2015 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2015, BOSTON, USA

JOINT TIME-FREQUENCY SCATTERING FOR AUDIO CLASSIFICATION

Joakim Andén

PACM
Princeton University, Princeton, NJ, USA
janden@math.princeton.edu

Vincent Lostanlen, Stéphane Mallat *

Département d'Informatique
École normale supérieure, Paris, France

ABSTRACT

We introduce the joint time-frequency scattering transform, a time shift invariant descriptor of time-frequency structure for audio classification. It is obtained by applying a two-dimensional wavelet transform in time and log-frequency to a time-frequency wavelet scalogram. We show that this descriptor successfully characterizes complex time-frequency phenomena such as time-varying filters and frequency modulated excitations. State-of-the-art results are achieved for signal reconstruction and phone segment classification on the TIMIT dataset.

Index Terms— audio classification, invariant descriptors, time-frequency structure, wavelets, convolutional networks

1. INTRODUCTION

Signal representations for classification need to capture discriminative information from signals while remaining invariant to irrelevant variability. This allows accurate classifiers to be trained using a limited set of labeled examples. In audio classification, classes are often invariant to time shifts, making time shift invariant descriptors particularly useful.

Mel-frequency spectral coefficients are time-frequency descriptors invariant to time shifts up to 25 ms and form the basis for the popular mel-frequency cepstral coefficients (MFCCs) [1]. These can be seen as the time-averaging of a wavelet scalogram, which is obtained by constant-Q wavelet filtering followed by a complex modulus [2]. The time scattering transform refines this while maintaining invariance by further decomposing each frequency band in the wavelet scalogram using another scalogram [2, 3]. The result can be seen as the output of a multilayer convolutional network [3]. Classification experiments have demonstrated the importance of this second layer, which captures amplitude modulation [2]. Yet because it decomposes each frequency band separately, it fails to capture more complex time-frequency structure such as time-varying filters and frequency modulation, which are important in many classification tasks.

*This work is supported by the ERC InvariantClass 320959.

Section 2 introduces the joint time-frequency scattering transform which extends the time scattering by replacing the second-layer wavelet transform in time with a two-dimensional wavelet transform in time and log-frequency. This is inspired by the neurophysiological models of S. Shamma, where the scalogram-like output of the cochlea is decomposed using two-dimensional Gabor filters [4]. Section 3 shows that joint time-frequency scattering better captures the time-frequency structure of the scalogram by adequately characterizing time-varying filters and frequency modulation. This is illustrated in Section 4, which presents signal reconstruction results from joint time-frequency scattering coefficients that are comparable to state-of-the-art algorithms and superior to time scattering reconstruction. In Section 5, the joint time-frequency scattering transform is shown to achieve state-of-the-art performance for phone segment classification on the TIMIT dataset, demonstrating the importance of properly describing time-frequency structure. All figures and numerical results are reproducible using a MATLAB software package available at `http://www.di.ens.fr/data/scattering/`.

2. JOINT TIME-FREQUENCY SCATTERING

The wavelet scalogram of a signal represents time-frequency structure through a wavelet decomposition, which filters a signal using a constant-Q wavelet filter bank. A time scattering transform captures the temporal evolution of each frequency band by another set of wavelet convolutions in time. It does not fully capture the time-frequency structure of the scalogram since it neglects correlation across frequencies. The joint time-frequency scattering remedies this by replacing the one-dimensional wavelet transform in time with a two-dimensional wavelet transform in time and log-frequency.

We denote the Fourier transform of a signal $x(t)$ by $\widehat{x}(\omega) = \int x(u)e^{-i\omega u}\,du$. An analytic mother wavelet is a complex filter $\psi(t)$ whose Fourier transform $\widehat{\psi}(\omega)$ is concentrated over the frequency interval $[1 - 2^{1/2Q}, 1 + 2^{1/2Q}]$. Dilations of this mother wavelet defines a family of filters

978-1-4673-7454-5/15/$31.00

centered at frequencies $\lambda_1 = 2^{j_1/Q}$ for $j_1 \in \mathbb{Z}$, given by

$$\psi_{\lambda_1}(t) = \lambda_1 \psi(\lambda_1 t) \quad \Longrightarrow \quad \widehat{\psi}_{\lambda_1}(\omega) = \widehat{\psi}({\lambda_1}^{-1}\omega) \; . \quad (1)$$

Letting $\log u$ denote the base-two logarithm of u, we observe that $\log \lambda_1 = j_1/Q$ samples each octave uniformly with Q wavelets. The temporal support of ψ_{λ_1} is approximately $2\pi Q/\lambda_1$, so to ensure that the support does not exceed some fixed window size T, we define ψ_{λ_1} using (1) only when $\lambda_1 \geq 2\pi Q/T$. The low-frequency interval $[0, 2\pi Q/T]$ is covered by linearly spaced filters of constant bandwidth $2\pi/T$. However, to simplify explanations, we shall treat all filters as dilations of ψ.

The wavelet transform convolves a signal x with a wavelet filter bank. Its complex modulus is the wavelet scalogram

$$x_1(t, \log \lambda_1) = |x * \psi_{\lambda_1}(t)| \; , \quad \text{for all } \lambda_1 > 0 \; , \quad (2)$$

an image uniformly sampled in t and $\log \lambda_1$. Here $x_1(t, \log \lambda_1)$ represents time-frequency intensity in the interval of duration $2\pi Q/\lambda_1$ centered at t and the frequency band of bandwidth λ_1/Q centered at λ_1. Figure 1(a) shows a sample scalogram.

While a rich descriptor of time-frequency structure, the scalogram is not time shift invariant. The scattering transform ensures invariance to time shifts smaller than T by time-averaging with a low-pass filter ϕ_T of support T, giving

$$S_1 x(t, \log \lambda_1) = x_1(\cdot, \log \lambda_1) * \phi_T(t) \; , \quad (3)$$

known as first-order scattering coefficients. These approximate mel-frequency spectral coefficients [2].

To recover the high frequencies lost when averaging by ϕ_T in (3), x_1 is convolved with a second set of wavelets ψ_{λ_2}. Computing the modulus gives

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1(\cdot, \log \lambda_1) * \psi_{\lambda_2}(t)| \; . \quad (4)$$

As before, averaging in time creates invariance and yields

$$S_2 x(t, \log \lambda_1, \log \lambda_2) = x_2(\cdot, \log \lambda_1, \log \lambda_2) * \phi_T(t) \; . \quad (5)$$

These are called second-order time scattering coefficients. They supplement the first order (and by extension mel-frequency spectral coefficients) by capturing the temporal variability of the scalogram [3]. Higher-order coefficients can also be computed by repeating the same procedure.

A representation similar to second-order time scattering is the constant-Q modulation spectrogram, which computes the spectrogram of each frequency band and averages using a constant-Q scale [5]. The cascade structure of alternating convolutions and modulus nonlinearities is also shared by convolutional neural networks, which enjoy significant success in many classification tasks [6, 7].

In addition to time shift invariance, the scattering transform is also stable to time warping due to the constant-Q structure of the wavelets [3]. This is useful in audio classification where small deformations do not alter class membership.

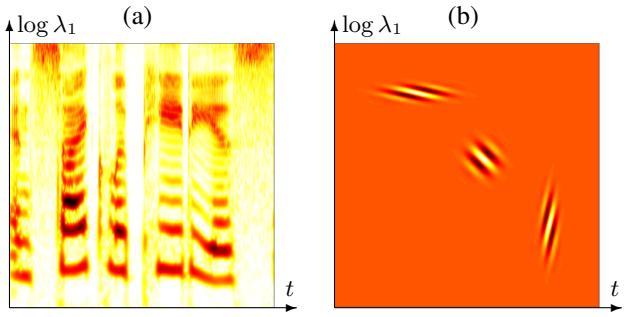

Fig. 1. (a) Scalogram of a woman saying the word "encyclopedias" with $Q = 8$ and $T = 32$ ms. (b) Three examples of real parts of wavelets Ψ_{λ_2} in the $(t, \log \lambda_1)$-plane.

In many audio classification tasks, such as speech recognition, classes are invariant to frequency transposition. In this case classifiers benefit from transposition-invariant descriptors. The time scattering transform is made invariant to transposition by computing a frequency scattering transform along $\log \lambda_1$, improving classification accuracy for such tasks [2].

While the time scattering transform successfully describes the average spectral envelope and amplitude modulation of a signal [2], it decomposes and averages each frequency band separately and so cannot capture the relationship between local temporal structure across frequency. Hence it does not adequately characterize more complex time-frequency phenomena, such as time-varying filters and frequency modulation.

To capture the variability of the scalogram across both time and log-frequency, we replace the one-dimensional wavelet transform in time with a two-dimensional wavelet transform in time and log-frequency. This follows the cortical model introduced by S. Shamma, where a sound is decomposed by the cochlea into a wavelet scalogram which is then convolved by two-dimensional Gabor filters in the auditory cortex [4]. Representations based on this cortical model have performed well in audio classification [8, 9], but often lack a mathematical justification.

Let us define the two-dimensional wavelet

$$\Psi_{\lambda_2}(t, \log \lambda_1) = \psi_\alpha(t) \, \psi_\beta(\log \lambda_1) \; , \quad (6)$$

where $\lambda_2 = (\alpha, \beta)$ for $\alpha \geq 0$ and $\beta \in \mathbb{R}$. The time wavelet $\psi_\alpha(t)$ is calculated with a dilation by α^{-1} for $\alpha \geq 2\pi/T$ as in (1), giving a Fourier transform centered at α. For these wavelets, $Q = 1$, although the notation remains the same. Similarly, we abuse notation and define the log-frequency wavelet by dilating a mother wavelet ψ to get

$$\psi_\beta(\log \lambda_1) = \beta \, \psi(\beta \log \lambda_1) \; . \quad (7)$$

The identity of the wavelet will be clear from context.

The Fourier transform of ψ_β is centered at the frequency β. We shall refer to this "frequency" parameter β associated

with the log-frequency variable $\log \lambda_1$ as a "quefrency," with units of cycles per octave. Note that this is different from the standard quefrency, which is measured in seconds.

Since the two-dimensional Fourier transform of Ψ_{λ_2} is centered at (α, β), it oscillates along the slope β/α. Its support in time and log-frequency is $2\pi/\alpha$ by $2\pi/\beta$. Sample wavelets are shown in Figure 1(b). To ensure invertibility of the wavelet transform, the Fourier transforms of Ψ_{λ_2} must cover a half-plane, hence the requirement that β take negative values. The sign of β determines the direction of oscillation.

The wavelet transform of x_1 is calculated through a two-dimensional convolution with Ψ_{λ_2}. Taking the modulus gives

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1 * \Psi_{\lambda_2}(t, \log \lambda_1)| \,, \qquad (8)$$

where $\log \lambda_2 = (\log \alpha, \log |\beta|, \operatorname{sgn} \beta)$. Similarly to (5), second-order time-frequency scattering coefficients are computed by time-averaging, which yields

$$S_2 x(t, \log \lambda_1, \log \lambda_2) = x_2(\cdot, \log \lambda_1, \log \lambda_2) * \phi_T(t) \,. \quad (9)$$

Higher-order coefficients are obtained as before by repeating the above process. In contrast to the time scattering transform, the joint descriptor successfully captures the two-dimensional structure of the scalogram at time scales below T.

To obtain frequency transposition invariance, it would suffice to average both $S_1 x$ and $S_2 x$ along $\log \lambda_1$ using a frequency window. However, the amount of invariance needed may differ between classes. Since the invariant is created through a linear mapping – averaging along $\log \lambda_1$ – a discriminative linear classifier can learn the proper amount of invariance for each class [2].

Just as time scattering is invariant to deformation in time, the two-dimensional wavelet decomposition ensures that the frequency-averaged joint scattering transform is invariant to deformation of the scalogram in time and log-frequency. This is useful for many audio classification tasks, where classes are often invariant under small deformations of the scalogram.

3. SCATTERING TIME-FREQUENCY STRUCTURE

We apply the joint time-frequency scattering transform to two signal models: a fixed excitation convolved with a time-varying filter and an unfiltered frequency-modulated excitation. Both represent non-separable time-frequency structure and are insufficiently captured by the time scattering transform but well characterized by joint time-frequency scattering. These models do not model more advanced structures such as polyphony and inharmonicity, but allow us to explore the basic properties of the joint scattering transform.

3.1. Time-varying filter

Let us consider a harmonic excitation

$$e(t) = \frac{2\pi}{\xi} \sum_n \delta\left(t - \frac{2\pi n}{\xi}\right) = \sum_k e^{ik\xi t} \qquad (10)$$

of pitch ξ. The signal is then given by applying a time-varying filter $h(t, u)$ to $e(t)$, defined as

$$x(t) = \int e(t-u) h(t, u) \, du \,. \qquad (11)$$

Parseval's theorem now gives

$$x(t) = \frac{1}{2\pi} \int \widehat{e}(\omega) \widehat{h}(t, \omega) e^{i\omega t} \, d\omega \,, \qquad (12)$$

where $\widehat{h}(t, \omega)$ is the Fourier transform of $h(t, u)$ along u. Thus $x(t)$ is the inverse Fourier transform of $\widehat{e}(\omega)$ multiplied by a time-varying transfer function $\widehat{h}(t, \omega)$. These transforms are also known as pseudo-differential operators.

Time-varying filters appear in many audio signals and carry important information. For example, during speech production the vocal tract is deformed to produce a sequence of phones. This produces amplitude modulation, but also shifts formants in the spectral envelope, which can be modeled by a time-varying filter. Similarly, much of the instrument-specific information in a musical note is contained in the attack, which is often characterized by a changing spectral envelope. For these reasons, it is important for an audio descriptor to adequately capture time-varying filters.

For a suitable choice of λ_1 we can show that

$$S_1 x(t, \log \lambda_1) \approx |\widehat{\psi}_{\lambda_1}(k\xi)| \, |\widehat{h}(\cdot, \lambda_1)| * \phi_T(t) \,, \qquad (13)$$

where $k = \lfloor \lambda_1 / \xi \rceil$ is the index of the partial closest to λ_1, while for small enough quefrencies $|\beta|$

$$\begin{aligned} &S_2 x(t, \log \lambda_1, \log \lambda_2) \\ &\quad \approx C \xi^{-1} |\widetilde{h} * \Psi_{\lambda_2}(\cdot, \log \lambda_1)| * \phi_T(t) \end{aligned} \,, \qquad (14)$$

where C does not depend on x. Here $\widetilde{h}(t, \log \omega)$ is a weighted and log-scaled version of $|\widehat{h}(t, \omega)|$ given by $\widetilde{h}(t, \log \omega) = \omega |\widehat{h}(t, \omega)|$. First-order coefficients thus provide the time-averaged amplitude of $\widehat{h}$ sampled at the partials $k\xi$ since $|\widehat{\psi}_{\lambda_1}(k\xi)|$ is non-negligible only for $\lambda_1 \approx k\xi$. Furthermore, the second order approximates the two-dimensional scattering coefficients of the modified filter transfer function $\widetilde{h}$, capturing its time-frequency structure.

In contrast, the time scattering transform only characterizes separable time-varying filters h that can be written as the product of an amplitude modulation in time and a fixed filter. In this case the model reduces to the amplitude-modulated, filtered excitation considered in [2]. Time scattering and joint time-frequency scattering thus differ in that the latter captures the non-separable structure of h while the former only describes its separable structure.

To justify (13) and (14), we proceed as in [2], convolving (12) with ψ_{λ_1} and taking the modulus to obtain

$$x_1(t, \log_2 \lambda_1) \approx |\widehat{h}(t, \lambda_1)| \sum_k |\widehat{\psi}_{\lambda_1}(k\xi)| \,, \qquad (15)$$

for $\widehat{h}(t,\omega)$ smooth enough and $\lambda_1/Q < \xi$. In this case at most one partial $k\xi \approx \lambda_1$ is found in the support of the wavelet so the sum only contains one non-negligible term when $k = \lfloor \lambda_1/\xi \rceil$. Averaging in time yields (13). Furthermore, we note that as a function of $\log\lambda_1$, the sequence of partials $\sum_k |\widehat{\psi}_{\lambda_1}(k\xi)|$ can be approximated at large scale by $C\lambda_1\xi^{-1}$. For small $|\beta|$, ψ_β is very regular in $\log\lambda_1$. If $|\widehat{h}(t,\omega)|$ is also smooth enough along ω, we can therefore replace the sum of partials by $C\lambda_1\xi^{-1}$ when convolving x_1 with Ψ_{λ_2}. Rewriting the convolution using $\widetilde{h}$ then yields

$$x_1 * \Psi_{\lambda_2}(t,\log\lambda_1) \approx C\xi^{-1}\widetilde{h} * \Psi_{\lambda_2}(t,\log\lambda_1)\,. \qquad (16)$$

Taking the modulus and averaging then gives (14).

3.2. Frequency modulation

We now consider an excitation of varying pitch

$$x(t) = \sum_k e^{ik\theta(t)}\,. \qquad (17)$$

At time t, x has instantaneous pitch $\theta'(t)$ and relative pitch variation $\theta''(t)/\theta'(t)$. This carries important information in many sounds, such as tonal speech, bioacoustic signals, and music (e.g. for vibratos and glissandi). A good audio descriptor should therefore adequately describe such pitch changes.

For appropriate λ_1 and T, we can show that

$$S_1x(t,\log\lambda_1) \approx |\widehat{\psi}_{\lambda_1}(k\theta'(\cdot))| * \phi_T(t)\,, \qquad (18)$$

where $k = \lfloor \lambda_1/\theta'(t) \rceil$ as before. Furthermore, for $|\beta|$ large,

$$\begin{aligned} &S_2x(t,\log\lambda_1,\log\lambda_2) \\ &\quad\approx C\left(S_1x(t,\cdot) * \phi_{2\pi/\beta}(\log\lambda_1)\right)\left|\widehat{\psi}\left(-\tfrac{\beta\theta''(t)}{\alpha\theta'(t)}\right)\right|\,, \end{aligned} \qquad (19)$$

where C is independent of x.

While first-order joint scattering coefficients provide an average of the instantaneous pitch $\theta'(t)$ over the interval of duration T, the second order describes the rate of pitch variation $\theta''(t)/\theta'(t)$. Indeed, for fixed t and λ_1, S_2x is maximized along the line $\alpha/\beta = -\theta''(t)/\theta'(t)$, and so captures this frequency modulation structure. The time scattering transform, in contrast, only captures the bumps in each frequency band induced by the varying pitch, ignoring its frequency structure.

To see why (18) and (19) hold, we linearize $\theta(t)$ over the support of ψ_{λ_1} when decomposing (17), which gives

$$x_1(t,\log\lambda_1) \approx |\widehat{\psi}_{\lambda_1}(k\theta'(t))|\,, \qquad (20)$$

provided that $\lambda_1/Q < \|\theta'\|_\infty$. As before, only the partial $k = \lfloor \lambda_1/\theta'(t) \rceil$ is contained in the frequency support of ψ_{λ_1}. Averaging in time gives (18). Each partial traces a curve along $\lambda_1 = k\theta'(t)$, so locally the scalogram x_1 can be approximated by sliding Dirac functions $C\delta(\log\lambda_1 - \log k\theta'(t))$ for some C. Convolving x_1 along $\log\lambda_1$ with ψ_β for $|\beta|$ large enough to capture only one line gives $C\psi_\beta(\log\lambda_1 - \log k\theta'(t))$. For a fixed λ_1, this is a complex exponential of instantaneous frequency $-\beta\theta''(t)/\theta'(t)$ multiplied by an envelope. Convolving this in time with a wavelet ψ_α on whose support the envelope is approximately constant then gives

$$\begin{aligned} &x_1 * \Psi_{\lambda_2}(t,\log\lambda_1) \\ &\quad\approx C\psi_\beta\left(\log\tfrac{\lambda_1}{k\theta'(t)}\right)\,\widehat{\psi}\left(-\tfrac{\beta\theta''(t)}{\alpha\theta'(t)}\right)\,, \end{aligned} \qquad (21)$$

Taking the modulus, we can replace $|\psi_\beta|$ with the low-pass filter $\phi_{2\pi/\beta}$. Assuming that $\theta''(t)/\theta'(t)$ is almost constant over an interval of duration T, averaging gives (19).

We note that the time-varying filter and frequency modulation models in (12) and (17) are complementary. For small quefrencies $|\beta|$, the joint scattering coefficients capture time-frequency structure over large frequency intervals, which is given by time-varying filters. Larger $|\beta|$ describe more localized behavior in log-frequency, like frequency modulation. This scale separation allows the joint scattering transform to simultaneously characterize both types of structures.

4. TIME-SHIFT INVARIANT RECONSTRUCTION

After having analyzed a given signal x with a scattering transform, synthesizing a new signal y from the invariant coefficients S_1x and S_2x highlights what information is captured in the representation — and, conversely, what is lost. In this section, we use a backpropagation algorithm on stationary audio textures to qualitatively compare the joint scattering transform with other architectures.

The reconstruction y is first initialized with random noise, and then iteratively updated to converge to a local minimum of the functional

$$\|Sx - Sy\|^2 = \|S_1x - S_1y\|^2 + \|S_2x - S_2y\|^2 \qquad (22)$$

with respect to y. Since the forward computation of scattering coefficients consists of an alternated sequence of linear operators (wavelet convolutions) and modulus nonlinearities, the chain rule for gradient backpropagation yields a sequence of closed-form derivatives in the reverse order. The modulus nonlinearities are backpropagated by applying $|z(t)|' = \mathrm{Real}(z'(t)\,|z(t)|/z(t))$. In turn, the backpropagation of the wavelet transforms consists of convolving each frequency band by the complex conjugate of the corresponding wavelet and summing across bands [10].

To illustrate, we have synthesized a bird song recording using different scattering transforms. Here $T = 375\,\mathrm{ms}$ and is of the order of three bird calls (see Figure 2(a)). First-order coefficients S_1x yield the reconstruction in Figure 2(c). This fits the averaged mel-frequency spectrum of the target sound. Although this is sufficient when x is the realization of a Gaussian process, it does not convey the typical intermittency in

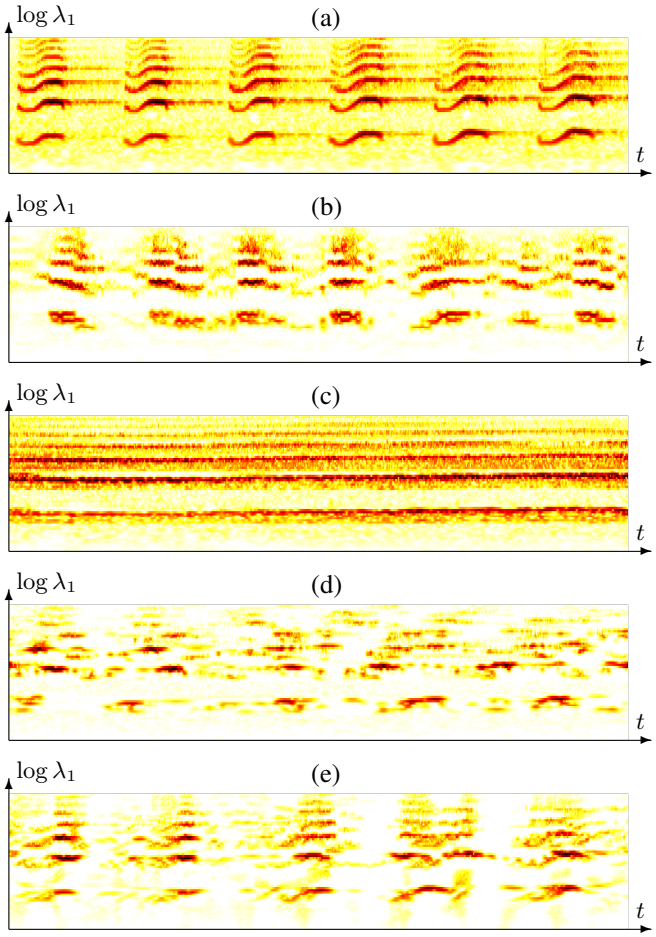

Fig. 2. Reconstructed bird calls from time-invariant coefficients. Top to bottom: (a) original $x(t)$, (b) from McDermott and Simoncelli representation [11], (c) from first-order scattering, (d) from first- and second-order time scattering, (e) from first- and second-order joint time-frequency scattering.

natural sounds. This is partly mitigated by adding second-order coefficients, giving the reconstruction in Figure 2(d), since these encode the amplitude modulation spectra in each acoustic subband. However, these spectra are not synchronized across subbands, so time scattering tends to synthesize auditory textures made of decorrelated impulses. In contrast, we observe that the reconstruction from joint scattering coefficients in Figure 2(e) is able to capture coherent structures in the time-frequency plane, such as joint modulations in amplitude and frequency. Notably, because of their chirping structure, bird calls are better synthesized with joint scattering. Indeed, recalling (19), chirps are represented with few nonzero coefficients in the basis of joint time-frequency wavelets. We believe that audio re-synthesis is greatly helped by this gain in sparsity. More experiments are available at `http://www.di.ens.fr/data/scattering/audio/`.

McDermott and Simoncelli [11] have built an audio texture synthesis algorithm based on a scattering-like transform along time, of which they compute cross-correlation statistics across λ_1 and across λ_2, as well as marginal moments (variance and skewness). Their representation is also able to synchronize frequency bands and recover amplitude modulation. Nevertheless, asymmetry in frequency modulation is lost. Indeed, while all bird calls from the original recording have an ascending instantaneous frequency, some of the chirps reconstructed with their method descend instead. Moreover, the higher-order statistics on which they rely are unstable to deformations and hence not suitable for classification purposes. In this section, we have shown that joint scattering may achieve comparable or better quality in audio re-synthesis, yet with only using stable features.

On the negative side, it must be noted that joint scattering is insufficient to capture temporal changes in harmonic structure. Indeed, partial tones which are several octaves apart are not likely to be correctly in tune — a limitation that we shall specifically address as a future work.

5. CLASSIFICATION

We evaluate the performance of the joint time-frequency scattering representation on phone segment classification using the TIMIT dataset [12]. The corpus consists of 6300 phrases, each of which has its constituent phone segments labeled with its position, duration, and identity. Given a position and duration, we want to identify the phone contained in the segment. This task is easier than the problem of continuous speech recognition, but provides a straightforward framework when evaluating signal representations for speech.

We follow the same setup as in [2]. Each phone is represented by a given descriptor applied to a 192-millisecond window centered on the phone along with the phone's log-duration. A Gaussian support vector machine (SVM) is used as a classifier through the LIBSVM library [13].

The SVM is a discriminatively trained, locally linear classifier. This means that, given enough training data, an SVM can learn the amount of averaging needed along $\log \lambda_1$ to gain the desired invariance [2]. We therefore present results for scattering transforms without averaging along $\log \lambda_1$.

Table 1 shows the results of the classification task. MFCCs calculated over the segment with a window size of 32 ms and concatenated to yield a single feature vector provide a baseline error rate of 18.3%. The non-scattering state of the art achieves 16.7% and is obtained using a committee-based hierarchical discriminative classifier on MFCC descriptors [14]. A convolutional network classifier applied to the log-scalogram with learned filters obtains 19.7% [7].

The time scattering transform is computed with $T = 32$ ms and $Q = 8$ up to the second order. As in previous experiments, we compute the logarithm of the scattering [2]. Since it better captures amplitude modulation, results improve with respect to MFCCs, achieving an error of 17.3%.

Representation	Error rate (%)
MFCCs	18.3
State of the art (excl. scattering) [14]	16.7
Time Scattering	17.3
Time Scattering + Freq. Scattering	16.1
Joint Time-Freq. Scattering	15.8

Table 1. Error rates in percent for the phone segment classification task. MFCCs and scattering transforms are computed with $T = 32$ ms and $Q = 8$.

Applying an unaveraged frequential scattering transform along $\log \lambda_1$ up to a scale of $K = 4$ octaves and computing the logarithm yields an error rate of 16.1%. As discussed earlier, transposition invariance counters speaker variability, and so improves performance. However, the frequency scattering is computed along $\log \lambda_1$ of a time scattering transform which has been averaged in time, so its discriminability also suffers from not capturing local correlations across frequencies.

Computing the joint time-frequency scattering transform for $K = 4$ octaves yields an error of 15.8%, an improvement compared to the time scattering transform with scattering along log-frequency. This illustrates the importance of the complex time-frequency structure that is captured by the joint scattering transform, and can be partly explained by the fact that the onset of many phones is characterized by rapid changes in formants, which can be modeled by time-varying filters. As we saw earlier, these are better described by time-frequency scattering compared to time scattering. However, the small window size T limits the loss of time-frequency structure in the time scattering transform. We therefore expect a greater improvement for tasks involving larger time scales.

The previous state of the art was obtained at 15.9% using a scattering transform with multiple Q factors [2]. This more ad hoc descriptor has many similarities with the joint scattering transform, but is difficult to study analytically.

6. CONCLUSION

We introduced the joint time-frequency scattering transform, which is a time-shift invariant representation stable to time-frequency warping. This representation characterizes time-varying filters and frequency modulation. Reconstruction experiments show how it successfully captures complex time-frequency structures of locally stationary signals. Finally, phone segment classification results demonstrate the value of adequately representing these structures for classification.

7. REFERENCES

[1] S.B. 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] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Sig. Proc.*, vol. 62, pp. 4114–4128, 2014.

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

[4] 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.

[5] J. Thompson and L. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," in *IEEE Int. Conf. on Acoust. Speech, and Sig. Proc.*, 2003, vol. 5, pp. V–397.

[6] Y. LeCun, K. Kavukvuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *IEEE Int. Symp. on Circuits and Syst.*, 2010.

[7] H. Lee, P. Pham, Y. Largman, , and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, 2009.

[8] 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.

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

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

[11] 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.

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

[13] 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, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] H. Chang and J. Glass, "Hierarchical large-margin gaussian mixture models for phonetic classification," in *Proc. ASRU*. IEEE, 2007, pp. 272–277.