ defined by

$$\hat{\boldsymbol{h}}_j(\omega) = \prod_{p=0}^{j-1} \hat{\boldsymbol{h}}(2^p \omega).$$

As a result, $\hat{\boldsymbol{h}}_j(\omega)$ is concentrated in $[-2^{-j-1}, 2^{-j-1}]$ and $\boldsymbol{h}_j[n]$ has approximate duration 2^{j+1}.

The high frequencies of $\boldsymbol{a}_{j-1}[n]$ lost when convolving with $\boldsymbol{h}[n]$ are captured by Q bandpass filters $\boldsymbol{g}_0[n], \ldots, \boldsymbol{g}_{Q-1}[n]$.

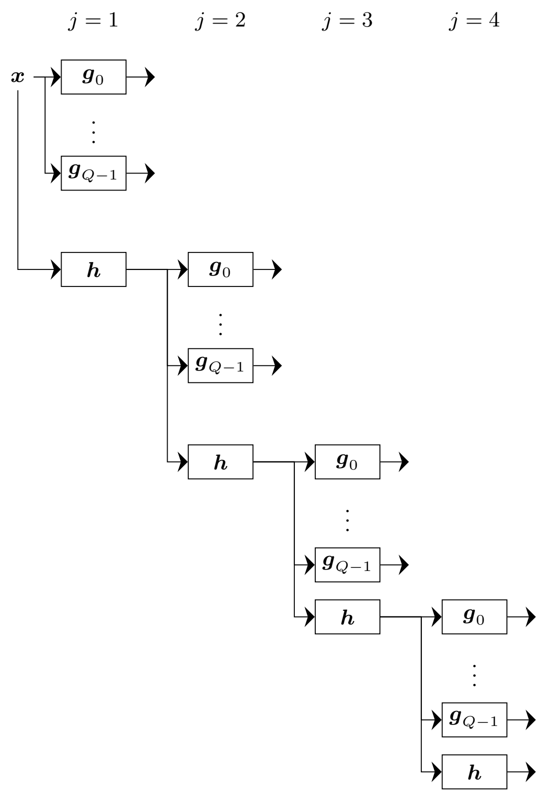


Fig. 3: Multirate filter bank computing wavelet coefficients for $J = 4$. Each block corresponds to a filter convolution subsampled by 2 where a boxed $\boldsymbol{h}$ is a low-pass filter and a boxed $\boldsymbol{g}_k$ is a band-pass filter. The depth corresponds to the octave index j while $k = 0, \ldots, Q-1$ is the suboctave index.

Each has a transfer function $\hat{\boldsymbol{g}}_k(\omega)$ concentrated in $[2^{-(k+1)/Q-1}, 2^{-k/Q-1}]$. After convolving $\boldsymbol{a}_{j-1}[n]$ with $\boldsymbol{g}_k[n]$, the result is subsampled by 2, yielding

$$\boldsymbol{d}_{j,k}[n] = \boldsymbol{a}_{j-1} * \boldsymbol{g}_k[2n], \tag{4}$$

for $j > 0$ and $0 \leq k < Q$. One may verify that

$$\boldsymbol{d}_{j,k}[n] = \boldsymbol{x} * \boldsymbol{g}_{j,k}[2^j n], \tag{5}$$

where $\hat{\boldsymbol{g}}_{j,k}(\omega) = \hat{\boldsymbol{h}}_{j-1}(\omega)\,\hat{\boldsymbol{g}}_k(2^{j-1}\omega)$. These filters are concentrated in intervals $[2^{-j-(k+1)/Q}, 2^{-j-k/Q}]$. In time, they have approximate duration $2^j Q$. Since we are only concerned with local variability below time scale T, we require $2^j Q \leq T$. This specifies the maximum depth $J = \log_2(T/Q)$ of the cascade.

Figure 3 illustrates this multirate filterbank cascade. Each box corresponds to a convolution and subsampling by 2 according to (3) or (4). First, $\boldsymbol{x}[n]$ is convolved with $\boldsymbol{g}_0[n], \ldots, \boldsymbol{g}_{Q-1}[n]$ and subsampled to yield the highest octave of bandpass coefficients $\boldsymbol{d}_{1,0}[n], \ldots, \boldsymbol{d}_{1,Q-1}[n]$. Convolving $\boldsymbol{x}[n]$ with $\boldsymbol{h}[n]$ and subsampling provides the remaining low frequencies, and the process is repeated. As we progress through this cascade, the depth corresponds to the octave index j.

Combining the bandpass outputs yields the discrete wavelet transform in (5) for $1 \leq j \leq J$ and $0 \leq k < Q$. This is similar to the output of the continuous wavelet transform. Indeed, if we sample a continuous band-limited signal $\boldsymbol{x}(t)$ at unit intervals, its discrete wavelet transform (5) approximates the continuous transform (2) for $\lambda = -j - k/Q \leq -1$ provided that $\hat{\boldsymbol{g}}_{j,k}(\omega) \approx \boldsymbol{\psi}_\lambda(\omega)$. Given the mother wavelet $\boldsymbol{\psi}(t)$, it is possible to construct filters $\boldsymbol{h}[n]$ and $\boldsymbol{g}_0[n], \ldots, \boldsymbol{g}_{Q-1}[n]$ such that this correspondence holds for large j [20]. The result is an approximation of the continuous wavelet transform using the convolutional network illustrated in Figure 3.

B. Mel-Spectrogram

The lack of time-shift invariance of the wavelet transform hinders its generalization power for classification. For most classification tasks, shifting a signal in time does not modify its class. To reduce variability when constructing models, the signal representation must therefore be made time-shift invariant.

The amplitude of the wavelet transform is the scalogram:

$$\mathbf{X}(t, \lambda) = |\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t)|. \tag{6}$$

Figure 1 shows two sample scalograms. Since the wavelets are analytic, applying the complex modulus performs a Hilbert demodulation, capturing the temporal envelope of each subband. The scalogram $\mathbf{X}(t, \lambda)$ therefore describes the time-frequency intensity of $\boldsymbol{x}(t)$ at time t and log-frequency λ.

Unfortunately, the scalogram is not time-shift invariant. Indeed, shifting a signal $\boldsymbol{x}(t) \mapsto \boldsymbol{x}_c(t) = \boldsymbol{x}(t-c)$ also shifts its scalogram $\mathbf{X}(t, \lambda) \mapsto \mathbf{X}_c(t, \lambda) = \mathbf{X}(t-c, \lambda)$. To ensure invariance, we average in time to obtain

$$\mathbf{M}\boldsymbol{x}(t, \lambda) = \mathbf{X}(\cdot, \lambda) * \boldsymbol{\phi}_T(t) = |\boldsymbol{x} * \boldsymbol{\psi}_\lambda| * \boldsymbol{\phi}_T(t), \tag{7}$$

where $\boldsymbol{\phi}_T(t) = T^{-1}\boldsymbol{\phi}(T^{-1}t)$ for some lowpass filter $\boldsymbol{\phi}(t)$ of duration 1, so $\boldsymbol{\phi}_T(t)$ has duration T. This is the mel-spectrogram $\mathbf{M}\boldsymbol{x}(t, \lambda)$ of $\boldsymbol{x}(t)$. For $|c| \ll T$, it satisfies $\mathbf{M}\boldsymbol{x}_c(t, \lambda) \approx \mathbf{M}\boldsymbol{x}(t, \lambda)$, so it is locally invariant to time-shifts. The underlying wavelet structure of the mel-spectrogram also ensures stability to time-warping deformations [12].

The mel-spectrogram was originally introduced for speech classification [1] and was motivated by psychoacoustic studies. It has since found widespread use in various audio classification tasks [21], [22], [23]. Traditionally, the mel-spectrogram is computed through frequency averaging of the windowed Fourier transform amplitude, also known as the spectrogram. However, it has recently been shown that they may be approximated by the time-averaged scalogram coefficients (7) [12], [24], [25]. We shall therefore use this wavelet-based variant of the mel-spectrogram in the following.

We now define the discrete mel-spectrogram using the discrete wavelet transform. The resulting convolutional network is shown in Figure 4. Instead of just convolving by $\boldsymbol{g}_k[n]$, this network also applies a modulus and subsamples by 2. The whole operation is denoted by a boxed $|\boldsymbol{g}_k|$. The result is then passed through a sequence of lowpass filters $\boldsymbol{h}^{(\mathrm{t})}[n]$ alternated with subsampling operators, approximating the convolution by $\boldsymbol{\phi}_T(t)$. The output is $JQ + 1$ signals of form $|\boldsymbol{x} * \boldsymbol{g}_{j,k}| * \boldsymbol{h}^{(\mathrm{t})}_{J-j}[2^J n]$, where j is the depth at which the modulus was applied. If the filters are chosen as in Section II-A, this approximates $\mathbf{M}x(t, \lambda)$ for a bandpass $\boldsymbol{x}(t)$.

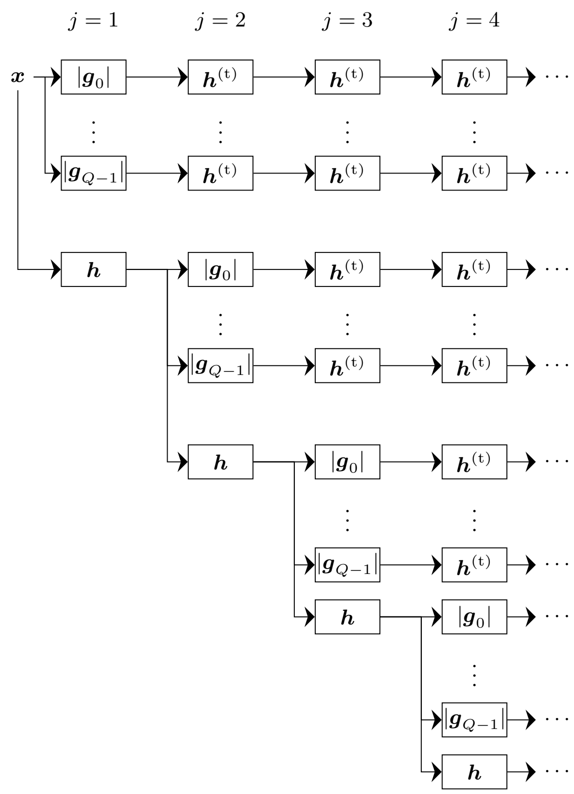


Fig. 4: Mel-spectrogram implemented as a convolutional network. Each $|\boldsymbol{g}_k|$ block convolves by a band-pass filter $\boldsymbol{g}_k[n]$, computes the modulus, and subsamples by 2. Blocks containing $\boldsymbol{h}$ or $\boldsymbol{h}^{(\mathrm{t})}$ convolve by a low-pass filter and subsample by 2.

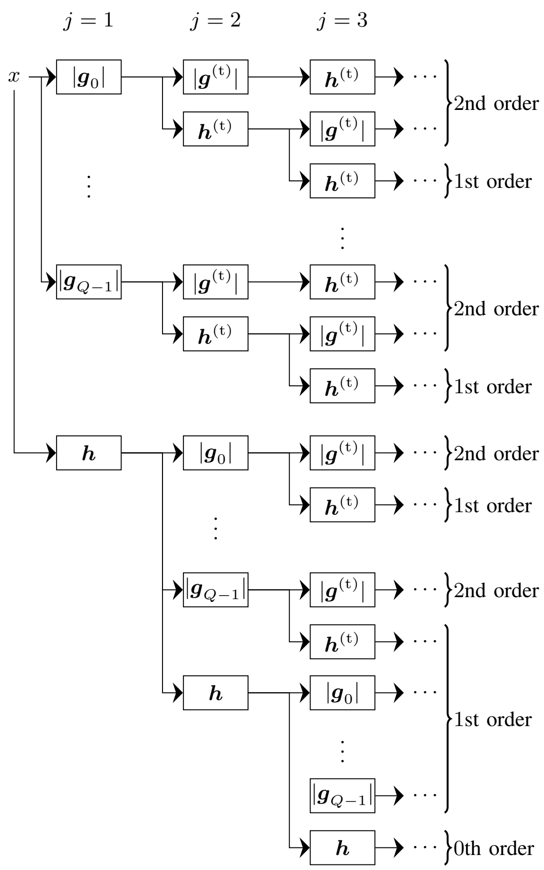


Fig. 5: A time scattering network. Each block with $|\boldsymbol{g}_k|$ or $|\boldsymbol{g}^{(\mathrm{t})}|$ outputs the modulus of the input convolved with a band-pass filter, subsampled by 2. Blocks with $\boldsymbol{h}$ and $\boldsymbol{h}^{(\mathrm{t})}$ convolves the input with a low-pass filter and subsample by 2.

For real $\boldsymbol{g}_k[n]$, we may replace the modulus with a rectified linear unit. Indeed, averaging a rectified bandpass signal approximates its Hilbert envelope, so the result is similar [26].

C. Time Scattering

The mel-spectrogram discards a large amount of potentially useful information when averaging $\mathbf{X}(t, \lambda)$ along t in (7), removing any high-frequency structure. The time scattering transform extends the mel-spectrogram and partially recovers this lost structure while maintaining invariance and stability [11], [12]. This is achieved by convolving the scalogram with a second set of wavelets, taking the modulus, and averaging.

The first-order time scattering coefficients coincide with the mel-spectrogram $\mathbf{M}\boldsymbol{x}(t, \lambda)$ and are given by

$$\mathbf{S}_1 \boldsymbol{x}(t, \lambda) = \mathbf{X}(\cdot, \lambda) * \boldsymbol{\phi}_T(t) \ .$$

The lost high frequencies of $\mathbf{X}(t, \lambda)$ are recovered by convolving with a new set of wavelets, defined from a Morlet mother wavelet $\boldsymbol{\psi}^{(\mathrm{t})}(t)$ by $\boldsymbol{\psi}_\mu^{(\mathrm{t})}(t) = 2^\mu \boldsymbol{\psi}^{(\mathrm{t})}(2^\mu t)$ for $\mu \in \mathbb{R}$. Each $\boldsymbol{\psi}_\mu^{(\mathrm{t})}(t)$ has a center frequency of approximately 2^μ, so we refer to μ as their log-frequency. Unlike their first-order counterparts $\boldsymbol{\psi}_\lambda(t)$, the second-order wavelets $\boldsymbol{\psi}_\mu^{(\mathrm{t})}(t)$ have $Q = 1$. As a result, they are better adapted to structures in $\mathbf{X}(t, \lambda)$, which are less oscillatory and more localized in time compared to those in $\boldsymbol{x}(t)$.

Convolving $\mathbf{X}(t, \lambda)$ with these wavelets along t, we obtain $\mathbf{X}(\cdot, \lambda) * \boldsymbol{\psi}_\mu^{(\mathrm{t})}(t)$. To ensure local invariance to translation, we take another modulus and average using $\boldsymbol{\phi}_T(t)$, which yields

$$\mathbf{S}_2 \boldsymbol{x}(t, \lambda, \mu) = \big|\, |\boldsymbol{x} * \boldsymbol{\psi}_\lambda| * \boldsymbol{\psi}_\mu^{(\mathrm{t})} \big| * \boldsymbol{\phi}_T(t) \ . \qquad (8)$$

These are the second-order time scattering coefficients. They describe the variability of $\mathbf{X}(t, \lambda)$ along t at frequency 2^μ, where λ is the first-order, or acoustic, log-frequency, while μ is the second-order, or modulation, log-frequency. As before, we limit ourselves to scales shorter than T by enforcing $2^{-\mu} \leq T$.

Concatenating all first- and second-order scattering coefficients $\mathbf{S}_1 \boldsymbol{x}$ and $\mathbf{S}_2 \boldsymbol{x}$ of $\boldsymbol{x}(t)$ yields the time scattering transform $\mathbf{S}\boldsymbol{x}$. Higher-order scattering coefficients may be defined [11], but these are of negligible energy [27] and do not greatly affect classification results [12]. The scattering transform satisfies the same invariance and stability properties of the mel-spectrogram described previously [11], [12]. It is more discriminative than the mel-spectrogram, however, since it captures amplitude modulations in $\mathbf{X}(t, \lambda)$ along t. As a result, the time scattering transform enjoys better performance for classification of audio [12], biomedical [13], and other types of time series [14], [15].

Other approaches capture temporal structure in the scalogram using Fourier transforms [2], [3] or second-order moments [4], [5], [28]. However, these lack the time-warping stability or noise robustness of the scattering transform [12], [11].

Extending the mel-spectrogram convolutional network of Figure 4, we define the network of a discrete time scattering transform. The result is shown in Figure 5. To implement the second-order wavelets $\boldsymbol{\psi}_{\mu}^{(\mathrm{t})}(t)$, we use the network of Figure 3, but with a single bandpass filter $\boldsymbol{g}^{(\mathrm{t})}[n]$ and a lowpass filter $\boldsymbol{h}^{(\mathrm{t})}[n]$. These are constructed to approximate convolutions with $\boldsymbol{\psi}_{\mu}^{(\mathrm{t})}(t)$ for $\mu = -j \leq -1$ as described in Section II-A.

As before, $\boldsymbol{x}[n]$ is first decomposed in the $|\boldsymbol{g}_k|$ boxes by convolution with $\boldsymbol{g}_0[n], \ldots, \boldsymbol{g}_{Q-1}[n]$ followed by modulus and subsampling by 2. However, instead of averaging their outputs, they are further convolved with $\boldsymbol{g}^{(\mathrm{t})}[n]$ followed by modulus and subsampling, denoted by $|\boldsymbol{g}^{(\mathrm{t})}|$ boxes. These coefficients are then averaged using lowpass filters $\boldsymbol{h}^{(\mathrm{t})}[n]$ which alternate with subsampling operators. This yields the second-order scattering coefficients of $\boldsymbol{x}[n]$ for the highest octave in λ and the highest octave in μ. We obtain lower octaves in μ by applying a sequence of convolutions with $\boldsymbol{h}^{(\mathrm{t})}[n]$ alternated with subsampling operators before convolving with $\boldsymbol{g}^{(\mathrm{t})}[n]$. Similarly, lower octaves in λ are obtained by applying a sequence of convolutions by $\boldsymbol{h}[n]$ and subsampling operators before the decomposition by $\boldsymbol{g}_0[n], \ldots, \boldsymbol{g}_{Q-1}[n]$. The outputs of this convolutional network approximate the continuous time scattering transform $\mathbf{S}\boldsymbol{x}$ of $\boldsymbol{x}(t)$.

III. Joint Representations in Time and Frequency

While successfully describing temporal modulation, the time scattering transform fails to capture more sophisticated time-frequency structure, as shown in Section III-A. It fails because it decomposes the scalogram as a set of one-dimensional time series. Section III-B introduces the joint time-frequency scattering transform, which instead decomposes the scalogram in both time and log-frequency. Its convolutional network representation introduces connections between nodes in each layer, increasing its discriminability. This property is demonstrated in Section III-C, where we show how the proposed transform accurately captures frequency-modulated excitations.

A. Loss of Time-Frequency Structure

The time scattering convolutional network in Figure 5 has a tree structure, that is, each node only has one parent. In contrast, a general convolutional network sums contributions from multiple nodes in a layer to produce a node in the next layer. Due to this tree structure, the time scattering transform is not sensitive to certain time-frequency deformations.

To see this, we suppose that $\boldsymbol{x}(t)$ is transformed into $\tilde{\boldsymbol{x}}(t)$ whose scalogram $\tilde{\mathbf{X}}(t, \lambda)$ is an approximate translation of $\mathbf{X}(t, \lambda)$ by $\boldsymbol{\tau}(\lambda)$ in each frequency band. In other words, $\tilde{\mathbf{X}}(t, \lambda) \approx \mathbf{X}(t - \boldsymbol{\tau}(\lambda), \lambda)$. Such transformations are illustrated in Figure 6 for a speech signal and a Dirac delta function. This time-frequency warping misaligns the speech harmonics and transforms the delta function into a chirp. Although $\boldsymbol{x}(t)$ differs

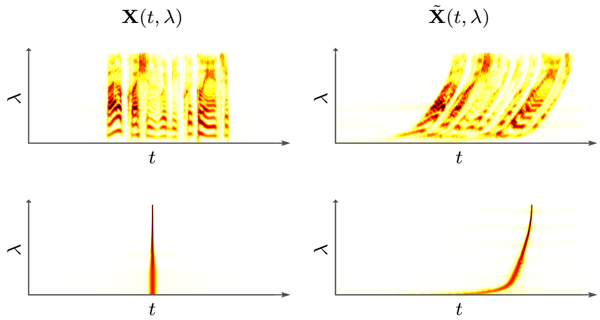


Fig. 6: Effect of frequency-dependent time-shifts $\boldsymbol{\tau}(\lambda)$ on scalograms of a speech recording (top) and a Dirac delta function (bottom). The two columns correspond to the original signal $\boldsymbol{x}(t)$ and the transformed signal $\tilde{\boldsymbol{x}}(t)$, respectively.

markedly from $\tilde{\boldsymbol{x}}(t)$, this is not detected by time scattering if $|\boldsymbol{\tau}(\lambda)| \ll T$. Indeed, the effect of the frequency-varying time shift disappears when averaging by $\boldsymbol{\phi}_T(t)$. Computing the scattering transforms $\mathbf{S}\boldsymbol{x}$ and $\mathbf{S}\tilde{\boldsymbol{x}}$ for T equal to the signal length yields relative differences $\|\mathbf{S}\tilde{\boldsymbol{x}} - \mathbf{S}\boldsymbol{x}\| / \|\mathbf{S}\boldsymbol{x}\|$ of 0.07 and 0.09 for the speech signal and the delta function, respectively.

Detection of time-frequency warping requires measurement of scalogram variability across frequency. In particular, the second-order wavelet convolution (8) in time must be replaced by a convolution in time and log-frequency.

B. Joint Time-Frequency Scattering

Existing methods for capturing a signal's time-frequency geometry are not always suitable for classification. For example, McDermott & Simoncelli [28] compute higher-order moments of the scalogram across frequencies. Through synthesis experiments, this representation is shown to provide a good model for audio textures. However, higher-order moments are not robust to noise, reducing the descriptor's usefulness for classification.

An alternative approach, motivated by neurophysiological studies in the audio cortex of ferrets, is the cortical transform of Shamma et al. [16]. It decomposes the scalogram in both time and log-frequency using two-dimensional Gabor wavelets. The cortical transform and related representations have brought significant improvements over mel-spectrograms in tasks from speech classification [29], [30] to timbre analysis [17], [31]. Unfortunately, the lack of time-shift invariance and time-warping stability limits the performance of this approach.

In the following, we adapt the cortical transform within the scattering framework, allowing us to address its invariance and stability. This also lets us analyze its discriminative power.

We first decompose the scalogram $\mathbf{X}(t, \lambda)$ using a two-dimensional wavelet transform. As before, we use Morlet wavelets. Two-dimensional Morlet wavelets are also used in the two-dimensional scattering transform, which enjoys significant success in natural image classification [32], [33]. For these images, however, wavelets are obtained by rotating and uniformly scaling a mother wavelet, which is not appropriate for the scalogram. Indeed, rotation does not preserve the

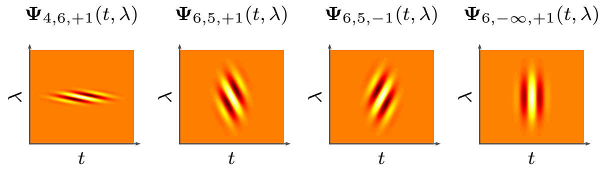


Fig. 7: Real parts of four time-frequency wavelets $\Psi_{\mu,\ell,s}(t,\lambda)$.

relationship between time and frequency–a rotated scalogram is generally not the scalogram of some other signal.

We instead define our wavelets separably, with independent scaling along time and log-frequency. The time-frequency mother wavelet $\Psi(t,\lambda) = \psi^{(t)}(t)\,\psi^{(f)}(\lambda)$ is the product of two one-dimensional functions in time and log-frequency. Both the time $\psi^{(t)}(t)$ and the frequency $\psi^{(f)}(\lambda)$ wavelets are Morlet wavelets with $Q = 1$. Dilating by $2^{-\mu}$ along t, dilating by $2^{-\ell}$ along λ, and reflecting according to s yields the wavelet

$$\Psi_{\mu,\ell,s}(t,\lambda) = 2^{\mu+\ell}\,\psi^{(t)}(2^{\mu}t)\,\psi^{(f)}(s2^{\ell}\lambda)\ , \tag{9}$$

where the spin $s = \pm 1$ specifies the oscillation direction (up or down). The frequency of the wavelet along t is 2^{μ}, so μ is the log-frequency of $\Psi_{\mu,\ell,s}(t,\lambda)$. Its frequency along λ is 2^{ℓ}, so we refer to it as a "quefrency." Consequently, ℓ is the "log-quefrency" of $\Psi_{\mu,\ell,s}(t,\lambda)$.

As before, μ satisfies $2^{-\mu} \leq T$. Along λ, we fix some maximum log-frequency scale F, measured in octaves, and let $2^{-\ell} \leq F$. At this maximum scale, we include a lowpass filter to capture average structure along λ. Specifically, we set

$$\Psi_{\mu,-\infty,+1}(t,\lambda) = 2^{\mu}\psi^{(t)}(2^{\mu}t)\,\phi_F(\lambda)\ . \tag{10}$$

Note that these are only defined for $s = +1$. Figure 7 shows a few sample two-dimensional wavelets $\Psi_{\mu,\ell,s}(t,\lambda)$.

The two-dimensional wavelet transform of the scalogram $\mathbf{X}(t,\lambda)$ computes convolutions $\mathbf{X} * \Psi_{\mu,\ell,s}(t,\lambda)$. It captures the joint variability of $\mathbf{X}(t,\lambda)$ at log-frequency μ and log-quefrency ℓ with spin s. To ensure time-shift invariance and time-warping stability, we take the complex modulus and average, obtaining the second-order joint time-frequency scattering coefficients

$$\mathbf{S}_2\boldsymbol{x}(t,\lambda,\mu,\ell,s) = |\mathbf{X} * \Psi_{\mu,\ell,s}(\cdot,\lambda)| * \phi_T(t). \tag{11}$$

These coefficients describe the time-frequency geometry of $\boldsymbol{x}(t)$ at time t and log-frequency λ. They have the same invariance and stability properties as second-order time scattering coefficients, but with increased discriminability.

Concatenating the first-order time scattering coefficients $\mathbf{S}_1\boldsymbol{x}$ and the second-order time-frequency scattering coefficients $\mathbf{S}_2\boldsymbol{x}$ yields the complete joint time-frequency scattering transform $\mathbf{S}\boldsymbol{x}$ of $\boldsymbol{x}(t)$. As for time scattering, we may define higher-order coefficients, but these are often of limited use for classification. For each t, there are $O(Q\log_2 T)$ first-order coefficients and $O(Q(\log_2 T)^2\log_2 F)$ second-order coefficients.

We now define a convolutional network to provide a discrete implementation of the joint scattering transform. In the time scattering network (see Figure 5), we approximate the convolution of $\mathbf{X}(t,\lambda)$ with $\psi_{\mu}^{(t)}(t)$ along t by cascading discrete filters

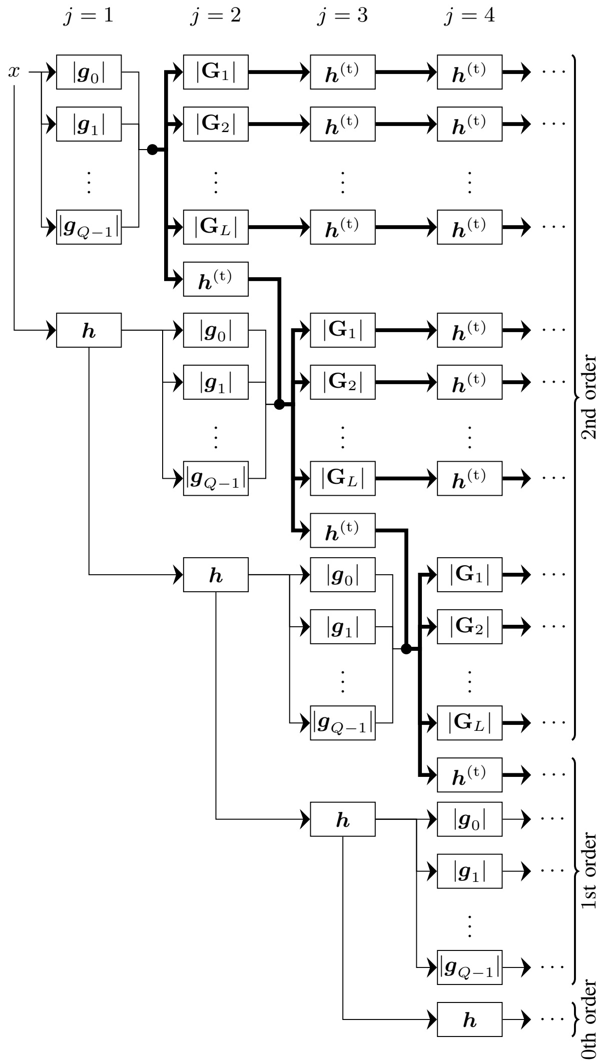


Fig. 8: A joint time-frequency scattering network. Each $|\boldsymbol{g}_k|$ block convolves a one-dimensional signal with the band-pass filter $\boldsymbol{g}_k$ and outputs its modulus. The outputs are aggregated into two-dimensional arrays shown by thick lines. A $|\mathbf{G}_\ell|$ block convolves a two-dimensional array with the band-pass filter $\mathbf{G}_\ell$ and outputs its modulus. The $\boldsymbol{h}$ and $\boldsymbol{h}^{(t)}$ blocks convolve only in time. All blocks subsample their output in time by 2.

$\boldsymbol{h}^{(t)}[n]$ and $\boldsymbol{g}^{(t)}[n]$, alternated with subsampling operators. The joint transform network incorporates additional filters along the discrete log-frequency $m = Q\lambda = -jQ - k \in \mathbb{Z}$, where, as before, j and k are the octave and subband indices of λ.

In a given layer, the modulus bandpass outputs of the previous layer are arranged along time n and log-frequency m into a two-dimensional array. This array is then filtered along m by different filters $2^{\ell}\psi^{(f)}(s2^{\ell}m/Q)$. It is also filtered by $\phi_F(m/Q)$ to account for $\ell = -\infty$. The sampling interval of the filters is $1/Q$, since this is the spacing of the discretized log-frequencies $\lambda = m/Q$. Each frequency-filtered array is then filtered by $\boldsymbol{g}^{(\mathrm{t})}[n]$ along n.

Combining these into two-dimensional filters, we get

$$\mathbf{G}_{\ell,s}[n,m] = \boldsymbol{g}^{(\mathrm{t})}[n]\, 2^{\ell} \boldsymbol{\psi}^{(\mathrm{f})}(s 2^{\ell} m/Q)$$
$$\mathbf{G}_{-\infty,+1}[n,m] = \boldsymbol{g}^{(\mathrm{t})}[n]\, \boldsymbol{\phi}_F(m/Q) \ ,$$

where $\ell \in \mathbb{Z}$ such that $-\log_2 F \le \ell \le \log_2 Q$ (to ensure that $1/Q \le 2^{-\ell} \le F$) and $s = \pm 1$. Abusing notation slightly, we renumber this set of discrete filters as $\mathbf{G}_1[n,m], \ldots, \mathbf{G}_L[n,m]$. These filters capture all log-quefrencies along λ, but only high frequencies along n. The missing low frequencies are absorbed by $\boldsymbol{h}^{(\mathrm{t})}[n]$, which averages along n, leaving m intact.

Using these filters, we construct the convolutional network shown in Figure 8, extending the time scattering network of Figure 5. Small circles denote aggregation of multiple time series into a two-dimensional array, while the arrays themselves are thick lines. We denote by a boxed $|\mathbf{G}_\ell|$ convolution with $\mathbf{G}_\ell[n,m]$ for $\ell = 1, \ldots, L$, followed by a complex modulus and subsampling by 2 along n. Similarly, a boxed $\boldsymbol{h}^{(\mathrm{t})}$ denotes lowpass filtering along n by $\boldsymbol{h}^{(\mathrm{t})}[n]$ followed by subsampling.

Starting with a signal $\boldsymbol{x}[n]$, we first compute its decomposition using the first-order blocks $|\boldsymbol{g}_0|, \ldots, |\boldsymbol{g}_{Q-1}|$, extracting the highest octave of the signal. We then combine these into a two-dimensional array which is decomposed by $|\mathbf{G}_1|, \ldots, |\mathbf{G}_L|$. The outputs of $|\mathbf{G}_1|, \ldots, |\mathbf{G}_L|$ are then forwarded to a succession of $\boldsymbol{h}^{(\mathrm{t})}$ blocks which implement the averaging by $\boldsymbol{\phi}_T[n]$. The original array is also decomposed by $\boldsymbol{h}^{(\mathrm{t})}$, and the result is concatenated to the first-order outputs of the second layer (that is, the second octave of the original signal). We then repeat the process on this array. As before, an appropriate choice of $\boldsymbol{g}^{(\mathrm{t})}[n]$ and $\boldsymbol{h}^{(\mathrm{t})}[n]$ ensures that the network accurately approximates the continuous joint scattering transform.

The important difference between this network and the time scattering network is the presence of within-layer connections. These break the tree structure, increasing discriminability through better characterization of time-frequency geometry. Returning to the frequency-warped signals of Figure 6, the joint network separates the original and transformed signals, with $\|\mathbf{S}\tilde{\boldsymbol{x}} - \mathbf{S}\boldsymbol{x}\| / \|\mathbf{S}\boldsymbol{x}\|$ of 0.41 and 0.90, compared to 0.07 and 0.09 for time scattering. This network therefore has same time-shift invariance as time scattering, but with better discriminability.

The output of a scattering network may be used input to another convolutional network whose filters are subsequently optimized for some classification task. This yields a large convolutional network taking raw waveforms as input and whose first few layers are fixed. By fixing certain layers, the network has fewer parameters to optimize and could then be trained using less data. Previous work training convolutional networks on raw waveforms have yielded mel-like filters in the first few layers [34], providing some support for this idea. In addition, the success of transfer learning [35], [36], [37] suggests that there exists certain universal representations which perform well for a wide range of tasks. The joint scattering network provides a way to construct such a representation satisfying certain invariance and discriminability conditions.

C. Frequency Modulation

The above construction is similar to that of traditional convolutional networks except that filters are not learned from data. These filters provide the time-shift invariance and time-warping stability of the time scattering transform, but the joint transform is also more discriminative. To illustrate this, we show how the joint time-frequency scattering transform is captures frequency modulation structure ignored by time scattering.

Let $\boldsymbol{x}(t) = \exp(2\pi \mathrm{i}\, \boldsymbol{\xi}(t))$ be a frequency-modulated excitation with instantaneous phase $\boldsymbol{\xi}(t)$. At time t, its instantaneous frequency is given by $\boldsymbol{\xi}'(t)$, while the relative change in this frequency, the (relative) chirp rate, is $\boldsymbol{\xi}''(t)/\boldsymbol{\xi}'(t)$. Frequency modulation occurs in a variety signals, such as speech, animal calls, music and radar signals [38].

We now consider a particular case of frequency modulation: the exponential chirp. Here $\boldsymbol{\xi}(t) = 2^{\alpha t}$, so it has instantaneous frequency $\boldsymbol{\xi}'(t) = \alpha \log(2) 2^{\alpha t}$ and constant chirp rate $\boldsymbol{\xi}''(t)/\boldsymbol{\xi}'(t) = \alpha \log(2)$. We note that an arbitrary frequency-modulated excitation may be locally approximated by an exponential chirp by setting $\alpha = (\log 2)^{-1}\, \boldsymbol{\xi}''(t)/\boldsymbol{\xi}'(t)$.

For exponential chirps, we have the following result.

Theorem 1. *Let $\boldsymbol{\psi}_\lambda(t)$ and $\boldsymbol{\Psi}_{\mu,\ell,s}(t,\lambda)$ be defined as in* (1) *and* (9). *We require that $\boldsymbol{\psi}(t)$ have compact support, that $\|\boldsymbol{\psi}\|_\infty$, $\|\boldsymbol{\psi}'\|_1$, $\int_{\mathbb{R}} |u| |\widehat{\boldsymbol{\psi}}(2^u)|\, \mathrm{d}u$, and $\|{\boldsymbol{\psi}^{(\mathrm{t})}}'\|_\infty$ are bounded, and that* $\operatorname{supp}\, \boldsymbol{\psi}^{(\mathrm{f})}(\lambda) \subset [-A, A]$ *for some* $A > 0$. *Further, we assume that $\boldsymbol{\psi}^{(\mathrm{t})}(t)$ is the product of a positive envelope $|\boldsymbol{\psi}^{(\mathrm{t})}|(t)$ and* $\exp(2\pi \mathrm{i}\, t)$. *Let $\boldsymbol{x}(t) = \exp(2\pi \mathrm{i}\, 2^{\alpha t})$ for some $\alpha \in \mathbb{R}$. The joint scattering transform* (11) *then satisfies*

$$\begin{aligned}&\mathbf{S}_2 \boldsymbol{x}(t, \lambda, \mu, \ell, s)\\ &\quad = \frac{c_0 E(t,\lambda)}{\alpha} \left| \widehat{\boldsymbol{\psi}}^{(\mathrm{f})} \left(-\frac{s 2^{\mu-\ell}}{\alpha} \right) \right| + \boldsymbol{\varepsilon}(t, \lambda, \mu, \ell, s) \ ,\end{aligned}$$

where

$$E(t,\lambda) = |\boldsymbol{\psi}_\mu^{(\mathrm{t})}| * \phi_T \left(t - \frac{\lambda}{\alpha} + \frac{\log\log 2^\alpha}{\log 2^\alpha} \right) \ ,$$

$$\|\boldsymbol{\varepsilon}\|_\infty < C \left(|\alpha| 2^{-\lambda + 2^{-\ell} A} + 2^{2\mu} |\alpha|^{-2} + 2^{2\mu-\ell} |\alpha|^{-2} \right) \ ,$$

for $C > 0$ depending only on $\boldsymbol{\psi}(t)$, $\boldsymbol{\psi}^{(\mathrm{t})}(t)$, and $\boldsymbol{\psi}^{(\mathrm{f})}(\lambda)$, and $c_0 = \int_{\mathbb{R}} |\widehat{\boldsymbol{\psi}}(2^u)|\, \mathrm{d}u$,

The proof is given in Appendix A. The result relies on approximating $\mathbf{X}(t,\lambda)$ by $|\widehat{\boldsymbol{\psi}}(\log(2^\alpha) 2^{-\lambda+\alpha t})|$. Since $|\widehat{\boldsymbol{\psi}}(\omega)|$ is maximized at $\omega = 1$, this forms a ridge $\lambda = \alpha t$ with slope α, as illustrated in Figure 9(a,b). In the joint transform, this ridge only activates certain second-order wavelets $\boldsymbol{\Psi}_{\mu,\ell,s}(t,\lambda)$. Indeed, only wavelets whose slope $-s 2^{\mu-\ell}$ aligns with $\lambda = \alpha t$ yield large coefficients. Taking the complex modulus and averaging in time preserves this slope information.

Let us consider another chirp $\tilde{\boldsymbol{x}}(t) = \exp(2\pi \mathrm{i}\, 2^{\tilde{\alpha} t})$. We may obtain $\tilde{\boldsymbol{x}}(t)$ from $\boldsymbol{x}(t)$ using a frequency-dependent time-shift of its scalogram $\mathbf{X}(t,\lambda)$ as in Section III-A. Here, we take

$$\boldsymbol{\tau}(\lambda) = \lambda \left(\frac{1}{\alpha} - \frac{1}{\tilde{\alpha}} \right) - \frac{\log\log 2^\alpha}{\log 2^\alpha} + \frac{\log\log 2^{\tilde{\alpha}}}{\log 2^{\tilde{\alpha}}} \ .$$

As we saw in Section III-A, the time scattering transform is not sensitive to such changes. In other words, the scattering transform discards information on slope, rendering it unsuitable for describing frequency modulation. The same applies to related representations which also decompose each subband of $\mathbf{X}(t,\lambda)$

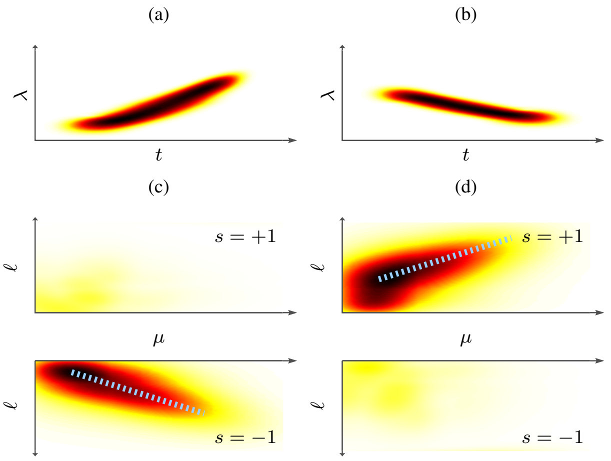


Fig. 9: Scalograms of two exponential chirps with chirp rates (a) $\alpha = 4$ and (b) $\alpha = -2$. (c, d) Corresponding second-order joint time-frequency scattering coefficients $\mathbf{S}_2\boldsymbol{x}(t, \lambda, \mu, \ell, s)$ for fixed t and λ. The dotted lines satisfy $s2^{\mu-\ell} = -\alpha$.

separately, such as mel-spectrograms, MFCCs, and modulation spectrograms. This information loss is fundamentally due to the tree structure of their convolutional networks.

Theorem 1 states that, for fixed t and λ, $\mathbf{S}_2\boldsymbol{x}(t, \lambda, \mu, \ell, s)$ is approximately proportional to $|\widehat{\boldsymbol{\psi}}^{(\mathrm{f})}(-s2^{\mu-\ell}\alpha^{-1})|$. Since $\widehat{\boldsymbol{\psi}}^{(\mathrm{f})}$ is concentrated around frequency 1, this is maximized for $-s2^{\mu-\ell}\alpha^{-1} = 1$. In other words, a ridge is present along $s2^{\mu-\ell} = -\alpha$. Frequency modulation structure in the form of the chirp rate α, is thus encoded in the second-order joint time-frequency scattering coefficients. Consequently, they are sensitive frequency-dependent time-shifts $\mathbf{X}(t, \lambda) \mapsto \mathbf{X}(t - \boldsymbol{\tau}(\lambda), \lambda)$ even when $|\boldsymbol{\tau}(\lambda)| \ll T$, since these change α.

Figure 9(c,d) displays a subset of the second-order joint scattering coefficients for the chirps whose scalograms are shown in Figure 9(a,b). These coefficients do indeed show a maximum along the predicted ridge. At low ℓ and high μ, the approximation does not hold, but for most of the frequency range, it is accurate. We thus see how the chirp rate α is captured by the joint scattering coefficients in a natural way.

IV. Audio Texture Synthesis

Section III-A showed how mel-spectrograms and time scattering transforms do not adequately capture time-frequency structure. As T increases, this problem becomes more serious, necessitating the introduction of the joint time-frequency scattering transform. In this section, we illustrate the representational power of this transform using texture synthesis experiments.

With the aim of generating realistic soundtracks of arbitrary duration, audio texture synthesis has many applications in virtual reality and multimedia design [39]. In computational neuroscience, it also offers a testbed for the comparative evaluation of biologically plausible models for auditory perception [28]. Given a signal $\boldsymbol{x}(t)$ and a time-shift invariant representation $\Phi\boldsymbol{x}$ of $\boldsymbol{x}$, the texture synthesis problem may be formulated as the minimization of the error $E(\boldsymbol{y}) = \|\Phi\boldsymbol{y} - \Phi\boldsymbol{x}\|^2$ between $\Phi\boldsymbol{x}$ and the representation $\Phi\boldsymbol{y}$ of the synthesized signal $\boldsymbol{y}(t)$. Here, Φ can be a scattering transform $\mathbf{S}$, a mel-spectrogram $\mathbf{M}$, or some other representation.

The state of the art in the domain is held by McDermott & Simoncelli [28], who define Φ using a set of summary statistics. These statistics are similar to time scattering transform as they are calculated using cascades of constant-Q filterbanks and pointwise nonlinearities. However, unlike the scattering transform, which simply averages in time, McDermott & Simoncelli also compute higher-order statistical moments: variance, skewness, kurtosis, and correlation coefficients across frequency bands. These coefficients are very sensitive to outliers in the data, which reduces their applicability to classification.

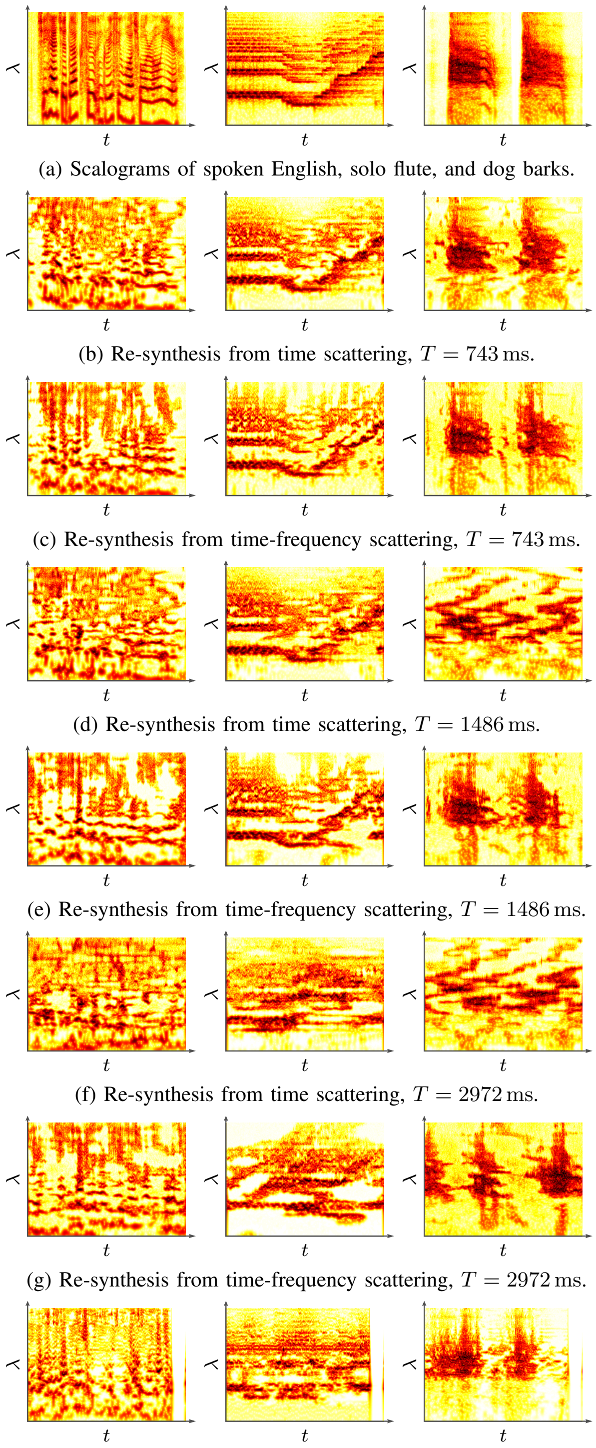
To synthesize $\boldsymbol{y}(t)$, we first initialize using random Gaussian noise with power spectral density matching the first-order scattering coefficients $\mathbf{S}_1\boldsymbol{x}(t, \lambda)$ of the target waveform $\boldsymbol{x}(t)$, since these coefficients are present in all the considered representations. We then iteratively refine the signal by gradient descent [40]. Because the modulus nonlinearity is not convex, the error $E(\boldsymbol{y})$ is not convex; consequently, gradient descent only converges towards a local minimum of $E(\boldsymbol{y})$. However, the local minimum is typically of low error, with $E(\boldsymbol{y})$ equal to around $0.02 \times \|\Phi x\|^2$ for typical audio recordings $\boldsymbol{x}(t)$. We found empirically that the convergence rate is increased using a fixed momentum term and a "bold driver" learning rate policy [41].

Gradient descent in a scattering network can be implemented by backpropagation from deeper to shallower layers. Like in a deep convolutional network, the gradient backpropagation of the convolution with each wavelet $\boldsymbol{g}_k(t)$ corresponds to a convolution with the adjoint filter $\boldsymbol{g}_k^{\dagger}(t) = \bar{\boldsymbol{g}}_k(-t)$, obtained by time reversal and complex conjugation of $\boldsymbol{g}_k(t)$.

Figure 10 shows the synthesized scalograms of three sounds for various values of T. Here, time-frequency scattering outperforms time scattering for T greater than 1 s. In particular, speech is more intelligible due to better reconstruction of articulations, individual notes in a musical scale are more salient, and broadband impulses such as dog barks keep their typical amplitude envelopes and inter-onset intervals. Compared to the representation of McDermott & Simoncelli [28], time-frequency scattering achieves similar quality, but does not have the same sensitivity to outliers. Indeed, the contractivity of the wavelet transform and the modulus ensures the scattering transform's robustness to additive noise [11], [12].

V. Supervised Classification

We evaluate the performance of the joint time-frequency scattering transform on various classification tasks. It is shown to enjoy significantly greater accuracy compared to baseline MFCC and time scattering approaches. In fact, the proposed transform performs comparably to state-of-the-art learned convolutional networks whose training requires significant computational resources and large training sets. As a result, the joint scattering transform provides a good alternative when such an expensive training step is infeasible or undesirable.



(a) Scalograms of spoken English, solo flute, and dog barks.

(b) Re-synthesis from time scattering, $T = 743\,\text{ms}$.

(c) Re-synthesis from time-frequency scattering, $T = 743\,\text{ms}$.

(d) Re-synthesis from time scattering, $T = 1486\,\text{ms}$.

(e) Re-synthesis from time-frequency scattering, $T = 1486\,\text{ms}$.

(f) Re-synthesis from time scattering, $T = 2972\,\text{ms}$.

(g) Re-synthesis from time-frequency scattering, $T = 2972\,\text{ms}$.

(h) Re-synthesis using McDermott & Simoncelli [28].

Fig. 10: Scalograms of audio re-synthesis using time scattering, time-frequency scattering, and McDermott & Simoncelli [28]. Synthesis is performed at various time scales T and inputs: spoken English (left), solo flute (middle), and dog barks (right).

A. Frequency Transposition Invariance

In addition to time-shifting and time-warping, signals are also transformed by frequency-shifting and frequency-warping. Frequency-shifting, also known as frequency transposition, changes the pitch but leaves subband envelopes intact. This shifts the scalogram $\mathbf{X}(t, \lambda)$ by a fixed amount η in log-frequency, giving $\mathbf{X}(t, \lambda - \eta)$. While certain tasks are sensitive to pitch, like speaker identification, others, like speech recognition in non-tonal languages, require invariance to transposition.

The time scattering transform is rendered transposition invariant and stable to frequency-warping by applying a second scattering transform along log-frequency λ. The result is the separable time and frequency scattering transform [12]. Note that we may skip the averaging step of this second scattering transform. Indeed, the averaging step is a linear map that can be learned by the classifier given enough training data [12].

To render the joint time-frequency scattering transform transposition invariant, we similarly apply a second scattering transform along λ for the first-order coefficients $\mathbf{S}_1\boldsymbol{x}$. For the second order $\mathbf{S}_2\boldsymbol{x}$, however, we simply average along λ, since the two-dimensional wavelet decomposition already captures the relevant frequency structure. The resulting representation then has the necessary transposition invariance and frequency-warping stability properties. Again, if the training set is large enough, the final averaging steps can be learned by the classifier.

B. Phone Segment Classification

An individual phone in speech is short, on average 40 ms in duration. For phone identification, we therefore require the invariance scale T to be of this order. Since T is small, there is less room for the type of misalignment seen in Section III-A. We therefore expect the joint time-frequency scattering to provide only limited improvement over time scattering.

To evaluate, we use the TIMIT dataset, which contains recordings of spoken phrases, each labeled with its constituent phones and their locations [42]. Given a phone segment, we wish to classify the phone according to the standard protocol [43], [44]. This task is simpler than continuous speech recognition, but provides a good framework for evaluating representations. The training and evaluation sets consist of 3696 and 192 phrases, respectively. We use a 400–phrase validation set to optimize hyperparameters (see Andén & Mallat [12]).

Instead of the raw scattering transform, we use their logarithm, known as the log-scattering transform, as input to the classifier [12]. We compute these coefficients over 192-second intervals centered on each segment with $T = 32$ ms. All coefficients are concatenated into a single vector together with the logarithm of the segment duration. This vector is then used for classification. The same processing is also performed for separable time and frequency scattering as well as joint time-frequency scattering. We set the maximum frequency scale F to 4 octaves. As a baseline, we compute Delta-MFCCs, which supplement standard MFCCs with first and second time derivatives [45]. These are computed with the same windows and concatenation as the log-scattering coefficients.

For each representation, we train a support vector machine (SVM) [46] with a Gaussian kernel. Here and in the following, we use a modified implementation of the LIBSVM library [47].

Representation	Error (%)
Delta-MFCCs	18.3
State of the art [48]	15.0
Time scattering	17.3
Separable time and freq. scattering	16.1
Joint time-freq. scattering	15.7

TABLE I: Error rates for phone segment classification. All representations are computed with $T = 32$ ms and $Q = 8$.

Representation	Error (%)
Delta-MFCCs	39.3
Time convolutional networks	38.2
Time-frequency convolutional networks	28.3
Spiral convolutional networks [55]	26.0
Time scattering	38.0
Time-frequency scattering	22.0

TABLE II: Error rates for musical instrument classification. All representations are computed with $T = 3$ s and $Q = 12$.

Results are shown in Table I. Delta-MFCCs have an error rate of 18.3%, while the state-of-the-art representation, a convolutional neural network, achieves 15.0% [48]. The time scattering transform obtains an error rate of 17.3%, which is improved by scattering along log-frequency to give 16.1%. Finally, we obtain an error of 15.7% for the joint time-frequency scattering transform. As mentioned earlier, the amount of time-frequency structure present in an individual phone is small, but there is enough to give a small improvement to the joint transform. This is partly due to the fact that certain phones (such as plosives) are characterized by their onset, which exhibits sophisticated time-frequency structure.

For this task, the joint time-frequency scattering transform does not outperform the state-of-the-art learned convolutional network. Note, however, that the only learning involved for the scattering transform is training the SVM. The scattering network weights are fixed, providing a simpler representation with acceptable performance. Combining these two approaches–a scattering transform as input to a more adaptive deep neural network–could yield even better performance as fewer parameters need to be estimated. Indeed, replacing mel-spectrograms by scattering transforms in deep neural networks have improved performance for several tasks [49], [50], [51].

C. Musical instrument recognition

The timbre of a musical instrument is essentially determined by its shape and materials. Both remain constant during a musical performance. Therefore, musical instruments may be modeled as dynamical systems with constant parameters. The task of musical instrument recognition is to retrieve these parameters while remaining invariant to changes in pitch, intensity, and expressive technique induced by the performer.

In a musical instrument, the response of the vibrating body to an excitation is typically nonlinear. As a result, sharp onsets produce distinctive time-frequency patterns which are not adequately captured by short-term audio descriptors, such as MFCCs, which operate on scales $T \approx 20\,\mathrm{ms}$. Joint time-frequency scattering, on the other hand, captures such patterns up to the scale $T \approx 3\,\mathrm{s}$ of a short musical phrase.

To illustrate this, we apply it to automatic instrument recognition in solo phrases with a taxonomy of eight instruments. In line with the cross-collection methodology of Bogdanov et al. [52], we train and validate all models on the MedleyDB v1.1 dataset [53] and test them on the solosDb dataset [54]. This is the evaluation setting of Lostanlen & Cella [55].

Results are shown in Table II. It appears that all models which do not explicitly decompose in both time and log-frequency (Delta-MFCCs, time scattering, and a convolutional network of temporal convolutions on the scalogram) perform comparably, with errors around 38%. On the other hand, joint time-frequency scattering obtains an error of 22.0%, outperforming state-of-the-art time-frequency convolutional networks and spiral convolutional networks, which achieve error rates of 28.3% and 26.0%, respectively [55]. The improved accuracy of the joint transform is partially explained by reduced overfitting for categories with little available data.

D. Acoustic Scene Classification

Environmental sounds and acoustic scenes are characterized by larger-scale time-frequency structures. These recordings typically stretch over several seconds, each composed of shorter sound events which characterize the scene. This could be birdsong in a park, car horns in a street, or the scraping of chairs in a café. To differentiate between different sequences of such events, we must characterize longer-range structures. As discussed above, this is not possible using standard representations, such as MFCCs or time scattering, which do not adequately capture time-frequency structure.

We evaluate the joint scattering transform on three acoustic scene datasets: UrbanSound8K (US8K) [56], ESC-50 [23], and DCASE2013 [57]. US8K and DCASE2013 have 10 classes each, while ESC-50 contains 50 classes, ranging from gun shots and subway stations to crying babies and supermarkets. Both US8K and ESC-50 contain several thousand recordings of approximate duration 4 s. DCASE2013, on the other hand, contains 100 (public) training samples and 100 (private) evaluation samples, each of duration 30 s. All recordings being relatively long, they may exhibit sophisticated time-frequency structures that are discriminative for classification.

For US8K and ESC-50, we compute scattering transforms with $Q = 8$ and $T = 4$ s. To ensure transposition invariance, we explicitly average the separable and joint transforms over $F = 1$ octave (US8K) or $F = 2$ octaves (ESC-50). Since T equals the clip duration, each clip yields a single scattering vector, which is fed into the classifier. For DCASE2013, we compute scattering transforms with $Q = 4$, $T = 1.5$ s, and frequency averaging over $F = 8$ octaves where applicable. Each clip yields multiple scattering vectors which are classified separately. The overall class is then obtained by majority voting. Delta-MFCCs are computed for all datasets as a baseline. For each representation, we train a linear SVM whose hyperparameters are optimized by cross-validation on the training subset.

The error for US8K and ESC-50 is calculated through cross-validation on pre-specified folds. For these datasets, we use the data augmentation scheme of Salamon and Bello [58], but

Representation	US8K	ESC-50	DCASE2013
Delta-MFCCs [56], [23]	46.0	56.0	42
Salamon & Bello [58]	21.0	–	–
SoundNet [36]	–	25.8	12
L^3 network [37]	–	20.7	7
Time scatt.	26.9 ± 4.1	39.3 ± 2.2	12
Separable time and freq. scatt.	22.8 ± 3.0	26.0 ± 2.7	6
Joint time-freq. scatt.	19.6 ± 2.9	21.8 ± 2.0	5

TABLE III: Average and standard deviation of error rates for scene classification on US8K, ESC-50, and DCASE2013.

without pitch-shifting, since transposition invariance is already enforced. We calculate the DCASE2013 error on the evaluation subset in accordance with previous work [36], [37].

Results are shown in Table III. The Delta-MFCCs have error rates of 46.0%, 56.0%, and 42% for US8K, ESC-50, and DCASE2013, respectively. State-of-the-art convolutional networks, on the other hand, obtain 21.0%, 20.7% and 7%.

The standard time scattering transform yields accuracies of 26.8% (US8K), 39.3% (ESC-50), and 12% (DCASE2013), improving on Delta-MFCCs by better capturing the temporal structure of each subband. Adding a scattering transform along the log-frequency axis improves results to 22.8% (US8K), 26.0% (ESC-50), and 6% (DCASE2013). This improvement is expected since these sounds exhibit significant pitch variability which is not discriminative to each class.

The joint time-frequency scattering transform performs even better, giving errors of 19.6% (US8K), 21.8% (ESC-50), and 5% (DCASE2013). This is partly because environmental sounds are often characterized by dynamic filters which evolve in time, creating a spectrotemporal filter. The mechanical and biological nature of these sounds also results in frequency modulation. Both phenomena are examples of time-frequency geometry which are well-characterized by the joint scattering transform.

Again, the joint scattering transform performs comparably to learned convolutional neural networks. However, learned networks require significant computational resources to train and certain expertise in designing the network. Both SoundNet and the L^3 network are pretrained on large external datasets, requiring several days of computation on graphics processing units. In contrast, the joint scattering transform has a fixed network structure, so the only training needed is for the SVM, requiring at most a few hours. By considering the invariances of the problem (time-shifting, frequency transposition) and the structures we would like to capture (joint time-frequency geometry), we obtain good performance without costly pretraining.

VI. Conclusion

We introduced a joint time-frequency scattering transform, a time-shift invariant descriptor with state-of-the-art classification performance for a wide range of audio datasets. Important improvements are obtained for classification tasks involving large-scale signal structures. Time-frequency scattering descriptors also recover complex signals including audio textures.

A joint time-frequency scattering has a computational structure similar to deep convolutional networks [59], but is calculated with fixed wavelet filters. It thus requires less training data to obtain accurate classification results. However, when more training examples are available, learned convolutional networks provide state-of-the-art results. Indeed, these networks adapt the representation to each classification problem. Taking into account prior information on time-frequency geometry could help improve their performance.

Acknowledgments

The authors would like to thank Justin Salamon for sharing his data augmentation code and Alex Barnett for helpful discussions on oscillatory integrals.

Appendix A

Lemma 1. *Let $\boldsymbol{\psi}_\lambda(t)$ be as defined in Theorem 1. For $\boldsymbol{x}(t) = \exp(2\pi\mathrm{i}\, 2^{\alpha t})$, we then have*

$$|\boldsymbol{x} * \boldsymbol{\psi}_\lambda|(t) = |\widehat{\boldsymbol{\psi}}\left(\log(2^\alpha)\, 2^{\alpha t - \lambda}\right)| + \boldsymbol{\varepsilon}(t, \lambda) \ , \tag{12}$$

where

$$|\boldsymbol{\varepsilon}(t, \lambda)| \leq C|\alpha| 2^{-\lambda} \tag{13}$$

for some constant $C > 0$ which only depends on $\boldsymbol{\psi}(t)$.

Proof. If $\alpha = 0$, $\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t) = \widehat{\boldsymbol{\psi}}(0)\exp(2\pi\mathrm{i})$. We therefore assume that $\alpha \neq 0$. If supp $\boldsymbol{\psi} \subset [-\Delta, \Delta]$, we have

$$\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t) = \int_{|u| \leq 2^{-\lambda}\Delta} \exp(2\pi\mathrm{i}\, 2^{\alpha(t-u)}) \boldsymbol{\psi}_\lambda(u)\, \mathrm{d}u \ . \tag{14}$$

For u close to zero, the derivative of $2^{\alpha(t-u)}$ is approximately $-\log(2^\alpha) 2^{\alpha t}$. We exploit this to integrate (14) by parts. Let $\boldsymbol{g}(u) = \exp(2\pi\mathrm{i}\, 2^{\alpha t}(2^{-\alpha u} + u\log(2^\alpha)))$. We then have

$$\begin{aligned}\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t) &= \int_{|u| \leq 2^{-\lambda}\Delta} \boldsymbol{g}(u) \exp(-2\pi\mathrm{i}\, u \log(2^\alpha) 2^{\alpha t}) \boldsymbol{\psi}_\lambda(u)\, \mathrm{d}u \\ &= \boldsymbol{g}(2^{-\lambda}\Delta) \widehat{\boldsymbol{\psi}}_\lambda(\log(2^\alpha)2^{\alpha t}) - \int_{|u| \leq 2^{-\lambda}\Delta} \boldsymbol{g}'(u) \mathbf{I}(u)\, \mathrm{d}u \ , \end{aligned} \tag{15}$$

where $\mathbf{I}(u) = \int_{-2^{-\lambda}\Delta}^{u} \exp(-2\pi\mathrm{i}\, v \log(2^\alpha) 2^{\alpha t}) \boldsymbol{\psi}_\lambda(v)\, \mathrm{d}v$.

The magnitude of the second term in (15) is bounded by

$$2\pi\, 2^{\alpha t} |\log(2^\alpha)| (1 - 2^{-|\alpha| 2^{-\lambda}\Delta}) 2^{1-\lambda}\Delta \max_{|u| \leq 2^{-\lambda}\Delta} |\mathbf{I}(u)| \ . \tag{16}$$

Integrating $\mathbf{I}(u)$ by parts and taking the modulus gives

$$|\mathbf{I}(u)| \leq \frac{\|\boldsymbol{\psi}_\lambda\|_\infty + \|\boldsymbol{\psi}'_\lambda\|_1}{2\pi |\log(2^\alpha)|\, 2^{\alpha t}} = 2^\lambda \frac{\|\boldsymbol{\psi}\|_\infty + \|\boldsymbol{\psi}'\|_1}{2\pi |\log(2^\alpha)|\, 2^{\alpha t}} \ ,$$

since $\alpha \neq 0$, $\|\boldsymbol{\psi}_\lambda\|_\infty = 2^\lambda \|\boldsymbol{\psi}\|_\infty$ and $\|\boldsymbol{\psi}'_\lambda\|_1 = 2^\lambda \|\boldsymbol{\psi}'\|_1$. Plugging this into (16), we obtain

$$2\Delta(1 - 2^{-|\alpha| 2^{-\lambda}\Delta})(\|\boldsymbol{\psi}\|_\infty + \|\boldsymbol{\psi}'\|_1) \ . \tag{17}$$

Given that $1 - 2^{-u} < \log(2) u$ for all $u > 0$, this simplifies to

$$2\log(2)\Delta^2(\|\boldsymbol{\psi}\|_\infty + \|\boldsymbol{\psi}'\|_1)\, |\alpha| 2^{-\lambda} \ . \tag{18}$$

Since $|\boldsymbol{g}(u)| = 1$, the modulus of the first term in (15) is $|\widehat{\boldsymbol{\psi}}_\lambda(\log(2^\alpha)2^{\alpha t})| = |\widehat{\boldsymbol{\psi}}(\log(2^\alpha)2^{\alpha t - \lambda})|$. The triangle inequality then establishes (12) with $|\boldsymbol{\varepsilon}(t, \lambda)|$ bounded by (18). $\square$

Lemma 2. *Define $\psi_\lambda(t)$, $\Psi_{\mu,\ell,s}(t,\lambda)$, and c_0 as in Theorem 1 and let*

$$t_0(\lambda) = \frac{\lambda}{\alpha} - \frac{\log\log 2^\alpha}{\log 2^\alpha}.$$

Given

$$\mathbf{Y}(t,\lambda) = |\widehat{\psi}\left(\log(2^\alpha)\,2^{\alpha t-\lambda}\right)| \ ,$$

its two-dimensional wavelet modulus decomposition satisfies

$$\begin{aligned}&|\mathbf{Y} * \Psi_{\mu,\ell,s}|(t,\lambda) \qquad (19)\\ &= \frac{c_0|\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda))}{\alpha}\left|\widehat{\psi}^{(\mathrm{f})}\left(-\frac{s2^{\mu-\ell}}{\alpha}\right)\right| + \varepsilon(t,\lambda,\mu,\ell,s) \ ,\end{aligned}$$

where $|\psi^{(\mathrm{t})}_\mu|(t) = 2^\mu|\psi^{(\mathrm{t})}|(2^\mu t)$, and

$$|\varepsilon(t,\lambda,\mu,\ell,s)| \leq C(2^{2\mu}|\alpha|^{-2} + 2^{2\mu-\ell}|\alpha|^{-2}) \ ,$$

for some $C > 0$ depending only on $\psi(t)$, $\psi^{(\mathrm{t})}(t)$, $\psi^{(\mathrm{f})}(\lambda)$.

Proof. Since $|\widehat{\psi}(\omega)|$ is maximized at $\omega = 1$, fixing λ, the maximum of $\mathbf{Y}(t,\lambda)$ is at $t_0(\lambda)$. For small enough μ, $\mathbf{Y}(t,\lambda)$ approximates a Dirac delta function centered at $t_0(\lambda)$. We exploit this when convolving $\mathbf{Y}(t,\lambda)$ by $\psi^{(\mathrm{t})}_\mu(t)$.

Approximating $\psi^{(\mathrm{t})}(u)$ with its value at $u = t - t_0(\lambda)$ gives

$$\begin{aligned}\mathbf{Y}(\cdot,\lambda) * \psi^{(\mathrm{t})}_\mu(t) &= \int_{\mathbb{R}} \mathbf{Y}(t-u,\lambda)\psi^{(\mathrm{t})}_\mu(u)\,\mathrm{d}u\\ &= \int_{\mathbb{R}} |\widehat{\psi}(\log(2^\alpha)2^{\alpha(t-u)-\lambda})| \times\\ &\quad (\psi^{(\mathrm{t})}_\mu(t-t_0(\lambda)) + \varepsilon_1(t,\lambda,\mu,u))\,\mathrm{d}u \ ,\end{aligned}$$

where $|\varepsilon_1(t,\lambda,\mu,u)| \leq |t - t_0(\lambda) - u|\,\|\psi^{(\mathrm{t})\prime}_\mu\|_\infty$. Setting $\varepsilon_2(t,\lambda,\mu) = \int_{\mathbb{R}} |\widehat{\psi}(\log(2^\alpha)2^{\alpha(t-u)-\lambda})|\varepsilon_1(t,\lambda,\mu,u)\,\mathrm{d}u$ gives

$$\mathbf{Y}(\cdot,\lambda) * \psi^{(\mathrm{t})}_\mu(t) = c_0\alpha^{-1}\psi^{(\mathrm{t})}_\mu(t-t_0(\lambda)) + \varepsilon_2(t,\lambda,\mu) \ , \quad (20)$$

using a change of variables, where $c_0 = \int_{\mathbb{R}} |\widehat{\psi}(2^u)|\,\mathrm{d}u < \infty$.

We bound $\varepsilon_2(t,\lambda,\mu)$ through

$$\begin{aligned}|\varepsilon_2(t,\lambda,\mu)| &\leq \int_{\mathbb{R}} |\widehat{\psi}(\log(2^\alpha)2^{\alpha(t-u)-\lambda})|\,|\varepsilon_1(t,\lambda,\mu,u)|\,\mathrm{d}u\\ &\leq \|\psi^{(\mathrm{t})\prime}_\mu\|_\infty \int_{\mathbb{R}} |\widehat{\psi}(\log(2^\alpha)2^{\alpha(t-u)-\lambda})|\,|t - t_0(\lambda) - u|\,\mathrm{d}u \ .\end{aligned}$$

The change of variables $t - t_0(\lambda) - u \mapsto \alpha^{-1}u$ now gives

$$|\varepsilon_2(t,\lambda,\mu)| \leq |\alpha|^{-2}2^{2\mu}\|\psi^{(\mathrm{t})\prime}\|_\infty \int_{\mathbb{R}} |\widehat{\psi}(2^u)|\,|u|\,\mathrm{d}u, \quad (21)$$

where we have used $\|\psi^{(\mathrm{t})\prime}_\mu\|_\infty = 2^{2\mu}\|\psi^{(\mathrm{t})\prime}\|_\infty$.

We now convolve (20) by $\psi^{(\mathrm{f})}_{\ell,s}(\lambda) = 2^\ell\psi^{(\mathrm{f})}(s2^\ell\lambda)$. At high ℓ, this wavelet will mostly capture phase variation. To see this, we factorize $\psi^{(\mathrm{t})}_\mu(t)$ into an envelope and a phase, yielding $|\psi^{(\mathrm{t})}_\mu|(t)\exp(2\pi\mathrm{i}\,2^\mu t))$. The convolution then becomes

$$\begin{aligned}&c_0\alpha^{-1}\psi^{(\mathrm{t})}_\mu(t-t_0(\cdot)) * \psi^{(\mathrm{f})}_{\ell,s}(\lambda)\\ &= c_0\alpha^{-1}\int_{\mathbb{R}} |\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda-\gamma)) \times\\ &\quad \exp(2\pi\mathrm{i}\,2^\mu(t-t_0(\lambda-\gamma)))\psi^{(\mathrm{f})}_{\ell,s}(\gamma)\,\mathrm{d}\gamma \ . \qquad (22)\end{aligned}$$

We now make the approximation

$$|\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda-\gamma)) = |\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda)) + \varepsilon_3(t,\lambda,\mu,\gamma) \ ,$$

where $|\varepsilon_3(t,\lambda,\mu,\gamma)| \leq \||\psi^{(\mathrm{t})}_\mu|'\|_\infty|t_0(\lambda-\gamma)-t_0(\lambda)|$. Plugging this into (22), we obtain

$$\begin{aligned}&c_0\alpha^{-1}|\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda)) \times \qquad (23)\\ &\int_{\mathbb{R}} \exp(2\pi\mathrm{i}\,2^\mu(t-t_0(\lambda-\gamma)))\psi^{(\mathrm{f})}_{\ell,s}(\gamma)\,\mathrm{d}\gamma + \varepsilon_4(t,\lambda,\mu,\ell,s) \ ,\end{aligned}$$

where $\varepsilon_4(t,\lambda,\mu,\ell,s)$ equals

$$c_0\alpha^{-1}\int_{\mathbb{R}} \varepsilon_3(t,\lambda,\mu,\gamma)\exp(2\pi\mathrm{i}\,2^\mu(t-t_0(\lambda-\gamma)))\psi^{(\mathrm{f})}_{\ell,s}(\gamma)\,\mathrm{d}\gamma \ .$$

Since $t_0(\lambda-\gamma) = t_0(\lambda) - \alpha^{-1}\gamma$, the first term in (23) is

$$c_0\alpha^{-1}|\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda))\mathrm{e}^{\mathrm{i}2^\mu(t-t_0(\lambda))}\widehat{\psi}^{(\mathrm{f})}_{\ell,s}(-2^\mu\alpha^{-1}) \ . \quad (24)$$

The same property of $t_0(\lambda)$ lets us bound $\varepsilon_4(t,\lambda,\mu,\ell,s)$ by

$$\begin{aligned}|\varepsilon_4(t,\lambda,\mu,\ell,s)| &\leq c_0|\alpha|^{-2}\||\psi^{(\mathrm{t})}_\mu|'\|_\infty \int_{\mathbb{R}} |\gamma||\psi^{(\mathrm{f})}_{\ell,s}(\gamma)|\,\mathrm{d}\gamma\\ &= c_0|\alpha|^{-2}2^{2\mu-\ell}\||\psi^{(\mathrm{t})}|'\|_\infty \int_{\mathbb{R}} |\gamma||\psi^{(\mathrm{f})}(\gamma)|\,\mathrm{d}\gamma \ , \qquad (25)\end{aligned}$$

which follows from change of variables and from $\||\psi^{(\mathrm{t})}_\mu|'\|_\infty = 2^{2\mu}\||\psi^{(\mathrm{t})}|'\|_\infty$. We must also convolve $\varepsilon_2(t,\lambda,\mu)$ with $\psi^{(\mathrm{f})}_{\ell,s}(\lambda)$. Since $\|\psi^{(\mathrm{f})}_{\ell,s}\|_1 = \|\psi^{(\mathrm{f})}\|_1$ for all ℓ, s, we have

$$|\varepsilon_2(t,\cdot,\mu) * \psi^{(\mathrm{f})}_{\ell,s}(\lambda)| \leq \|\varepsilon_2(t,\cdot,\mu)\|_\infty\,\|\psi^{(\mathrm{f})}\|_1 \ . \quad (26)$$

Combining (20) with (24) and taking the modulus yields (19) since $\widehat{\psi}^{(\mathrm{f})}_{\ell,s}(\omega) = \widehat{\psi}^{(\mathrm{f})}(s2^{-\ell}\omega)$, where the bound on $\varepsilon(t,\lambda,\mu,\ell,s)$ follows from (21), (25), (26), and the triangle inequality. □

Proof of Theorem 1. Lemma 1 gives

$$\mathbf{X}(t,\lambda) = |x * \psi_\lambda|(t) = |\widehat{\psi}(\log(2^\alpha)2^{\alpha t-\lambda})| + \varepsilon_1(t,\lambda) \ ,$$

where $|\varepsilon_1(t,\lambda)| \leq C_1|\alpha|2^{-\lambda}$ for some $C_1 > 0$. We now convolve $\mathbf{X}(t,\lambda)$ with $\psi^{(\mathrm{t})}_\mu(t)$ in time $\psi^{(\mathrm{f})}_{\ell,s}(\lambda)$ in log-frequency and take the modulus. Lemma 2 approximates the convolution of the first term. For the second term, we observe that

$$\|\varepsilon_1(\cdot,\lambda) * \psi^{(\mathrm{t})}_\mu\|_\infty \leq \|\varepsilon_1(\cdot,\lambda)\|_\infty\|\psi^{(\mathrm{t})}\|_1 \leq C_2|\alpha|2^{-\lambda} \ ,$$

for some $C_2 > 0$, since $\|\psi^{(\mathrm{t})}_\mu\|_1 = \|\psi^{(\mathrm{t})}\|_1$ for all μ. Now,

$$\begin{aligned}|\varepsilon_1 * \Psi_{\mu,\ell,s}(t,\lambda)| &\leq C_2|\alpha|\int_{\mathbb{R}} 2^{-(\lambda-\mu)}|\psi^{(\mathrm{f})}_{\ell,s}(\mu)|\,\mathrm{d}\mu\\ &= C_2|\alpha|2^{-\lambda}\int_{|\mu|\leq 2^{-\ell}A} 2^\mu|\psi^{(\mathrm{f})}_{\ell,s}(\mu)|\,\mathrm{d}\mu\\ &\leq C_2|\alpha|2^{-\lambda}2^{2^{-\ell}A}\|\psi^{(\mathrm{f})}_{\ell,s}\|_1 = C_3|\alpha|2^{-\lambda+2^{-\ell}A} \ ,\end{aligned}$$

for some $C_3 > 0$, since $\psi^{(\mathrm{f})}$ is supported on $[-A, A]$.

As a result,

$$\begin{aligned}&|\mathbf{X} * \Psi_{\mu,\ell,s}(t,\lambda)| \qquad (27)\\ &= \frac{c_0}{\alpha}|\psi^{(\mathrm{t})}_\mu|(t-t_0(\lambda))\left|\widehat{\psi}^{(\mathrm{f})}\left(-\frac{s2^{\mu-\ell}}{\alpha}\right)\right| + \varepsilon_2(t,\lambda,\mu,\ell,s) \ ,\end{aligned}$$

where

$$|\varepsilon_2(t,\lambda,\mu,\ell,s)| \leq C(|\alpha|2^{-\lambda+2^{-\ell}A} + |\alpha|^{-2}2^{2\mu} + |\alpha|^{-2}2^{2\mu-\ell}) \ .$$

Since this bound is constant in t and $\|\phi_T\|_1 = \|\phi\|_1$ for all T, it still holds after convolving (27) with $\phi_T(t)$. □

REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," in *Proc. ASRU*. IEEE, Dec 1997, pp. 140–147.

[3] J. Thompson and L. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," in *Proc. ICASSP*, vol. 5. IEEE, 2003, pp. 397–400.

[4] M. Slaney and R. Lyon, "On the importance of time–A temporal representation of sound," in *Visual representations of speech signals*, S. B. M. Cooke and M. Crawford, Eds. Wiley, 1993, pp. 95–116.

[5] R. D. Patterson, "Auditory images: How complex sounds are represented in the auditory system," *Acoustical Science and Technology*, vol. 21, no. 4, pp. 183–190, 2000.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[8] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 6645–6649.

[9] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. ISCS*. IEEE, 2010.

[10] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[11] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[12] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, pp. 4114–4128, 2014.

[13] V. Chudáček, J. Andén, S. Mallat, P. Abry, and M. Doret, "Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1100–1108, 2014.

[14] R. Talmon, S. Mallat, H. Zaveri, and R. R. Coifman, "Manifold learning for latent variable inference in dynamical systems," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3843–3856, 2015.

[15] J. Sulam, Y. Romano, and R. Talmon, "Dynamical system classification with diffusion embedding for ECG-based person identification," *Signal Processing*, vol. 130, pp. 403–411, 2017.

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

[17] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, 2006.

[18] J. C. Brown, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Am.*, vol. 92, no. 5, p. 2698, 1992.

[19] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[20] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1999.

[21] B. Logan, "Mel frequency ceptral coefficients for music modeling," in *Proc. ISMIR*, 2000.

[22] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov 2006.

[23] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. MM*. ACM, 2015, pp. 1015–1018.

[24] M. Dörfler, R. Bammer, and T. Grill, "Inside the spectrogram: Convolutional neural networks in audio processing," in *Proc. SampTA*, 2017, pp. 152–155.

[25] M. Dörfler, T. Grill, R. Bammer, and A. Flexer, "Basic filters for convolutional neural networks applied to music: Training or design?" 2017, submitted. arXiv:1709.02291.

[26] A. Papoulis, *Signal Analysis*. McGraw-Hill Education, 1977.

[27] I. Waldspurger, "Exponential decay of scattering coefficients," in *Proc. SampTA*, 2017, pp. 143–146.

[28] J. McDermott and E. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[29] M. Schädler, B. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[30] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction." in *Proc. Interspeech*, 2002.

[31] K. Siedenburg, I. Fujinaga, and S. McAdams, "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," *J. New Music Res.*, vol. 45, no. 1, pp. 27–41, 2016.

[32] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, 2013.

[33] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *Proc. CVPR*. IEEE, 2013, pp. 1233–1240.

[34] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Proc. Interspeech*, 2014.

[35] P. Hamel, M. E. P. Davies, K. Yoshii, and M. Goto, "Transfer learning in MIR: Sharing learned latent representations for music audio classification and similarity," in *Proc. ISMIR*, 2013.

[36] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. NIPS*, 2016, pp. 892–900.

[37] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. ICCV*. IEEE, 2017.

[38] M. A. Richards, *Fundamentals of radar signal processing*. McGraw-Hill Education, 2005.

[39] D. Schwarz, "State of the art in sound texture synthesis," in *Proc. DAFx*, 2011.

[40] J. Bruna and S. Mallat, "Audio texture synthesis with scattering moments," 2013, unpublished. arXiv:1311.0407.

[41] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, 2013.

[42] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[43] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, 1989.

[44] P. Clarkson and P. Moreno, "On the use of support vector machines for phonetic classification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, pp. 585–588, 1999.

[45] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[46] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[47] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intell. Syst. and Technol.*, vol. 2, pp. 27:1–27:27, 2011.

[48] H. Tang, W. Wang, K. Gimpel, and K. Livescu, "Discriminative segmental cascades for feature-rich phone recognition," in *Proc. ASRU*. IEEE, 2015, pp. 561–568.

[49] V. Peddinti, T. N. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep scattering spectrum with deep neural networks," in *Proc. ICASSP*. IEEE, 2014, pp. 210–214.

[50] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum–deep siamese network pipeline for unsupervised acoustic modeling," in *Proc. ICASSP*. IEEE, 2016, pp. 4965–4969.

[51] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proc. ICCV*, 2017.

[52] D. Bogdanov, A. Porter, P. Herrera, and X. Serra, "Cross-collection evaluation for music classification tasks," in *Proc. ISMIR*, 2016.

[53] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. ISMIR*, 2014.

[54] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 174–186, 2009.

[55] V. Lostanlen and C.-E. Cella, "Deep convolutional networks in the pitch spiral for music instrument classification," in *Proc. ISMIR*, 2016.

[56] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. MM*. ACM, 2014.

[57] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1733–1746, 2015.

[58] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

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