Interpretable and Learnable Super-Resolution Time-Frequency Representation

Randall Balestriero RANDALLBALESTRIERO@GMAIL.COM
ECE Dept., Rice University, Houston, TX, USA

Hervé Glotin HERVE.GLOTIN@UNIV-TLN.FR
Université de Toulon, Aix Marseille Univ., CNRS, LIS, DYNI, INPS, Marseille, France

Richard G. Baraniuk RICHB@RICE.EDU
ECE Dept., Rice University, Houston, TX, USA

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

Abstract

We develop a novel interpretable and learnable time-frequency representation (TFR) that produces a super-resolved quadratic signal representation for time-series analysis; the proposed TFR is a Gaussian filtering of the Wigner-Ville (WV) transform of a signal parametrized with a few interpretable parameters. Our approach has two main hallmarks. First, by varying the filters applied onto the WV, our new TFR can interpolate between known TFRs such as spectrograms, wavelet transforms, and chirplet transforms. Beyond that, our representation can also reach perfect time and frequency localization, hence super-resolution; this generalizes standard TFRs whose resolution is limited by the Heisenberg uncertainty principle. Second, our proposed TFR is interpretable thanks to an explicit low-dimensional and physical parametrization of the WV Gaussian filtering. We demonstrate that our approach enables us to learn highly adapted TFRs and is able to tackle a range of large-scale classification tasks, where we reach higher performance compared to baseline and learned TFRs. Ours is to the best of our knowledge the first learnable TFR that can continuously interpolate between super-resolution representation and commonly employed TFRs based on a few learnable parameters and which preserves full interpretability of the produced TFR, even after learning.

Keywords: Learnable Time-Frequency Representation, Time-Series, Cohen Class, Wigner-Ville, Spectrogram, Wavelet, Chirplet, Gabor Transform, Audio Classification, Interpretability, Explainability, Super-Resolution, Speech Recognition, Environmental Sounds, biosonar, bioacoustics

1. Introduction

With the recent deep learning advances (LeCun et al., 2015; Goodfellow et al., 2016) there has been an exponential growth in the use of Deep Networks (DNs) on various time-series, including promising results on transients (Ferrari et al., 2020). However, the vast majority of DNs do not directly observe the time-series data but instead a handcrafted, a priori designed representation. Indeed, the vast majority of state-of-the-art methods combine DNs with some variant of a *Time-Frequency Representation* TFR (Lattner et al., 2019; Purwins et al., 2019; Liu et al., 2019). A TFR is an *image* representation of a time-serie obtained by convolving the latter with a filter-bank, such as wavelets or localized complex sinusoids (e.g., Gabor transform). Different filter-banks lead to different TFR families; that is, the produced *image* will highlight differently the events present in the time-serie input. It is for example common to employ wavelet transforms on biological signals and spectrogram on electrical or machine signals. In short, the different representations all inform about which frequency components are present at different times in the signal, but the *precision* of that information will vary.

The TFR-DN combination is powerful due to three major reasons: (i) the TFR contracts small transformations of the time-serie internal events such as translation, time and/or frequency warping (Bruna and Mallat, 2013) leading to more stable learning and faster DN convergence; (ii) the image representation allows to treat TFRs as a computer vision task where current DNs excels; (iii) the representation of the features of interests, such as phonemes for speech (Waibel et al., 1989), form very distinctive shapes in the TFR image, with dimensionality much smaller than the event's time-serie representation. In fact, a single coefficient of the TFR can encode information of possibly thousands of contiguous bins in the time-serie representation (Coifman et al., 1994; Le Pennec and Mallat, 2000; Logan et al., 2000). A coherent choice of TFR based on the data and task at hand will greatly affect the importance of each of the above points; hence, the choice of TFR has the potential to dim, or amplify, the above benefits further pushing the need to provide a TFR that can adapt to the data and task at hand.

Choosing the "best" TFR is a long lasting research problem in signal processing (Coifman and Wickerhauser, 1992; Jones and Baraniuk, 1994; Donoho, 1994). While TFR selection and adaptation was originally driven by signal reconstruction and compression (Xiong et al., 1998; Do and Vetterli, 2000; Cosentino et al., 2016), the recent developments of large supervised time-series datasets have led to novel learnable solutions that roughly fall into four camps. First, methods relying on the Wavelet Transform (WT) (Meyer, 1992). A WT is TFR with a constant-Q filter-bank based on dilations of a mother wavelet. In Balestriero et al. (2018) the learnability of the mother wavelet is introduced by means of a cubic spline parametrization of the mother wavelet to learn the mother wavelet shape. Second, methods relying on band-pass filters without center-frequency to bandwidth (Q) constraint such as the Short-Time Fourier Transform (STFT) (Allen, 1977). Khan and Yener (2018) propose to independently learn the center frequencies and bandwidths of a collection of Morlet wavelets (learning of those coefficients independently breaks the constant-Q property). Ravanelli and Bengio (2018) relies on learning the start and cutoff frequency of a bandpass sinc filter apodized with an hamming window. Those two methods similarly learn the location of the bandpass but use a different apodization window of a complex sine (Gaussian or hamming). In Cosentino and Aazhang (2020), it is proposed to learn a filter-bank through nonlinear transformations of a fixed filter, due to this, recovery of WT and chirp like filters was possible as special cases. Third, Zeghidour et al. (2018) proposes to learn Mel filters that are applied onto a spectrogram (modulus of STFT). Those filters linearly combine adjacent filters in the frequency axis which can be interpreted as learning a linear frequency subsampling of the spectrogram; learning the apodization window used to produce the spectrogram has also been developed in Jaillet and Torrésani (2007); Pei and Huang (2012). Finally, there are also methods relying on unconstrained DN layers applied on the time-series but with designed parameter initialization such that the induced representation (layer output) resembles (before training has started) an a priori determined target TFR. This has been done for chirplet transforms (Baraniuk and Jones, 1996; Glotin et al., 2017; Balestriero and Glotin, 2019) and for Mel-Spectrograms (Çakir and Virtanen, 2018).

All the above methods for learning the "best TFR" suffer from at least one of the three following limitations: (i) the inability to interpolate between different TFR families due to family specific parametrization of the learnable filter-banks; (ii) the inability to maintain interpretability of the filter-bank/TFR after learning; (iii) the inability to reach super-resolution in time and frequency to allow more precise representation. There is thus a need to *provide a universal learnable formulation able to interpolate between and within TFRs while preserving interpretability of the learned representation and with the ability to reach super-resolution*. The crucial component that allows us to produce such a learnable TFR is that as opposed to current solutions, we do not learn a filtering of the waveform signal, but a filtering of the Wigner-Ville transform (Wigner, 1932; Flandrin, 1998), which is a bilinear representation of the signal and from which common TFRs can be obtained through Gaussian filtering of that representation. Through an explicit parametrization of those Gaussians we allow adaptivity of those filters while maintaining interpretability of the filters and the resulting TFR.

We validate our method on multiple large scale datasets of speech, bird and marine bioacoustic and general sound event detection, and demonstrate that the proposed representation outperforms current learnable TFR techniques as well as fixed baseline TFRs regardless of the DN employed on top of those representations.

We summarize our contributions as follows:

[C1] We develop a Wigner-Ville Distribution based TFR with explicit interpretable parametrization able to reach super-resolution (Sec. 3.1), able to maintain interpretability of the filters and the representation at any point in the learning phase (Sec. 3.2), able to continuously interpolate between state-of-the-art TFRs (Sec. 3.3) and able to adapt its sensitivity to input transformations (Sec. 3.4).

[C2] We provide an efficient implementation allowing us to compute the proposed representation solely by means of Short-Time Fourier Transforms (Sec. 4.2). This allows GPU friendly computation and applicability of the method to large scale time-series dataset. We study the method complexity and provide details on our implementation (Sec. 4.1).

[C3] We validate our model and demonstrate how the proposed method outperforms other learnable TFRs as well as fixed expert based transforms on various datasets and across multiple DN architectures. We interpret the learned representations hinting at the key features of the signals needed to solve the task at hand (Sec. 4.3).

2. Background on Time-Frequency Representations

In this section we briefly recall the standard time-frequency representations that are commonly employed when dealing with time-series data.

Fourier and Spectrogram. Motivated by the understanding of physical phenomena, mathematical analysis tools have been created, notably the Fourier transform (Bracewell and Bracewell, 1986). Any signal $\boldsymbol{x}$ in $L^2(\mathbb{R})$ can be expressed in any basis of $L^2(\mathbb{R})$ (Mallat, 2008), the Fourier transform of a signal expresses $\boldsymbol{x}$ in the orthogonal basis formed by complex exponentials as $\mathcal{F}_x(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt$, providing a powerful representation for stationary signals. For non stationary signal analysis, where the observed signal carries different information dynamics throughout its duration, co-existence of the time and frequency variables in the representation is needed. One solution is offered by the Short Time Fourier Transform (STFT) (Allen, 1977) defined as follows

$$\text{STFT}_{x,g}(t,f) = \int_{-\infty}^{\infty} g(t-\tau)x(\tau)e^{-if\tau}d\tau, \tag{1}$$

with g an apodization window which vanishes when moving away from 0. This representation thus only assumes stationarity within the effective support of g which we denote by $\sigma_{\rm t}$. The squared modulus of the STFT is called the spectrogram as $\mathrm{SP}_{x,g}(t,f) = |\mathrm{STFT}_{x,g}(t,f)|^2$. The simplicity and efficiency of its implementation makes the spectrogram one of the most widely used TFR for non-stationary signals. Nonetheless, the spectrogram has a fundamental antagonism between its temporal and frequency resolution which depends on the apodization window spread σ_{t}. In a spectrogram, large σ_{t} allows high frequency resolution and poor time resolution and conversely for small σ_{t}. Most applications employ a Gabor (or truncated/approximated) apodization window, in which case $g(u) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-t^2/(2\sigma_{\mathrm{t}}^2)}$. Such spectrograms are denoted as Gabor transforms (Gabor, 1946). We will denote such a Gaussian window by $g_{\sigma_{\mathrm{t}}}$.

Wavelet Transform. A wavelet filter-bank is obtained by dilating a mother filter $\psi_0 \in L^2(\mathbb{R})$ with various *scales* s; the relationship between scale and center frequency f of the dilated filter is given by $s = 2^{S(1-\frac{f}{\pi})}$ with S the largest scale to be analyzed. Application of those dilated filters onto a signal leads to the wavelet transform WT (Mallat, 2008)

$$\mathrm{WT}_x(t,s) = (x \star \psi_s)(t), \text{ where } \psi_s(t) = \frac{1}{\sqrt{s}}\psi_0\left(\frac{t}{s}\right),$$

with $s > 0$. The relative position of the mother wavelet center frequency is not relevant as the scales can be adapted as desired, let consider here that ϕ_0 is placed at the highest frequency to be analyzed. Note that the above is often referred as the *continuous* wavelet transform due to the use of analytical mother filter ϕ_0. On the other hand there also exists a *discrete* wavelet transform (Haar, 1909; Daubechies, 1992) in which the discrete mother wavelet is often obtained by solving a system of equations (Jensen and la Cour-Harbo, 2001) which we do not consider in this study. To provide a representation invariant to the phase of the signal's internal events, it is common to apply a complex modulus after the convolution, or a squared complex modulus (Lostanlen et al., 2020), we thus consider in our study $\mathrm{WT}_x(t,s) = |x \star \psi_s|^2(t)$. As opposed to the spectrogram, the resolution of the WT varies with frequencies as the filters have a constant bandwidth to center frequency ratio. In a WT, high frequency atoms ψ_s with s close to 1 are localized in time offering a good time resolution but low frequency resolution. Conversely, for low frequency atoms with $s \gg 1$, the time resolution is low but the frequency resolution is high. As natural signals tend to be of small time duration when they are at high frequencies and of longer duration at low frequencies (Daubechies, 1990), the scalogram is one of the most adapted TFR for natural biological signals (Meyer, 1992).

Wigner-Ville Transform. The Wigner-Ville (WV) transform (or quasi probability or distribution) (Wigner, 1932) was originally derived for quantum mechanics (Moyal, 1949) and was only proposed as a TFR representation in Ville (1948). As opposed to the above TFRs which linearly transform $\boldsymbol{x}$ through a given basis or filter-bank, the WV transform is bilinear in $\boldsymbol{x}$. The transform combines complex sine filters as the Fourier transform and auto-correlations of the signal as follows

$$\mathrm{WV}_x(t,f) = \int_{-\infty}^{\infty} x\left(t - \frac{\tau}{2}\right) x^*\left(t + \frac{\tau}{2}\right) e^{-if\tau} d\tau. \tag{2}$$

Due to the auto-correlation term, computing the WV is demanding but provides a representation with perfect time and frequency localization as opposed to the spectrogram, wavelet transform, or any representation obtained through a linear filtering of the signal. This increased localization comes at the cost of introducing artifacts, or interference, in the representation (Daubechies and Planchon, 2002) and led to many variants of the WT with goal to reduce those artifacts e.g. the Pseudo Wigner Ville (Flandrin and Escudié, 1984) or the Smoothed Pseudo Wigner-Ville (Hlawatsch et al., 1995).

Filtering the Wigner-Ville Transform: (Affine) Cohen Class. One of the greatest property of the WV transform beyond perfect time-frequency localization lies in the ability to recover any TFR obtained from a linear filtering of a signal with a filter-bank by instead performing a linear filtering of the WV representation with a low-pass filter we denote as $\Pi \in L^2(\mathbb{R}^2)$ as in $(\mathrm{WV}_x \star \Pi)$ with $\star$ the 2-dimensional convolution. Collecting all possible low-pass filtered WV representation produces the Cohen class (Cohen, 1989) which we denote simply as $C_{\boldsymbol{x}}$ and defined as

$$\mathrm{C}_x = \{\mathrm{WV}_x \star \Pi | \Pi \in L^2(\mathbb{R}^2), \Pi \text{ low-pass } \}, \tag{3}$$

and where we have as special cases $\mathrm{SP}_{\boldsymbol{x},g} \subset \mathrm{C}_x$ for any Gaussian g. We thus see that any Gabor transform lives in this class, as well as many other known TFRs as we will discuss in the following section. An important extension of the Cohen class is given by the affine Cohen class (Flandrin and Rioul, 1990; Flandrin, 1993; Daubechies and Planchon, 2002) where not a single filter Π is employed with a 2-dimensional convolution but instead a time only convolution is employed with a collection of filters $\{\Pi_f, f \in \mathbb{R}^*\}$ with the constraint that $\Pi_f(\tau, \omega) = \Pi_0(f\tau, f\omega)$ given a "mother" low-pass filter Π_0. One should notice the close resemblance to the WT, in fact, WTs belong to the affine Cohen class.

This paper extends the Cohen class and the affine Cohen class by removing the affine transformation constraint relating the kernels Π_f, and by imposing a parametrization of those filters to allow learning, computational efficiency and interpretability of the filters and produced representation. Removing the affine constraint is key in allowing the WV transform filtering to reach a larger set TFRs e.g. chirplet transforms which are fundamental for many applications (Gribonval, 2001; Yin et al., 2002).

3. Learnable, Universal and Stable Wigner-Ville Based Signal Representation

In this section we develop our proposed signal representation which builds upon the Wigner-Ville Distribution. We first define our representation and study some key properties to finally propose a physics based parametrization that will allow for interpretability and robust learning. Throughout our development, we consider continuous time and frequency and will consider the finite sample, descrete case in the next section.

3.1. Wigner-Ville Based Signal Representation

We now define our transformation, coined the *K-transform*, which corresponds to applying a 2-dimensional truncated Gaussian kernel $\Phi_f \in L^2(\mathbb{R} \times [0, 2\pi))$ onto the Wigner-Ville transform $\mathrm{WV}_x \in L^2(\mathbb{R} \times [0, 2\pi))$ of the signal $x \in L^2(\mathbb{R})$ (recall (2)), each kernel employs independent parameters for different frequencies $f \in [0, 2\pi)$.

Definition 1 (K-transform) *The K-transform of a signal x with kernel Φ is defined as*

$$K_x(t, f) = \int_{\mathbb{R} \times [0,2\pi)} WV_x(\tau, \omega)\Phi_f(t - \tau, \omega) d\tau d\omega, \tag{4}$$

where the 2-dimensional Gaussian density function is parametrized by

$$\Phi_f(\tau, \omega) = \mathcal{N}\left(\begin{pmatrix}\tau \\ \omega\end{pmatrix}; \begin{pmatrix}0 \\ \mu_{\mathrm{f}}(f)\end{pmatrix}, \begin{pmatrix}\sigma_{\mathrm{t}}(f)^2 & \rho(f) \\ \rho(f) & \sigma_{\mathrm{f}}(f)^2\end{pmatrix}\right) \mathbb{1}_{\{\omega \in [0,2\pi)\}}, \tag{5}$$

where $\mathcal{N}$ is the Gaussian multivariate density function.

First, one should notice that the above extends the definition of the Cohen's class (recall (3)) and the affine Cohen class. The K-transform falls back to the Cohen class iff the kernels Φ_f are identical for any frequency position f (in that case $\mathrm{K}_{\boldsymbol{x}}$ can be obtained with a single 2-dimensional convolution between $\mathrm{WV}_{\boldsymbol{x}}$ and the single Gaussian kernel). The K-transform falls back to the affine Cohen class iff the mean and covariance parameters are tied together as in $\sigma_{\mathrm{t}}(f)^2 \propto 1/\mu_{\mathrm{f}}(f)$ and $\sigma_{\mathrm{f}}(f)^2 \propto \mu_{\mathrm{f}}(f)$. That is, the filters have bandwidths varying with their mean and center-frequencies (in the WV domain). One important family of TFRs known as variable Q-transform (Huang et al., 2015) can be reached by the K-transform but not by the affine Cohen class as opposed to constant Q-transforms TFRs (Brown, 1991) that can be reached by the K-transform and the affine Cohen class.

Second, the K-transform is real valued as both the Wigner-Ville representation $\mathrm{WV}_{\boldsymbol{x}}$ and the Gaussian kernel are real valued. We now discuss the advantages of such a parametrization. Each filter has 4 degrees of freedom that fully characterize the type of events being captured and allow for interpretability, we discuss this further below. As the parameters of different filters $\Phi_f, \Phi_{f'}, f \neq f'$ are independent, the number of degrees of freedom of the transform depends (linearly) on the desired number of filters (a finite number in practice). For common machine learning tasks and datasets we rarely see the number of filters exceed 128 which leads to 512 degrees of freedom. In this paper we will learn those parameters with some flavors of gradient descent as the transform is differentiable w.r.t. the Gaussian parameters. This will allow to adapt the filters and thus the transform to the data and task at hand. All the implementation details as well as a fast computation method will be provided in the next section.

Lastly, a key property of most TFRs reside in their time-frequency resolution, that is, how precise will be the representation into reflecting the frequency content at each time step present in the studied signal. Standard TFRs (and their learnable versions) have limited resolutions constrained by the uncertainty principle. In short, TFRs such as scalograms or STFTs all employ time-frequency filters that have their effective support's area lower bounded (when viewed in the WV domain). And while changing the hyper-parameters of those TFRs alter the supports' shape of those filters, the area constraint prevents them from being highly localized in both time and frequency. We present an illustration of this in Fig. 4. The WV has perfect time and frequency localization, and thus the K-transform can reach super-resolution by adapting the covariance matrices of the filters. In the limit when the Gaussian filters become Dirac functions ($\rho(f) = 0, \sigma_{\mathrm{t}}(f) = \sigma_{\mathrm{f}}(f) = \epsilon$), the K-transform falls back to the WV transform. We will study the sensitivity of the learned representation as a function of the Gaussian covariance matrices in the next section (Prop. 2), where we show that reaching super-resolution will make the representation highly sensitive to input noise and deformations; the ability to learn and adapt to the task and data at hand is thus crucial to ensure that super-resolution is reached only if beneficial to increase performances.

3.2. Interpretability

The Gaussian parametrization of the kernel Φ (recall (5)) plays many crucial roles; the first being the ease of interpretability of the filters and of the produced representation. The frequency mean $\mu_{\mathrm{f}}(f)$ represents the frequency center of the events captured by this filter. Studying the mapping μ will also inform on the concentration of those center-frequencies and if for example they follow a logarithmic growth as in the WT or a linear growth as in the STFT. The diagonal of the covariance matrix $(\sigma_{\mathrm{t}}(f)^2, \sigma_{\mathrm{f}}(f)^2,)$ encodes the frequency and time bandwidths of the filter. That is, how precise

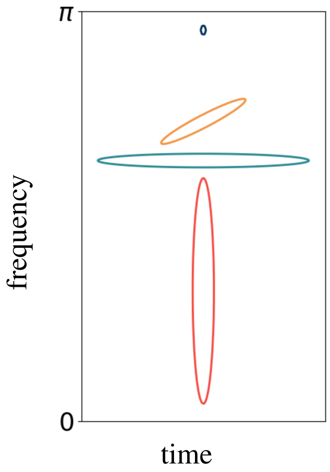

Figure 1: Visualization of four different Gaussian filters Φ in the time-frequency plane (the Wigner-Ville domain) each depicted with a different color. Interpretability of the filters (as detailed in Sec. 3.2) can be done directly by looking at the mean (position) of the Gaussian and the covariance matrix. In this case, the **black/top** filter is highly localized in time and frequency near the Nyquist frequency (notice the much higher localization than allowed in a WT or STFT; the **orange/anisotropic** filter is encoding chirp events (linear change in time of the instantaneous frequency), this is typical of bats or birds'calls; the **green/flat** filter is highly localized in frequency and encodes long duration event, this can be seen with insects emitting relatively high frequency sounds for a long duration; the **red/bottom** filter on the contrary is localized in time and encodes large band frequency events, this is typical of transient events such as clicks of dolphins.

in time and in frequency is the filter. For example when capturing a stationary signal with long time duration an appropriate $\sigma_{\rm t}(f)$ to best capture the event would be greater than when capturing a transient (short duration, burst like) event. Hence, studying the learned parameters (and filters) allow to study the physical properties of the encoded events which is crucial for example in bioacoustics (Dawson, 1991) or geophysics (Seydoux et al., 2020). The off-diagonal of the covariance, $\rho(f)$, encodes the (linear) chirpness of the filter; that is, how the instantaneous frequency of the filter changes with time. This chirpness parameter plays a crucial when capturing events as generated by bats, birds (Barclay, 1999; Capus and Brown, 2003) or radar and imaging devices (Baraniuk and Steeghs, 2007; Luo et al., 2009). Beyond this direct interpretability of the types of event being encoded by each filter, this parametrization is also general enough to reach most of the current TFRs as we now demonstrate.

3.3. Universality

A second key benefit of the proposed Gaussian parametrization resides in its ability to reach most of the employed TFRs. In fact, it has been shown through various studies how TFRs such as the chirplet transform, the WT with Morlet wavelet or the Gabor transform (Nuttall, 1988; Flandrin and Rioul, 1990; Jeong and Williams, 1990; Baraniuk and Jones, 1996; Talakoub et al., 2010; Gillespie and Atlas, 2001) all belong to the (affine) Cohen class where the convolution kernels are 2-dimensional Gaussian where only the mean and covariance parameters are to be changed. We formalize this result below.

Proposition 1 (Universality) *Any Gabor transform, Gabor wavelet transform, Gabor chirplet transform are reachable by the K-transform. (Proof in App. C.2.)*

The above result is crucial as it demonstrates how the proposed formulation covers most of the current state-of-the-art deterministic TFRs. In addition, most learnable frameworks simply adapt the parameters of some fixed TFRs, as such, those learned representations are also reachable by the K-transform. We depict in Fig. 2 some specific cases of means and covariances that can be imposed on the parameters of the Gaussians Φ_f to make the K-transform fall back to known TFRs; we also give the analytically parameters for each case in Table 3 in the Appendix.

We should highlight that not all representations can be reached by the K-transform. For example, a sinc based filter-bank can not be reached as it would require not a Gaussian filter applied onto $\text{WV}_{\boldsymbol{x}}$

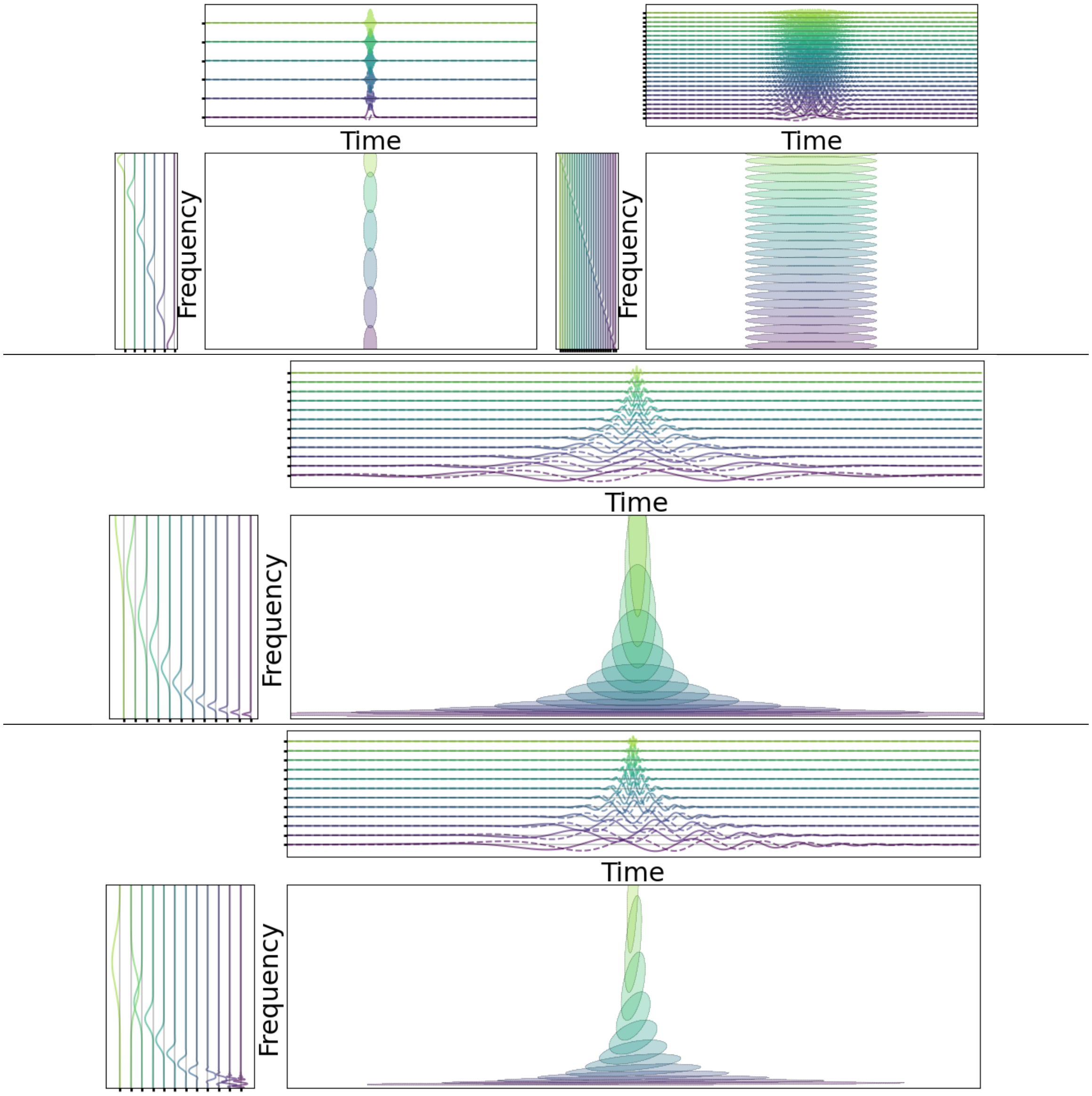

Figure 2: Depiction of various filter-banks in the frequency domain, time domain and time-frequency plane (Wigner-Ville domain) for a STFT (**top**) with small Gaussian apodization window (left) and large Gaussian apodization window (right); for a wavelet transform (**middle**) with Morlet wavelets, and for a chirplet transform (**bottom**). First, all those TFRs can be obtained through a Gaussian filtering of the Wigner-Ville representation each with a specific set of means and covariances for the Gaussians. By changing those parameters it is possible to move continuously from one TFR to another (Theorem 1). Second, the larger is the Gaussian (covariance) in one direction or the other (time or frequency) the more stable will be the produced TFR (Prop. 2), with an highly localized Gaussian (small covariance entries) super-resolution is achieved providing the produced TFR increased time and/or frequency precision of the events present in the input signal. Third, learnability of the Gaussian parameters allow to let the data, classification/regression pipeline and the loss function at hand drive the design of the TFR to maximize performances.

but a rectangular window. Also, any TFR employing a time-varying basis can not be recovered by the K-transform. This type of "signal adapted" basis (Ramchandran and Vetterli, 1993; Abramovich et al., 1998) is not common in regression and classification tasks as those tasks heavily rely on translation invariance of the representation. A simple rule determining if a TFR can be reached by the K-transform is formalized below as a direct application of Moyal theorem (see Theorem 8 in Bahri and Ashino (2015)).

Remark 1 *Any signal representation R that can be expressed as $R(.,f) = |\boldsymbol{x} \star \psi_f|^2$ where $\psi_1, \psi_2, \dots$ are possibly independent filters from each other can be reached by the K-transform if the WV transform of the signals $WV_{\psi_f}, \forall f$ are 2-dimensional Gaussian.*

3.4. Continuity and Stability

In this section we highlight some fundamental properties of the K-transform concerning its stability to changes in the Gaussian parameters of the filters, and changes in the input signal $\boldsymbol{x}$. In particular our first result demonstrate how the proposed formulation allows to continuously move from one TFR to another (recall Prop. 1) when changing the Gaussian parameters.

Continuity. For convenience, we encapsulate all the Gaussian parameters as a single operator $\theta : [0, 2\pi) \mapsto \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}$ defined as

$$\theta(f) \triangleq \left(\mu_{\mathrm{f}}(f), \sigma_{\mathrm{t}}(f)^2, \sigma_{\mathrm{f}}(f)^2, \rho(f)\right)^T, \tag{6}$$

where to simplify notation we assume $\rho(f) \in \mathbb{R}$, we discuss the valid support that ensures invertibility of the covariance matrix in the next section; we also explicit the K-transform as $\mathrm{K}_{\boldsymbol{x},\theta}$ to make the dependency on θ explicit. We can now derive the Lipschitz continuity result of the proposed representation in term of its parameters.

Lemma 2 (Lipschitz continuity) *Given two parameter mappings θ and θ' in $L^2(\mathbb{R})$ as defined by (6), the distance between the two induced representations is upper bounded by the distance between those mappings as in*

$$\|K_{x,\theta} - K_{x,\theta'}\|^2_{L^2(\mathbb{R}^2)} \leq \kappa \|x\|^2_{L^2(\mathbb{R})} \left(\int_0^{2\pi} \|\theta(f) - \theta'(f)\|_2^2 df\right)^{\frac{1}{2}},$$

with κ the Lipschitz constant of a standard 2D Gaussian (≈ 0.2422). (Proof in App. C.1.)

From the above result we see how one can compare different K-transforms simply by comparing their associated parameters, as close parameters imply close representations of the same inputs. As such, if the learned parameters are say close to the ones of a WT, then one can confidently assume the learned representation to be a WT. Lipschitz continuity also naturally provides the following.

Theorem 1 (TFR Interpolation) *The representation $K_{\boldsymbol{x},\theta}$ moves continuously with θ allowing to continuously interpolate between any reachable TFR. (Proof in App. C.3.)*

The above result demonstrates that the transform is Lipschitz continuous with respect to its parameters θ, hence, moving from any representation to another is done in a continuous fashion when continuously moving the parameters. This property is also important in our context of learning θ. In fact, when performing updates of those parameters in a gradient based fashion (small increments) we are guaranteed to not produce abrupt changes in the representations. This is highly beneficial to have stability during training. We conclude with an important result derived when the affine Cohen class was generalized to produce wavelet transforms in Flandrin and Rioul (1990) (Prop. 4) stating that the Gaussian parametrization we employ is not only sufficient but also necessary.

Remark 3 *A continuous passage from a spectrogram to a scalogram by filtering the WV$_{\boldsymbol{x}}$ is possible iff the filtering is done with Gaussians filters.*

The above result should demonstrate along with Sec. 3.2 and 3.3 that the Gaussian parametrization is not only providing interpretability and universality but is also the only parametrization allowing a continuous interpolation between TFRs.

Stability. We now study how does the representation change when the input $\boldsymbol{x}$ is perturbed (as opposed to the previous paragraph that focused on perturbations of the parameters θ). The first direct result concerns the equivariance of the representation with translation of the input in time. That is, translating the input in time translates the representation. This results comes from the fact that the WV is itself translation equivariant, and since we apply a time convolution on this representation, itself being translation equivariant, we have that the K-transform maintains this property. Maintaining this property is important especially for tasks without a priori knowledge in order to maintain enough information about the input signal (Kondor and Trivedi, 2018; Cohen and Welling, 2016; Cohen et al., 2018).

On the other hand, we also want to ensure that small changes in the input do not yield high changes in the representation. This stability to input deformation is also crucial to provide easier stochastic optimization and improved generalization (Mallat, 2016). This type of stability is often achieved from handcrafted designs such as a time averaging (Bruna and Mallat, 2013). We now demonstrate that in the K-transform, the sensitivity to input transformations is directly controlled by the covariance of the Gaussian filters, and that during training, the filters will adapt to produce stable representations only if needed based on the data and task specific perturbations. Denote a transformation u applied on the signal x and formally characterize the induced perturbation amount in the K-transform by $\|\mathrm{K}_{x,\Phi} - \mathrm{K}_{u(x),\Phi}\|_{L^2(\mathbb{R}^2)}$, let also denote $\det(\Phi_f) \triangleq \sigma_{\mathrm{t}}(f)^2\sigma_{\mathrm{f}}(f)^2 - \rho(f)^2$.

Proposition 2 (Stability to deformation) *A transformation of the signal $D(x)$ implies a change in the representation proportional to the inverse of the covariance determinant as*

$$\|K_{x,\Phi} - K_{D(x),\Phi}\|_{L^2(\mathbb{R}^2)} \leq \frac{\kappa \, \|x - D(x)\|_{L^2(\mathbb{R})}}{\min_f \det(\Phi_f)},$$

with κ the Lipschitz constant of the WV which exists and is finite for bounded domain. (Proof in App. C.6.)

Based on the data and task at hand, the K-transform can adapt its parameters and in particular its covariance matrix to reach super-resolution or to reach stability to input deformations. The extreme cases range from Dirac-like filters producing a highly sensitive TFR to constant filters (infinite limit variance) producing a completely invariant representation. This highlighted trade-off demonstrates how current TFRs (e.g. WT, SP) which have $\det(\Phi_f)$ lower bounded by the uncertainty principle provide some stability in their representation. Furthermore, specific invariants can be obtained as highlighted below allowing to recover for example time invariance à la scattering network (Mallat, 2012).

Remark 4 *Time invariance, $\|K_{x,\Phi} - K_{D(x),\Phi}\|_{L^2(\mathbb{R}^2)} = 0$ with D any translation operator, is reached as $\sigma_{\mathrm{t}}(f) \to \infty, \forall f$, frequency invariance, $\|K_{x,\Phi} - K_{D(x),\Phi}\|_{L^2(\mathbb{R}^2)} = 0$ with D any frequency shift operator, is reached as $\sigma_{\mathrm{f}}(f) \to \infty, \forall f$.*

Noise Sensitivity. One last important study concerns the sensitivity of the proposed transform in the presence of noise in the signal. That is, we concentrate the above analysis to the case where D would produce an observation x formed by an underlying signal y corrupted with additive random noise ϵ. In the case of i.i.d. noise variables with finite first two moments we obtain the following result.

Lemma 5 *The K-transform of a noisy signal $x = y + \epsilon$, with i.i.d. noise (in time) ϵ, is a random variable with mean given by*

$$\mathbb{E}\left[K_x(t,f)\right] = K_y(t,f) + \int_0^{2\pi} S(\omega)\phi_f\left(\omega; \mu_{\mathrm{f}}(f), \sigma_{\mathrm{f}}^2(f)\right) d\omega, \tag{7}$$

with $S(\omega)$ the noise power density function and ϕ_f is the density function of a 1-dimensional Gaussian with given mean and variance. (Proof in App. C.5.)

From the above, we obtain that the K-transform of noisy signals will be biased based on the noise power density function and its correlation with the employed Gaussian kernels. The variance of the noisy transform $\mathrm{K}_x(t,f)$ can also be obtained analytically as we demonstrate in Lemma 5's proof. While we will focus on classification problems in this study, the above result shall open the door to applying the K-transform and learning the Gaussian kernels Φ_f for denoising applications. We now demonstrate how the K-transform can be computed efficiently solely from Fast Fourier Transforms, for other properties of the transform such as characterization of the interference based on the covariance please see Appendix D.

4. Fast Fourier Transform Computation of the K-transform

We carefully develop in this section the implementation of our method as well as present an interesting computation method allowing to scale the K-transform to large scale dataset and signal with long time duration (high number of recorded bins). We then employ our method on various large scale benchmarks and compare with alternative learnable TFRs and demonstrate how the K-transform attains higher accuracies.

4.1. Gaussian Parameter Learning.

Recall from Def. 1 that the K-transforms involves a Gaussian filtering of the $\mathrm{WV}_{\boldsymbol{x}}$ representation. We propose to learn the Gaussian parameters for each filter of the finite filter-bank

$$\{\Phi_1, \dots, \Phi_F\},$$

where the values of $\theta(1), \dots, \theta(F)$ from (6) are considered as the learnable parameters of the model. In order to ensure stability during training and avoid learning parameters producing non positive semidefinite (PSD) covariance matrices we impose

$$\sigma_{\mathrm{t}}(i) > 0, \sigma_{\mathrm{f}}(i) > 0, \rho(i) \in \left(-0.95\sqrt{\sigma_{\mathrm{t}}(i)\sigma_{\mathrm{f}}(i)}, 0.95\sqrt{\sigma_{\mathrm{t}}(i)\sigma_{\mathrm{f}}(i)}\right), i = 1, \dots, F,$$

ensuring that we can perform unconstrained gradient descent optimization of our objective function without producing covariance matrices that are not PSD. We do not impose constraints on $\mu_{\mathrm{f}}(i)$, from the experiments we observed that no divergence of this parameter occurred; however one could enforce that $\mu_{\mathrm{f}}(i) \in [0, \pi)$.

4.2. The Short-Time-Fourier-Transform Trick

The definition of the K-transform leverages the $\mathrm{WV}_{\boldsymbol{x}}$ transform of a signal $\boldsymbol{x}$. For a discrete signal of length N, this requires the computation of a $N \times N$ matrix obtained by doing N Fourier transforms of length N. Its computational complexity is thus quadratic ($\mathcal{O}(N^2 \log(N))$) making it unsuited for large scale tasks when N can easily reach $100,000$ or more. Once again, we will leverage the Gaussian parametrization to greatly speed-up the K-transform computation.

Isotropic covariance case. In order to provide a fast implementation, we will first consider the case of Φ_f employing an isotropic covariance matrix, that is, $\rho(f) = 0$ and $\sigma_{\mathrm{t}}(f) = \sigma_{\mathrm{f}}(f), \forall f$. In that specific case, the K-transform can be obtained by (i) computing $\mathrm{STFT}_{x,\sigma_{\mathrm{t}}(f)}$, (ii) doing a spectral autocorrelation of $\mathrm{STFT}_{x,\sigma}$ and (iii) doing a 2D convolution with a Gaussian with diagonal covariance. We formalize this result below.

Lemma 6 *Any K-transform (recall (4)) with Φ_f having isotropic covariance matrix $\sigma^2 I_2, \forall f$ can be obtained via*

$$
\begin{aligned}
K_{\boldsymbol{x}}(t,f) &= \left(WV_{\boldsymbol{x}} \star \mathcal{N}\left(.; \begin{pmatrix} 0 \\ \mu_{\mathrm{f}}(f) \end{pmatrix}, \sigma^2 I_2 \right) \right)(t,f) = \left(WV_{\boldsymbol{x}} \star \mathcal{N}\left(.; \mathbf{0}, \sigma^2 I_2\right)\right)(t, \mu_{\mathrm{f}}(f)) \\
&= \int_{-\infty}^{\infty} g_{\sigma^{-1}-\sigma}(\omega) e^{j2\pi\omega t} STFT_{\boldsymbol{x},\sigma^{-1}}\left(t, \mu_{\mathrm{f}}(f) + \frac{\omega}{2}\right) STFT^{*}_{\boldsymbol{x},\sigma^{-1}}\left(t, \mu_{\mathrm{f}}(f) - \frac{\omega}{2}\right) d\omega,
\end{aligned}
$$

where g_σ is a 1-dimensional Gaussian function with spread σ. (Proof in App. C.7.)

It is clear from the above result that σ greatly impacts the speed of computation of $\mathrm{K}_{\boldsymbol{x}}$ as (i) it will limit the amount of cross-correlation to perform (not computation needed outside of the effective support of g which decreases as σ grows), and (ii) the STFT computation will be perform on smaller windows as σ increases. For details on the Gaussian window truncation please see Appendix E. As a result, the above provide a convenient solution to produce a transform with diagonal covariance matrix in the filters Φ_f and the computational cost increases as σ reduces. In the discrete case, the above turns the complexity from $\mathcal{O}(N^2 \log(N))$ to $\mathcal{O}(NM \log(M))$ with M the window size (effective support in bins of a Gabor apodization with standard deviation σ^{-1}; where $M \to N$ as $\sigma \to 0$. We now move to the general case of arbitrary covariance matrices.

General case. In the general case we do not impose any constraint on the filter Φ_f covariance matrices. Nevertheless, it should be clear that any 2-dimensional Gaussian with arbitrary covariance can be rewritten as the convolution between a Gaussian width diagonal covariance matrix and a Gaussian with arbitrary covariance matrix.

Theorem 2 *Any K-transform can be obtained from convolving the WV with a diagonal covariance Gaussian and a full covariance Gaussian as*

$$
K_x(t,f) = \left(\underbrace{WV_x \star \mathcal{N}\left(.; \mathbf{0}, \sigma^2 I_2\right)}_{\text{does not depend on } f} \star \mathcal{N}\left(.; \mathbf{0}, \Sigma(f)\right) \right)(t, \mu_{\mathrm{f}}(f)),
$$

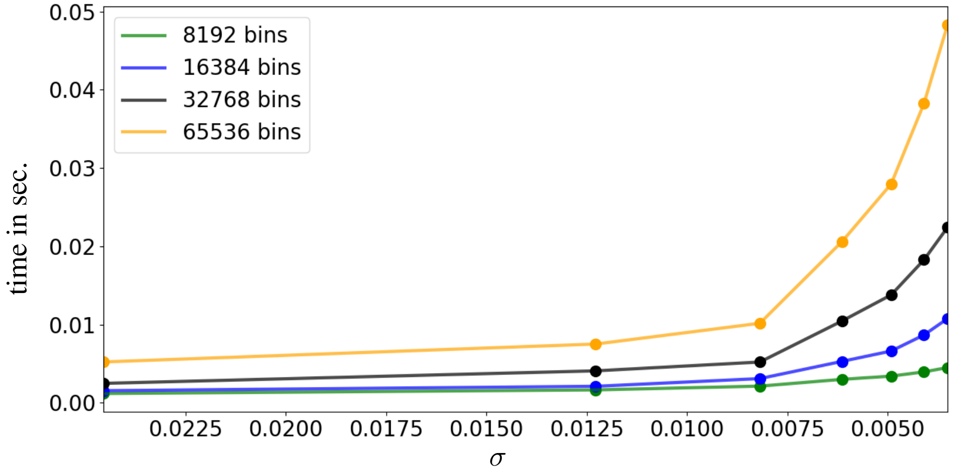

Figure 3: Computation time (in sec.) to evaluate $\mathrm{WV}_x \star \mathcal{N}\left(.;\mathbf{0},\sigma^2 I_2\right)$ on signals x of different lengths depicted with different colors and with varying σ. The smaller σ the more localized in time and frequency is the representation, in the limit case of $\sigma \rightarrow 0$ the produced representation falls back to the WV transform. An increase in the length of the signal x incurs a linear increase in computation time. The outer left case has similar computation time than the corresponding STFT with same σ for its frequency bandwidth.

with $\Sigma(f)$ a positive semidefinite matrix and σ^2 the largest value such that

$$\Sigma(f)+\sigma^2 I_2 = \begin{pmatrix} \sigma_{\mathrm{t}}^2(f) & \rho(f) \\ \rho(f) & \sigma_{\mathrm{f}}(f) \end{pmatrix}.$$

(Proof in Appendix C.8.)

By employing the above and Lemma 6 we see how one can compute a priori the representation $\mathrm{WV}_x \star \mathcal{N}\left(.;\mathbf{0},\sigma^2 I_2\right)$ that is then used to produce the desired K-transform. We propose in Fig. 3 computation times of $\mathrm{WV}_x \star \mathcal{N}\left(.;\mathbf{0},\sigma^2 I_2\right)$ for different signal lengths and different values of σ. Computing the K-transform for various kernels can be done from a base representation which is not the WV but the already convolved WV. If one a priori imposes a minimal value for the covariance matrix then this method can greatly speed up computation as opposed to employing the analytical $\mathrm{WV}_{\boldsymbol{x}}$ from (2). Furthermore, the employed STFT is efficiently implemented in most (CPU/GPU) software allowing an efficient computation of the transform.

Remark 7 *The K-transform time and/or frequency resolution is inversely proportional to its speed of computation.*

Pseudo-code. We provide below the explicit pseudo code that summarizes all the involved steps and their impact of the final representation obtained (see App. G for the computational complexity):
1. Do a Gabor transform of $\boldsymbol{x}$ with a Gaussian window g_σ for apodization leading to $\mathrm{STFT}_{\boldsymbol{x},\sigma}$;this parameter will determine the final frequency resolution of the transform ($\propto \sigma$) (Prop. 7, Fig. 2).
2. Do the spectral auto-correlation of $\mathrm{STFT}_{\boldsymbol{x},\sigma}$ with a gaussian spectral apodization window $g_{1/\sigma-\sigma}$ to obtain $\left(\mathrm{WV}_{\boldsymbol{x}} \star \mathcal{N}\left(.;\mathbf{0},\sigma^2 I_2\right)\right)$ from Lemma 6. This maintains the frequency resolution of $\mathrm{STFT}_{\boldsymbol{x},\sigma}$ while increasing the original time resolution $\frac{1}{\sigma}$ by σ.
3. Compute the K-transform as per Theorem 2.

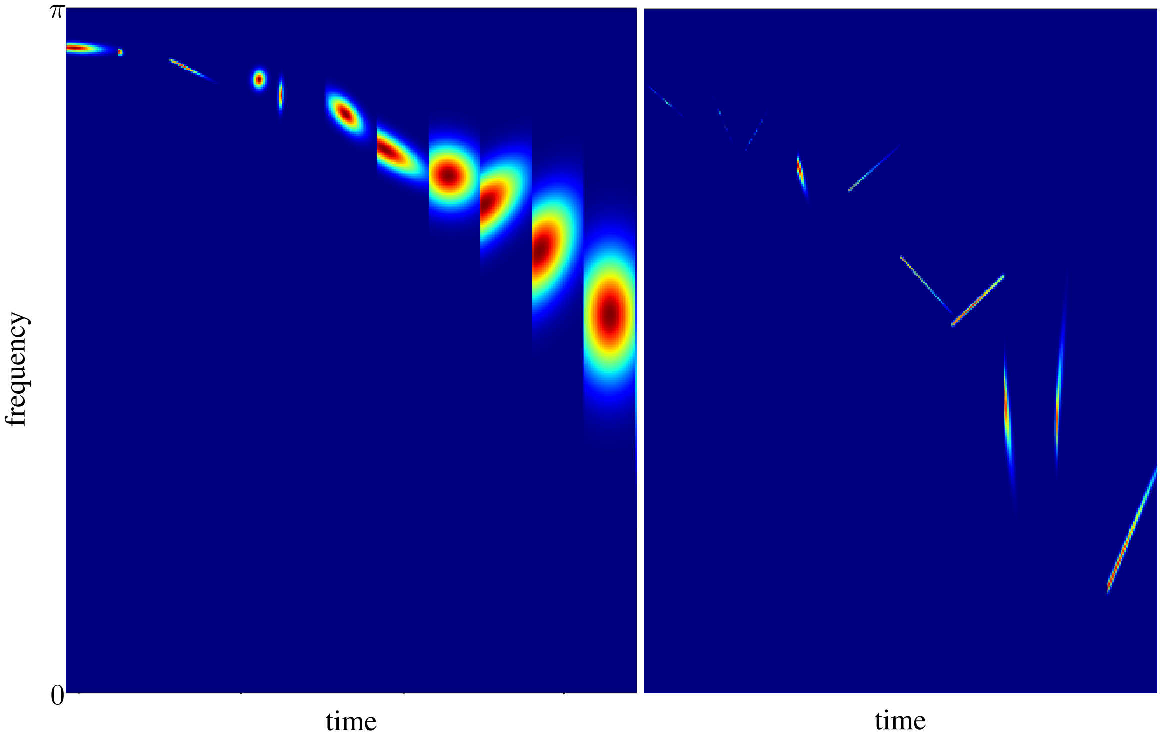

Figure 4: Depiction of the learned filters $\Psi[., f]$ (Def. 1). The full filters banks can be found in Appendix B, the vertical over horizontal axes ratio is 1. **Left, our learned representation on AudioMNIST dataset (Becker et al., 2018)**: This noiseless speech pushes the filters with high frequency resolution (easily seen from the filter covariance as per (6)). For some of these filters the time resolution is also very high (reaching super-resolution), while others favor local translation invariance. For all the medium to low frequency filters, great frequency and time invariance is preferred with large gaussian support, with a slight chirpness for the medium frequency filters. Thus, the low frequency filters tend to favor time resolution. **Right, our learned representation on Birdvox dataset (Lostanlen et al., 2018)**: We see that the detection of bird songs heavily relies on chirps. Indeed the characteristic sound of birds is increasing or decreasing in frequency over time. Moreover, the learned representation demonstrate filters that reach super-resolution. Thus it encodes with high precision the time and/or frequency position of the events per Prop. 2) as the cost of being more sensitive to input deformation. It fits with the extreme time-frequency accuracy of bird audition and acoustic production (Dooling and Lohr, 2006).

4.3. Experimental Validation

We propose to validate our method on various classification tasks. We briefly describe below the used dataset and DN architectures and provide the accuracy results averaged over 10 runs and over multiple learning rates in Table 1. All dataset are described in details in Appendix H. For each dataset we experiment composing our learnable TFR with three DN architectures (for detailed description of those architectures and additional hyper parameters choices please see Appendix A). First, we consider a simple model that corresponds to a **Linear Scattering** where we have TFR-global time averaging-linear classifier. This model will fully rely on the TFR as no additional (non)linear transformation processes the inputs prior reaching the linear classifier. Second, we employ a slightly more complex model involving a nonlinear classifier instead. That is, we use the pipeline TFR-global time averaging-MLP; we denote it as **Nonlinear Scattering**. Lastly, we extended the Nonlinear Scattering case by employing a convolutional layer prior the time averaging

Table 1: Average over 10 runs of classification result using three architectures without data augmentation, one layer scattering followed by a linear classifier (Linear Scattering), one layer scattering followed by a two layer neural network (Nonlinear Scattering) and a two layer scattering with joint (2D) convolution for the second layer followed by a linear classifier (Linear Joint Scattering). For each architecture and dataset, we experiment with the baseline, Morlet wavelet fitler-bank (morlet), and learnable frameworks being ours (lwvd), learnable sinc based filters (sinc) and learnable Morlet wavelet (lmorlet). As can be seen, across the dataset, architectures and learning rates, the proposed method provides significant performance gains (stds in Tab. 2). *The increased gap in performances for the linear case and the ability of the K-transform to reach high accuracies in that linear case confirm the ability to produce highly adapted representations across datasets.* For the reader's curiosity we provide some banchmarks on those datasets, however direct comparison is not possible at different data splitting, data augmentation (in our case no data augmentation is employed) and the likes are different for each study. For Audio MNIST using AlexNet on spectrogram leads to 95% (Becker et al., 2018) accuracy; for biosonar classification in DOCC10 challenge raw audio CNN lead to 71.13% test accuracy (Ferrari et al., 2020); for BirdVox, using carefully designed spectrograms with PCA and kernel SVM lead to 87.77% accuracy, and employing instead 3 layers of 2D convolution and 2 fully connected layers lead to 90.48% (and 94.85% with data augmentation) test accuracy (Lostanlen et al., 2018); for the Google command dataset, a DenseNet 121 without pretraining nor data augmentation reaches 80% test accuracy (de Andrade et al., 2018) using a spectrogram.

		Linear Scattering				Nonlinear Scattering				Linear Joint Scattering			
	le. rate	*morlet*	**lwvd**	sinc	lmorlet	*morlet*	**lwvd**	sinc	lmorlet	*morlet*	**lwvd**	sinc	lmorlet
DOCC10	0.0002	14.3	63	31.1	29.7	54.1	84.7	74.4	74.9	70.7	**83.7**	82.4	75.8
	0.001	12.7	65.5	26.0	28.3	50.1	**87.9**	77.4	77.4	70.1	80.6	80.8	73.2
	0.005	13.0	**65.9**	17.1	27.0	51.8	87.1	43.3	83.2	65.9	78.0	70.5	80.8
BirdVox	0.0002	63.8	77.9	69.6	65.4	84.7	92.9	88.1	85.8	82.1	**90.5**	87.2	84.3
	0.001	65.0	80.0	67.2	64.3	85.0	**94.2**	88.1	86.6	80.3	88.7	86.8	83.1
	0.005	65.2	**80.4**	67.3	66.9	84.8	94.2	86.0	87.2	78.1	87.5	78.3	82.8
MNIST	0.0002	43.9	68.4	52.2	44.0	82.3	85.3	10.4	83.0	95.3	97.6	22.1	95.4
	0.001	41.5	**68.8**	43.5	42.2	83.2	**89.8**	87.1	85.4	89.7	**97.8**	93.2	90.4
	0.005	34.6	68.8	23.9	36.0	82.7	22.1	68.7	88.1	81.1	12	64.4	80.2
command	0.0002	8.1	24.9	9.5	7.6	33.9	38.2	36.2	33.4	65.8	**76.7**	3.7	66.8
	0.001	7.5	**26.1**	8.0	8.2	33.5	**42.9**	35.5	33.7	53.6	71.8	27.9	51.9
	0.005	7.3	25.7	6.2	6.5	33.0	17.0	28.9	34.8	32.1	35.2	17.2	32.9
fsd	0.0002	9.7	15.3	10.3	9.0	22.9	23.1	2.3	27.9	40.1	38.8	1.6	42.0
	0.001	9.8	16.7	10.4	10.6	24.2	27.4	13.1	**31.1**	38.9	**44.9**	2.1	42.3
	0.005	9.0	**17.4**	5.5	10.2	24.2	28.8	16.9	30.4	25.0	31.5	17.0	33.2

and then using a linear classifier as in TFR-Conv 2D-global time averaging-linear classifier, and is denoted as **Linear Joint Scattering**). While we provide those different settings to represent the plurality of situations where learnable TFRs can be used in practice, the Linear Scattering model is the one relying the most in the TFR and is thus the case where the most significant trends should occur.

In addition to comparing different classification networks, we also compare for each of those networks and each dataset our learnable TFR (the K-transform) that we abbreviate as **lwvd**, to the learned Morlet filter-bank (Khan and Yener, 2018) denoted as **lmorlet**, and to the learnable sinc based filter-bank (Ravanelli and Bengio, 2018) denoted as **lsinc**, which are the current state-of-the-art techniques proposing a learnable TFR. In order to calibrate all the results we also compare with a fixed Morlet filter-bank which is the one that is often seen as the most adapted when dealing with speech and bird signals. We do not employ any data-augmentation technique. In all cases the K-transform outperforms other learnable TFRs and the a priori optimal one across dataset and optimization settings, offering significant performance gains. We propose in Fig. 4 and Appendix B all the figures of the learned filters and their interpretation. We provide the results in Tab. 1 and Tab. 2 averaged over 10 runs, for each of the runs, the same data split, DN initialization and parameters are used across all TFRs to allow exact performance comparisons. We also provide the results across various learning rate to perceptually measure the sensitivity of each method to this parameter. The first key observation is that the learnable methods are much more sensitive to the learning rate than when using a fixed TFR. Nevertheless, the proposed K-transform is able to outperform all methods across the datasets and for any DN. This comes from the extreme adaptivity of the produced TFR. Notice that the fixed morlet TFR reaches reasonable accuracy especially on speech data without noise (audio MNIST). This is another key feature of learnable TFR, the ability to learn more robust representations. Another key observation comes from the ability of the proposed method to reach state-of-the-art performances while leveraging a simple (few layer) DN in the Linear Joint Scattering case.

5. Conclusions

We proposed a novel approach to learn generic WVD based TFR, derived an efficient implementation and demonstrated its ability to outperform standard and other learnable TFR techniques across dataset and architecture settings. In addition of learning any desired TFR, our framework is interpretable allowing one to easily understand the physical properties being captured by the learned TFR as well as positioning the learned TFR in the realm of the conventional TFRs such as spectrograms and scalograms. Our study opens many interesting research directions. First, it is possible to perform statistical analysis of noisy signals in order to better design and constrain the Gaussian filters of the K-transform to minimize the impact of noise into the final representation. This is a powerful scope that would allow interpretable and theoretically grounded regularization techniques to be obtained for the K-transform filters as has been done in the case of the WV transform in Amirmazlaghani and Amindavar (2009, 2013); Levy et al. (2020). Another extension of the proposed method is model-based signal approximation, such as using a tree model Baraniuk (1999). Yet another extension consists of using an additional learnable parameter for the time dimension of the Gaussian filters mean parameters. Doing so would allow further flexibility in the learned representation at the cost of reduced interpretability as each frequency dimension will have its own time alignment. Such a compromise of representation power versus interpretability should offer more options for practitioners. Lastly, it is possible to leverage advances in computational methods for performing convolutions with Gaussian filters (Getreuer, 2013) in order to speed up the proposed method or enable CPU and/or memory efficient computations.

Acknowledgments: Randall Balestriero and Richard Baraniuk are supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571, N00014-20-1-2787, and N00014-20-1-2534; AFOSR grant FA9550-18-1-0478; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. Hervé Glotin is supported by AI Chair ADSIL ANR-20-CHIA-0014, SMILES ANR-18-CE40-0014.

References

Felix Abramovich, Theofanis Sapatinas, and Bernard W Silverman. Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749, 1998.

Jonathan Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977.

Maryam Amirmazlaghani and Hamidreza Amindavar. Modeling and denoising wigner-ville distribution. In *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, pages 530–534. IEEE, 2009.

Maryam Amirmazlaghani and Hamidreza Amindavar. Statistical modeling and denoising wigner–ville distribution. *Digital Signal Processing*, 23(2):506–513, 2013.

Mawardi Bahri and Ryuichi Ashino. Convolution and correlation theorems for wigner-ville distribution associated with linear canonical transform. In *2015 12th International Conference on Information Technology-New Generations*, pages 341–346. IEEE, 2015.

Randall Balestriero and Hervé Glotin. Wavelet learning by adaptive hermite cubic splines applied to bioacoustic chirps. In *OCEANS 2019 - Marseille*, pages 1–5, 2019. doi: 10.1109/OCEANSE.2019.8867410.

Randall Balestriero, Romain Cosentino, Herve Glotin, and Richard Baraniuk. Spline filters for end-to-end deep learning. In *Proc. 35th Int. Conf. on Machine Learning*, volume 80, pages 364–373, 10–15 Jul 2018.

Richard Baraniuk. Optimal tree approximation with wavelets. In *Wavelet Applications in Signal and Image Processing VII*, volume 3813, pages 196–208. International Society for Optics and Photonics, 1999.

Richard Baraniuk and Douglas Jones. Wigner-based formulation of the chirplet transform. *IEEE Transactions on signal processing*, 44(12):3129–3135, 1996.

Richard Baraniuk and Philippe Steeghs. Compressive radar imaging. In *2007 IEEE radar conference*, pages 128–133. IEEE, 2007.

Robert Barclay. Bats are not birds—a cautionary note on using echolocation calls to identify bats: a comment. *Journal of Mammalogy*, 80(1):290–296, 1999.

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

Ronald Newbold Bracewell and Ronald Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

Judith Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Emre Çakir and Tuomas Virtanen. End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input. In *JCNN*, pages 1–7. IEEE, 2018.

Chris Capus and Keith Brown. Short-time fractional fourier methods for the time-frequency representation of chirp signals. *The Journal of the Acoustical Society of America*, 113(6):3253–3263, 2003.

Leon Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941–981, 1989.

Leon Cohen. *Time-frequency analysis*, volume 778. Prentice hall, 1995.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. *CoRR*, abs/1801.10130, 2018.

Ronald Coifman and Victor Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on information theory*, 38(2):713–718, 1992.

Ronald Coifman, Yves Meyer, Steven Quake, and Victor Wickerhauser. Signal processing and compression with wavelet packets. In *Wavelets and their applications*, pages 363–379. Springer, 1994.

Elena Cordero and Fabio Nicola. Sharp integral bounds for wigner distributions. *International Mathematics Research Notices*, 2018(6):1779–1807, 2018.

Romain Cosentino and Behnaam Aazhang. Learnable group transform for time-series. In *International Conference on Machine Learning*, pages 2164–2173. PMLR, 2020.

Romain Cosentino, Randall Balestriero, and Behnaam Aazhang. Best basis selection using sparsity driven multi-family wavelet transform. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 252–256. IEEE, 2016.

Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Ingrid Daubechies and Fabrice Planchon. Adaptive gabor transforms. *Applied and Computational Harmonic Analysis*, 13(1):1–21, 2002.

Stephen Dawson. Clicks and communication: the behavioural and social contexts of hector's dolphin vocalizations. *Ethology*, 88(4):265–276, 1991.

Douglas Coimbra de Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. A neural attention model for speech command recognition. *arXiv preprint arXiv:1808.08929*, 2018.

Minh Do and Martin Vetterli. Orthonormal finite ridgelet transform for image compression. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 2, pages 367–370. IEEE, 2000.

David Donoho. On minimum entropy segmentation. In *Wavelet Analysis and Its Applications*, volume 5, pages 233–269. Elsevier, 1994.

Robert Dooling and Bernard Lohr. Auditory temporal resolution in the Zebra Finch (Taeniopygia guttata): A model of enhanced temporal acuity. *Ornithological Science*, 5(1):15 – 22, 2006. doi: 10.2326/osj.5.15.

Maxence Ferrari, Hervé Glotin, Ricard Marxer, and Mark Asch. Docc10: Open access dataset of marine mammal transient studies and end-to-end cnn classification. In *Int. Joint Conf. on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Patrick Flandrin. Temps–fréquence (traité des nouvelles technologies, série traitement du signal). 1993.

Patrick Flandrin. *Time-frequency/time-scale analysis*. Academic press, 1998.

Patrick Flandrin and Bernard Escudié. An interpretation of the pseudo-wigner-ville distribution. *Signal Processing*, 6(1):27–36, 1984.

Patrick Flandrin and Oliver Rioul. Affine smoothing of the wigner-ville distribution. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2455–2458. IEEE, 1990.

Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902*, 2018.

Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

Pascal Getreuer. A survey of gaussian convolution algorithms. *Image Processing On Line*, 2013: 286–310, 2013.

Bradford Gillespie and Les Atlas. Optimizing time-frequency kernels for classification. *IEEE Transactions on Signal Processing*, 49(3):485–496, 2001.

Hervé Glotin, Julien Ricard, and Randall Balestriero. Fast chirplet transform injects priors in deep learning of animal calls and speech. In *ICLR (Workshop)*, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, volume 1. MIT Press, 2016.

Rémi Gribonval. Fast matching pursuit with a multiscale dictionary of gaussian chirps. *IEEE Transactions on signal Processing*, 49(5):994–1001, 2001.

Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1909.

Werner Heisenberg. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift fur Physik*, 43(3-4):172–198, Mar 1927. doi: 10.1007/BF01397280.

Franz Hlawatsch, Thulasinath Manickam, Rüdiger Urbanke, and William Jones. Smoothed pseudo-wigner distribution, choi-williams distribution, and cone-kernel representation: Ambiguity-domain analysis and experimental comparison. *Signal Processing*, 43(2):149–168, 1995.

Dong-Yan Huang, Minghui Dong, and Haizhou Li. A real-time variable-q non-stationary gabor transform for pitch shifting. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Florent Jaillet and Bruno Torrésani. Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered gabor expansions. *Int. J. of Wavelets, Multiresolution and Inf. Proc.*, 5(02):293–315, 2007.

Arne Jensen and Anders la Cour-Harbo. *Ripples in mathematics: the discrete wavelet transform*. Springer Science & Business Media, 2001.

Jechang Jeong and William Williams. Variable-windowed spectrograms: connecting cohen's class and the wavelet transform. In *Fifth ASSP Workshop on Spectrum Estimation and Modeling*, pages 270–274. IEEE, 1990.

Douglas Jones and Richard Baraniuk. A simple scheme for adapting time-frequency representations. *IEEE Transactions on Signal Processing*, 42(12):3530–3535, 1994.

Haidar Khan and Bulent Yener. Learning filter widths of spectral decompositions with wavelets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Inf. Proc. Sys. 31*, pages 4601–4612. 2018.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. 2018.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

Stefan Lattner, Monika Dörfler, and Andreas Arzt. Learning complex basis functions for invariant representations of audio. *arXiv preprint arXiv:1907.05982*, 2019.

Erwan Le Pennec and Stéphane Mallat. Image compression with geometrical wavelets. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 661–664. IEEE, 2000.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Chagai Levy, Monika Pinchas, and Yosef Pinhasi. Characterization of nonstationary phase noise using the wigner–ville distribution. *Mathematical Problems in Engineering*, 2020, 2020.

Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, and Shenglan Liu. Bottom-up broadcast neural network for music genre classification. *arXiv preprint arXiv:1901.08928*, 2019.

Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11, 2000.

Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Birdvox-full-night: A dataset and benchmark for avian flight call detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2018.

Vincent Lostanlen, Alice Cohen-Hadria, and Juan Pablo Bello. One or two components? the scattering transform answers. *arXiv preprint arXiv:2003.01037*, 2020.

Ying Luo, Qun Zhang, Cheng-wei Qiu, Xian-jiao Liang, and Kai-ming Li. Micro-doppler effect analysis and feature extraction in isar imaging with stepped-frequency chirp signals. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4):2087–2098, 2009.

Stéphane Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Stéphane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

Stéphane Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, July 2012.

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

Ives Meyer. *Wavelets and applications*, volume 31. Masson Paris, 1992.

José Moyal. Quantum mechanics as a statistical theory. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 99–124. Cambridge University Press, 1949.

Albert Nuttall. Wigner distribution function: Relation to short-term spectral estimation, smoothing, and performance in noise. Technical report, Naval Underwater sys. center New London Lab, 1988.

Soo-Chang Pei and Shih-Gu Huang. Stft with adaptive window width based on the chirp rate. *IEEE Transactions on Signal Processing*, 60(8):4065–4080, 2012.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE J. Sel. Topics in Signal Proc.*, 13(2):206–219, 2019.

Kannan Ramchandran and Martin Vetterli. Best wavelet packet bases in a rate-distortion sense. *IEEE Transactions on Image Processing*, 2(2):160–175, 1993.

Mirco Ravanelli and Yoshua Bengio. Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*, 2018.

Debashis Sen. The uncertainty relations in quantum mechanics. *Current Science*, pages 203–218, 2014.

Léonard Seydoux, Randall Balestriero, Piero Poli, Maarten De Hoop, Michel Campillo, and Richard Baraniuk. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, 11(1):1–12, 2020.

Ljubisa Stankovic and Srdjan Stankovic. Wigner distribution of noisy signals. *IEEE Transactions on Signal Processing*, 41(2):956–960, 1993.

Omid Talakoub, Jie Cui, and Willy Wong. Approximating the time-frequency representation of biosignals with chirplets. *EURASIP Journal on Advances in Signal Processing*, 2010:1–10, 2010.

Jean Ville. Theorie et application dela notion de signal analytique. *Câbles et transmissions*, 2(1): 61–74, 1948.

Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

Pete Warden. Speech commands: A public dataset for single-word speech recognition. *Tensrflow*, 2017.

Eugene Wigner. On the Quantum Correction For Thermodynamic Equilibrium. *Physical Review*, 40(5):749–759, Jun 1932. doi: 10.1103/PhysRev.40.749.

Zixiang Xiong, Kannan Ramchandran, and Michael T Orchard. Wavelet packet image coding using space-frequency quantization. *IEEE transactions on image processing*, 7(6):892–898, 1998.

Qinye Yin, Shie Qian, and Aigang Feng. A fast refinement for adaptive gaussian chirplet decomposition. *IEEE transactions on signal processing*, 50(6):1298–1306, 2002.

Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux. Learning filterbanks from raw speech for phone recognition. In *ICASSP*, pages 5509–5513. IEEE, 2018.

Supplementary Material

This appendix proposes to first review the implementation details and visual results of the paper, studying the learned filters, we conclude with all the proofs of the theoretical results.

Appendix A. DN topology

We leverage a time translation covariant form of the proposed learnable model as we aim at solving classification tasks based on audio clips. Hence the representation should be translation invariant. We keep the unconstrained frequency dimensions and thus do not impose any frequency shift invariance as in the Cohen class family of representations. The kernels are parametrized as given in the main text and the networks are given as follows:

```
TF: any representation (morlet, lwvd, ...)
the first mean(3) represents time pooling


- onelayer_nonlinear_scattering
     input = T.log(TF.mean(3).reshape([N, -1])+0.1)
     Dropout(0.3)
     Dense(256)
     BatchNormalization([0])
     LeakyReLU
     Dropout(0.1)
     Dense(n_classes)


 - onelayer_nonlinear_scattering:
     input = T.log(TF.mean(3).reshape([N, -1])+0.1)
     Dropout(0.1)
     Dense(n_classes)

 - joint_linear_scattering:
     feature = T.log(TF.mean(3).reshape([N, -1])+0.1)

     input = T.log(TF+0.1)
     Conv2D(64, (32,16))
     BatchNormalization([0,2,3])
     AbsoluteValue
     Concatenate(AbsoluteValue, feature)
     Dropout(0.1)
     Dense(n_classes)
```

all training is done with the Adam optimizer, same initialization and data splitting.

Appendix B. Additional Figures

We represent in this section the learned filters/kernels Φ applied on the smoothed pseudo Wigner-Ville distribution, for clarity we only depict one every 4 filters, concatenated horizontally. We do so for three dataset and provide analysis in the caption of each figures.

B.1. Samples of learnt filters

We propose in Fig. 5 and Fig. 8 and Fig. 7 the filters after learning for each dataset with their analysis.

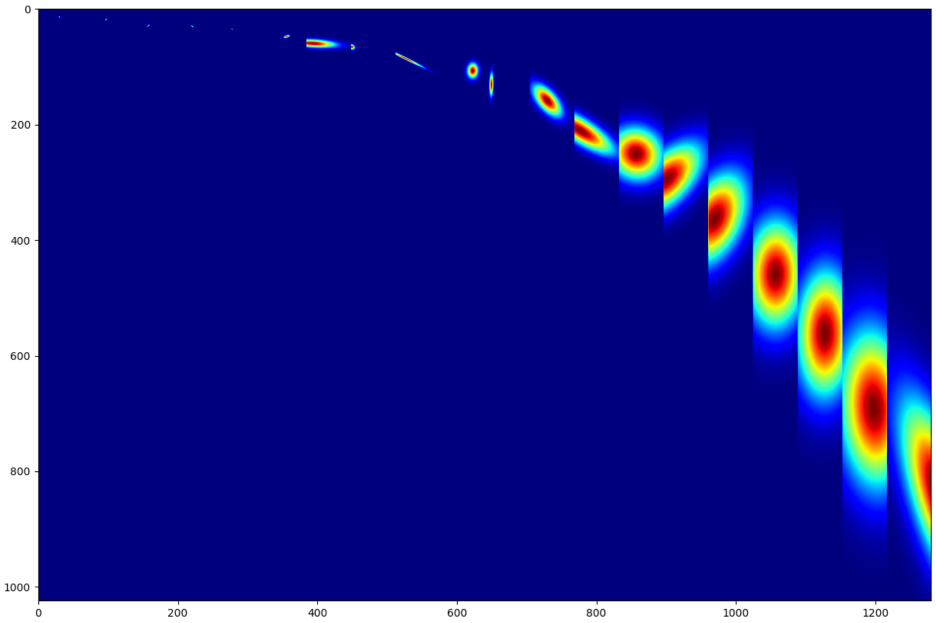

Figure 5: Audio MNIST: this dataset deals with spoken digit classification. A few key observations: the high frequency filters tend to take an horizontal shape greatly favoring frequency resolution, for some of the filters the time resolution is also very high (reaching super-resolution) while others favor local translation invariance. For all the medium to low frequency filters, great frequency and time invariance is preferred (large gaussian support) with a slight chirpness for the medium frequency filters. The low frequency filters tend to favor time resolution.

Appendix C. Proofs

In this section we present in details all the proofs of the main paper results, as well as providing some additional ones.

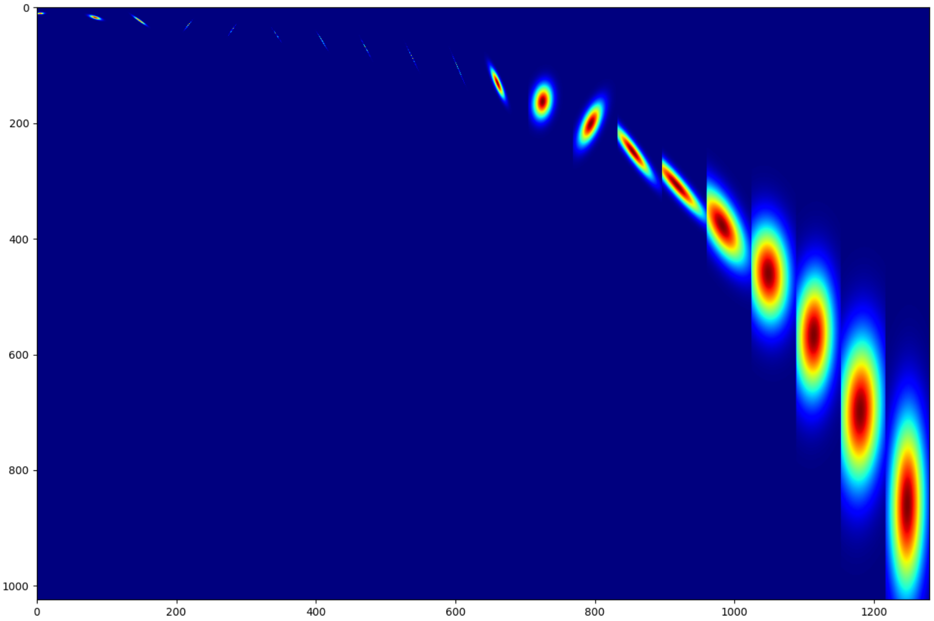

Figure 6: FreeSound: this dataset contains various different classes ranging on different frequencies and without an a priori prefered form of the events in the WV space. As opposed to the AudioMNIST case, we can see that the kernels tend to have smaller covariance (support) hence preferring time and frequency resolution to invariance. This becomes especially true for the high frequency atoms. We also see the clear chirpness for the medium/high frequency kernels with a specific (-30) angle, with decreasing slope. This might be specific to some particular events involving moving objects such as train, cars and so on.

C.1. Proof of Lemma 2

Let first prove the general case with arbitrary kernels

Lemma 8 *The norm of the difference of two representations obtained from kernel Φ and Φ' is bounded above as $\|K_{x,\Phi} - K_{x,\Phi'}\|_{L^2(\mathbb{R}^2)} \leq \|x\|^2_{L^2(\mathbb{R})}\|\Phi - \Phi'\|_{L^2(\mathbb{R}^4)}$. with $\|\Phi - \Phi'\|_{L^2(\mathbb{R}^4)} = \sqrt{\int_{t,f} \|(\Phi - \Phi')[t,f]\|^2_{L^2(\mathbb{R}^2)}}$.*

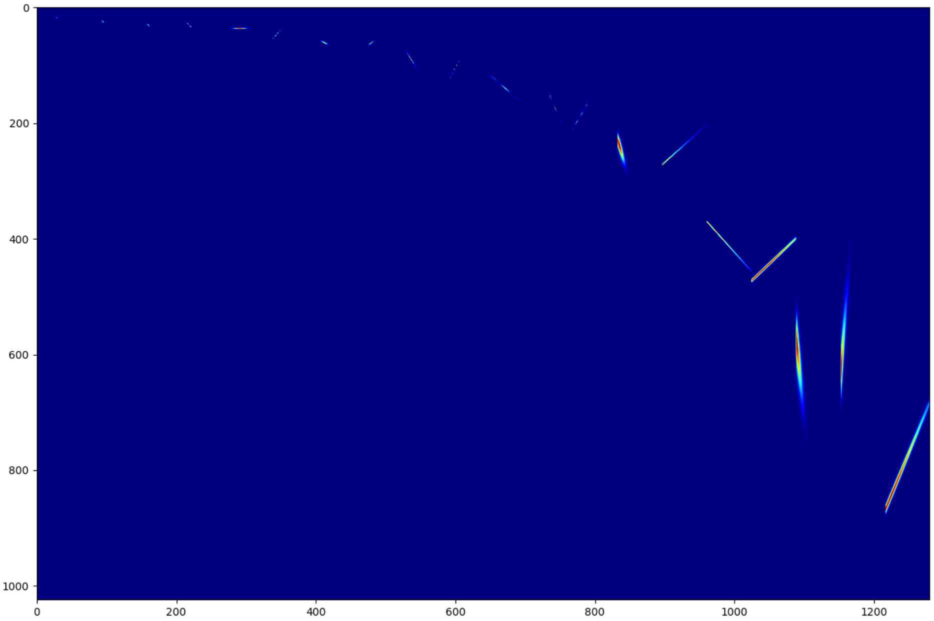

Figure 7: Bird: This dataset proposes to predict the presence or absence of a bird in short audio clips. A priori, detection of such events heavily relies on chirps, the characteristic sound of birds with increasing/decreasing frequency over time. We can see from the learned filters how the learned representation indeed focuses on such patterns, reach super-resolution and thus extreme sensitivity to the time and frequency position of the events, in particular for medium to low frequency kernels. For high frequency kernels, the time and frequency resolution is further increased.

Proof First, one can easily derive $\|\mathrm{WV}_x\|_{L^2(\mathbb{R})} = \|x\|^2_{L^2(\mathbb{R}^2)}$ (see for example Cordero and Nicola (2018)). Given this, and the definition of the K-transform, we obtain that

$$\begin{aligned}
\|\mathrm{K}_{x,\Phi} - \mathrm{K}_{x,\Phi'}\|_{L^2(\mathbb{R}^2)} =& \sqrt{\int_{t,f} (\langle \mathrm{WV}_x, \Phi[t,f]\rangle_{L^2(\mathbb{R}^2)} - \langle \mathrm{WV}_x, \Phi'[t,f]\rangle_{L^2(\mathbb{R}^2)})^2} \\
=& \sqrt{\int_{t,f} \langle \mathrm{WV}_x, \Phi[t,f] - \Phi'[t,f]\rangle^2_{L^2(\mathbb{R}^2)}} \\
\leq& \sqrt{\int_{t,f} \|\mathrm{WV}_x\|^2_{L^2(\mathbb{R}^2)} \|\Phi[t,f] - \Phi'[t,f]\|^2_{L^2(\mathbb{R}^2)}} \quad \text{Cauchy-Schwarz Ineq.} \\
=& \|\mathrm{WV}_x\|_{L^2(\mathbb{R}^2)} \sqrt{\int_{t,f} \|\Phi[t,f] - \Phi'[t,f]\|^2_{L^2(\mathbb{R}^2)}} \\
=& \|x\|^2_{L^2(\mathbb{R})} \sqrt{\int_{t,f} \|\Phi[t,f] - \Phi'[t,f]\|^2_{L^2(\mathbb{R}^2)}} \\
=& \|x\|^2_{L^2(\mathbb{R})} \|\Phi - \Phi'\|_{L^2(\mathbb{R}^4)}
\end{aligned}$$

Table 2: Std over 10 runs of classification result using three architectures, one layer scattering followed by a linear classifier (Linear Scattering), one layer scattering followed by a two layer neural network (Nonlinear Scattering) and a two layer scattering with joint (2D) convolution for the second layer followed by a linear classifier (Linear Joint Scattering). For each architecture and dataset, we experiment with the baseline, Morlet wavelet fitler-bank (morlet), and learnable frameworks being ours (lwvd), learnable sinc based filters (sinc) and learnable Morlet wavelet (lmorlet). As can be seen, across the dataset, architectures and learning rates, the proposed method provides significant performance gains.

		Linear Scattering				Nonlinear Scattering				Linear Joint Scattering			
	le. rate	*morlet*	lwvd	sinc	lmorlet	*morlet*	lwvd	sinc	lmorlet	*morlet*	lwvd	sinc	lmorlet
DOCC10	0.0002	3	0.2	1.1	1.1	0.8	0.2	24.6	0.8	0.2	0.1	40	0.1
	0.001	1.8	0.1	1.5	1.4	1.5	0.2	25.6	0.6	0.8	0.2	26.7	0.3
	0.005	1.4	0.9	1.6	0.6	1.2	28.8	14.3	0.5	0.4	25.8	34.2	0.5
BirdVox	0.0002	0.3	0	1	0.2	0.2	0.2	42.6	0.1	0.3	0.3	0.3	0.2
	0.001	0.7	0.2	1.9	1.6	0.2	0.2	29.1	0.3	0.4	0.4	0.8	0.3
	0.005	0.8	0.9	1.1	0.9	0.2	47.1	28.5	0.3	0.4	37.9	25.9	0.5
MNIST	0.0002	0.8	0.4	0.6	1	0.1	0.1	27.5	0.2	0.2	0.1	38.6	0.2
	0.001	1.9	0.4	1.6	1.5	0.3	0.3	0.6	0.3	0.6	0.1	0.6	1.1
	0.005	2.4	0.9	3.8	2.5	0.3	38.4	4.8	0.2	1.4	31.9	25.7	2.1
command	0.0002	0.3	0.1	0.7	0.5	0.3	0.4	0.2	0.3	0.7	0.2	8.1	0.6
	0.001	0.7	0.2	0.4	0.8	0.2	0.2	0.2	0.1	1.8	0.5	24.8	3
	0.005	0.5	0.4	0.6	0.1	0.1	18.8	0.8	0.3	3.4	30.6	14.9	1.1
fsd	0.0002	0.5	0.2	2.9	0.2	0.4	0.3	7	0.5	1	0.8	4.7	1
	0.001	0.9	0.6	1.1	0.5	0.4	0.6	16.1	0.8	0.9	0.6	6.2	1.3
	0.005	1.1	1.7	1.3	1.1	0.6	0.7	1.2	0.7	0.8	1.2	5.9	1.2

■

Now the proof for the special case of a 2D Gaussian kernel follows the exact same procedure as the one above for the Lipschitz constant of the general K-transform.

Proof In the same way we directly have

$$\|\mathrm{K}_{x,\Phi} - \mathrm{K}_{x,\Phi'}\|_{L^2(\mathbb{R}^2)} \leq \|x\|^2_{L^2(\mathbb{R})} \sqrt{\int_{t,f} \|\Phi[t,f](\theta) - \Phi[t,f](\theta')\|^2_{L^2(\mathbb{R}^2)}}$$

now the only different is that we need a last inequality to express the distance in term of the θ parameters and not the kernels Φ. To do so, we leverage the fact that the parametric kernel is a 2-dimensional Gaussian with zero mean and unit variance and leverage the following result:

$$\|f(\theta) - f(\theta')\| \leq \max_u \|\nabla_f(u)\| \|\theta - \theta'\|$$

and simply denote by κ the maximum of the Gaussian gradient norm, leading to the desired result by setting f the 2-dimensional Gaussian. ■

C.2. Proof of Prop. 1

First we prove the more general result which follows.

Lemma 9 *For any TFR or signal adapted TFR, there exists a kernel Φ such that $K_{x,\Phi}$ (recall (4)) is equal to it.*

Proof The proof is a direct application of Moyal Theorem which states that given a signal x and a filter y the application of the filter onto the signal and squaring the result can be expressed as the inner product between the Wigner-Ville transforms of the signal and filter as in

$$|\int_{-\infty}^{\infty} x(t)y^{*}(t)dt|^{2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{WV}_{x}(\tau,\omega)\mathrm{WV}_{y}(\tau,\omega)d\tau d\omega.$$

From the above we can see that any time frequency representation (time invariant or not) can be recovered simply by setting $\Phi[t,f] = \mathrm{WV}_{y}$ of some desired filter y. ■

Now for the special case of the Gabor based transform we can leverage the above result and proof with the following.

Proof This result is a direct application of the analytically derived kernels from Nuttall (1988); Flandrin and Rioul (1990); Jeong and Williams (1990); Baraniuk and Jones (1996); Talakoub et al. (2010) for various analytical time-frequency representations. For the Gabor transform, Morlet wavelet and Morlet based Chirplet atom, all the analytically derived kernels that are applied onto a Wigner-Ville distribution as a parametric form of the 2-dimensional Gaussian kernel proving the result. ■

C.3. Proof of Theorem 1

Proof The proof follows directly from the above result. In fact, we obtained above that the Lipschitz constant of the K-transform w.r.t. the Φ parameter is obtained by $\|x\|^{2}_{L^{2}(\mathbb{R})}$. Thus, the Lipschitz continuity implies continuity of the transform w.r.t. the Φ parameters which proves the result (see for example Kreyszig (1978)). ■

C.4. Proof of Lemma on Invariant

Proof

We provide the proof for each case below:

- Time translation of the signal x by τ to have $y(t) = x(t-\tau)$ leads to

$$\begin{aligned}\mathrm{WV}_{y}(t,\omega) &= \int_{-\infty}^{\infty} y(t+\frac{\tau}{2})y^{*}(t-\frac{\tau}{2})e^{-i\tau\omega}d\tau\\ &= \int_{-\infty}^{\infty} x(t+\frac{\tau}{2}-\tau)x^{*}(t-\frac{\tau}{2}-\tau)e^{-i\tau\omega}d\tau\\ &= \mathrm{WV}_{x}(t-\tau,\omega)\end{aligned}$$

Thus, the time equivariance of the representation defined as $\mathrm{K}_{y,\Phi}(t,f) = \mathrm{K}_{x,\Phi}(t-\tau,f)$ is obtained iff

$$\begin{aligned}\mathrm{K}_{y,\Phi}(t,f) =& \langle \mathrm{WV}_y, \Phi[t,f]\rangle = \langle \mathrm{WV}_x, \Phi[t,f](.+\tau,.)\rangle = \langle \mathrm{WV}_x, \Phi[t-\tau,f]\rangle \\ &\iff \Phi[t,f](.+\tau,.) = \Phi[t-\tau,f]\end{aligned}$$

as a result the above condition demonstrates that filters of different times $\Phi[t-\tau,f], \forall\tau$are just time translations of each other giving the desired result.

- Frequency modulation/shift with frequency ω_0 to have $y(t) = x(t)e^{i\omega_0 t}$ leads

$$\begin{aligned}\mathrm{WV}_y(t,\omega) &= \int_{-\infty}^{\infty} y(t+\frac{\tau}{2})y^*(t-\frac{\tau}{2})e^{-i\tau\omega}d\tau \\ &= \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})e^{i\omega_0(t+\frac{\tau}{2})}x^*(t-\frac{\tau}{2})e^{-i\omega_0(t-\frac{\tau}{2})}e^{-i\tau\omega}d\tau \\ &= \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})e^{i\omega_0\tau)}x^*(t-\frac{\tau}{2})e^{-i\tau\omega}d\tau = W_x(t-\tau,\omega) \\ &= \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})e^{-i\tau(\omega+\omega_0)}d\tau = W_x(t-\tau,\omega) \\ &= \mathrm{WV}_x(t,\omega+\omega_0)\end{aligned}$$

now leveraging the same result that for the time equivariance, we obtain the desired result.

- For completeness, we also demonstrate here how the Wigner-Ville behaves under rescaling of the input $y(t) = x(t/a)$

$$\begin{aligned}\mathrm{WV}_y(t,\omega) &= \int_{-\infty}^{\infty} y(t+\frac{\tau}{2})y^*(t-\frac{\tau}{2})e^{-i\tau\omega}d\tau \\ &= \int_{-\infty}^{\infty} x((t+\frac{\tau}{2})/a)x^*((t-\frac{\tau}{2})/a)e^{-i\tau\omega}d\tau \\ &= \int_{-\infty}^{\infty} x(t/a+\frac{\tau}{2a}))x^*(t/a-\frac{\tau}{2a})e^{-i\tau\omega}d\tau \\ &= \int_{-\infty}^{\infty} x(t/a+\frac{\tau}{2a}))x^*(t/a-\frac{\tau}{2a})e^{-i\tau\omega}d\tau \\ &= a\int_{-\infty}^{\infty} x(t/a+\frac{\tau}{2}))x^*(t/a-\frac{\tau}{2})e^{-i\tau a\omega}d\tau \\ &= a\mathrm{WV}_x(t/a,a\omega)\end{aligned}$$

■

C.5. Proof of Lemma 5

Proof The proof consists of first computing the analytical form of the noisy Wigner-Ville transform. Since the K-transform is a linear transformation of the Wigner-Ville transform we will obtain our result from that. Consider the noisy signal x made of the true underlying signal y augmented with additive noise ϵ as in $x = y + \epsilon$. By using Eq. 4 from Stankovic and Stankovic (1993) we obtain that

$$\begin{aligned}\mathbb{E}\left[\mathrm{K}_x(t,f)\right] =&\mathbb{E}\left[\int_{\mathbb{R}\times[0,2\pi)} \mathrm{WV}_x(\tau,\omega)\Phi_f(t-\tau,\omega)d\tau d\omega\right]\\ =&\int_{\mathbb{R}\times[0,2\pi)} \mathbb{E}\left[\mathrm{WV}_x(\tau,\omega)\right]\Phi_f(t-\tau,\omega)d\tau d\omega\\ =&\int_{\mathbb{R}\times[0,2\pi)} \left(\mathrm{WV}_y(\tau,\omega)+S(\omega)\right)\Phi_f(t-\tau,\omega)d\tau d\omega\\ =&\mathrm{K}_y(t,f)+\int_{\mathbb{R}\times[0,2\pi)} S(\omega)\Phi_f(t-\tau,\omega)d\tau d\omega\\ =&\mathrm{K}_y(t,f)+\int_0^{2\pi} S(\omega)\int_{\mathbb{R}} \Phi_f(t-\tau,\omega)d\tau d\omega\\ =&\mathrm{K}_y(t,f)+\int_{[0,2\pi)} S(\omega)\phi_f(\omega)d\omega\end{aligned}$$

where the last equality comes from the fact that integrating a two-dimensional Gaussian with respect to one dimension (the time one in this case) produces a one-dimensional Gaussian with mean and variance equal to the ones of the remaining dimension. For details on this step we refer the reader to Theorem 4 in http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html leading to the desired result. Be exploiting Eq. 5 from Stankovic and Stankovic (1993) it is also possible to derive the variance of the noisy K-transform using the same recipe. ■

C.6. Proof of Prop 2

Proof The proof is obtained by using the fact that the maximum of the 2-dimensional Gaussian with given covariance matrix will always be smaller than $\frac{1}{\det(\Sigma)}$ and is always nonnegative, and then simply upper bounding the norm difference by this value times the representation. Now that we have the upper bound in term of $\|WV_x - WV_{D(X)}\|_{L^2(\mathbb{R}^2)}$ simply apply the Lipschitz constant inequality with κ the constant of the WVD which exists as long as x is bounded which is the case as we have $\mathbb{L}^2(\mathbb{R})$ leading to the desired result. ■

C.7. Proof of Lemma 6

Proof We will obtain the desired result by unrolling the following equations and see that it coincides with the one of the Theorem with the given kernel applied on the WV representation:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{ST}_x(t, \omega + \frac{\eta}{2})\mathrm{ST}_x^*(t, \omega - \frac{\eta}{2})\mathcal{F}_g(\eta)e^{j2\pi\eta t}d\eta \\
&= \int_{-\infty}^{\infty} \mathrm{ST}_x(t, \omega + \frac{\eta}{2})\mathrm{ST}_x^*(t, \omega - \frac{\eta}{2}) \int_{-\infty}^{\infty} g(\xi)e^{-i\xi\eta}d\xi e^{j2\pi\eta t}d\eta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w^*(t-\tau)x(\tau)w(t-\theta)x^*(\theta)e^{-i\omega(\tau-\theta)} \int_{-\infty}^{\infty} g(\xi) \int_{-\infty}^{\infty} e^{j2\pi\eta(t-\frac{\tau}{2}-\frac{\theta}{2}-\xi)}d\eta d\xi d\tau d\theta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w^*(t-\tau)x(\tau)w(t-\theta)x^*(\theta)e^{-i\omega(\tau-\theta)}g(\xi)\delta(t-\frac{\tau}{2}-\frac{\theta}{2}-\xi)d\xi d\tau d\theta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w^*(t-\tau)x(\tau)w(\tau+2\xi-t)x^*(2t-\tau-2\xi)g(\xi)e^{-i\omega(2\tau-2t+2\xi)}d\xi d\tau \\
&= \int_{-\infty}^{\infty} g(\xi)\int_{-\infty}^{\infty} w^*(\xi-\mu/2)x(t+\mu/2-\xi)w(\xi+\mu/2)x^*(t-\mu/2-\xi)e^{-i\omega(2\mu)}d\tau d\xi \\
&\qquad\qquad (\mu = 2(\tau - t + \xi)) \\
&= \int g(\xi)(W_w(\xi,.) \star W_x(t-\xi,.))(\omega)/2d\xi \implies \Pi'(\tau,\omega) = g(\tau)W_w(\tau,\omega)/2
\end{aligned}$$

with $\Pi'(t,\omega) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{t^2}{\sigma_t^2/(\sigma_\omega^2\sigma_t^2+1)}-\frac{\omega^2\sigma_t^2}{2}}$.

■

C.8. Proof of Thm. 2

Proof From the above (lemma) result, it then follows directly that performing convolutions with two gaussians can be rewritten as a single Gaussian convolution with the given parameters based on both Gaussians. ■

Appendix D. Interferences

Finally, we now consider the study of interference that can arise in the K-transform whenever multiple events occur in the signal x Sen (2014). In fact, the Gabor limit also known as the Heisenberg Uncertainty principle Heisenberg (1927), corresponds to the optimal limit one can achieve in time and frequency resolution without introducing interference in the representation. In our case, no constraints are imposed onto $\Phi[t, f]$, and while learning will allow to reach any desired kernel for the task at hand, we propose some conditions that would prevent the presence of interference. In fact, there is a general necessary condition ensuring that the representation does not contain any interference.

Proposition 3 *A sufficient condition to ensure absence of interference in the K-transform is to have a nonnegative representation as in* $K_{x,\Phi}(t, f) \geq 0, \forall t, f$.

Proof We only consider signals that can be expressed as a linear combination of oscillatory atoms (natural signals) Mallat (1989). Due to the form of the Wigner-Ville distribution (with the quadratic term) the interference are oscillatory terms and thus can not be positive only. In short, there can not be a positive coefficient appearing due to an interference without its negative counterpart. As a result, a nonnegative representation can not have any interference. See (Cohen (1995)) for more details and background on Wigner-Ville distribution and interference. ■

Lastly, thanks to the Gaussian form, we can directly obtain the shape and in particular area of the logon (the inverse of the joint time and frequency resolution) as follows.

Proposition 4 *A sufficient condition to ensure absence of interference in the K-transform is to have the effective area of the 2D Gaussian greater that $\frac{1}{4\pi}$, that is, $\sigma'_{\rm T}\sigma'_{\rm F} \geq \frac{1}{4\pi}, \forall t, f$ with*

$$\sigma'_{\rm T} = \sigma_{\rm T}\cos^2(\theta) + 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\sin^2(\theta) \tag{8}$$

$$\sigma'_{\rm F} = \sigma_{\rm T}\sin^2(\theta) - 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\cos^2(\theta), \tag{9}$$

$$\theta = \frac{\arctan(\frac{2\rho}{\sigma_{\rm T}-\sigma_{\rm F}})}{2} \tag{10}$$

Proof The proof applies the Uncertainty principle which provides the minimal area of the Logon to ensure absence of interference (Gabor (1946)), this is then combined with the above result on the area of the Logon in the case of our 2-dimension Gaussian to obtain the desired result. So first, the condition is that $\sigma_{\rm T}\sigma_{\rm F} \geq \frac{1}{4\pi}$, however in our case we also have the possible chirpness parameter. But we can obtain the new rotated covariance matrix s.t. the chirpness is removed by rotating the time-frequency place instead. This can be done easily (see for example `http://scipp.ucsc.edu/~haber/ph116A/diag2x2_11.pdf`) to obtain the new parameter as

$$\begin{aligned}\sigma_{\rm T} =& \sigma_{\rm T}\cos^2(\theta) + 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\sin^2(\theta)\\ \sigma_{\rm F} =& \sigma_{\rm T}\sin^2(\theta) - 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\cos^2(\theta),\\ \theta =& \frac{\arctan(\frac{2\rho}{\sigma_{\rm T}-\sigma_{\rm F}})}{2}\end{aligned}$$

now the constraint can be expressed as

$$\left(\sigma_{\rm T}\cos^2(\theta) + 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\sin^2(\theta)\right)\left(\sigma_{\rm T}\sin^2(\theta) - 2\rho\cos(\theta)\sin(\theta) + \sigma_{\rm F}\cos^2(\theta)\right) \geq \frac{1}{4\pi}$$

■

Appendix E. Gaussian Truncation

Notice however that the STFT has to be done by padding the signal windows to allow interpolation, this is common when dealing with such transformations. Given some parameters, we convert them into discrete bins to get actual window sizes as follows

$$N_\sigma(\epsilon) = -g_\sigma^{-1}(\epsilon) \times 2Fs$$

with g_σ^{-1} the inverse of the gaussian density distribution with 0 mean and σ standard deviation, taken only on the negative part of its support. As such, N is a function that maps a given tolerance and standard deviation to the window length in bins s.t. the apodization window at the boundaries of this window are of no more than ϵ.

Table 3: Various hand picked K-transform parameters leading to known time invariant TFRs with their respective parameters. For the adaptive versions, such as the wavelet tree with various Gabor mother wavelet parameters σ_0, then at each time t, the corresponding parameter is any of the optimal one based on the desired criterion.

	$\mu_{\text{time}}(t,f)$	$\mu_{\text{freq}}(t,f)$	$\sigma_{\text{time}}(t,f)$	$\sigma_{\text{freq}}(t,f)$	$\rho(t,f)$ (chirpness)
Spectrogram	t	f	σ_t	σ_t^{-1}	0
Melscale Spectrogram	t	$2^{S(1-f/\pi)}$	σ_t	$2^{S(f/\pi-1)}\sigma_t^{-1}$	0
Scalogram	t	$2^{S(1-f/\pi)}$	$2^{S(1-f/\pi)}\sigma_0$	$2^{S(f/\pi-1)}\sigma_0^{-1}$	0
Scattering Layer	t	$2^{S(1-f/\pi)}$	$s2^{S(1-f/\pi)}\sigma_0$	$2^{S(f/\pi-1)}\sigma_0^{-1}$	0
Chirpogram	t	$2^{S(1-f/\pi)}$	$2^{S(1-f/\pi)}\sigma_0$	$2^{S(f/\pi-1)}\sigma_0^{-1}$	$\rho(t,f) \neq 0$

Appendix F. Example of K-transform parameterization

We propose in Table 3 different configurations of the parameter θ corresponding to some standard and known TFRs.

Appendix G. Computational Complexity

The computational complexity of the method varies with the covariance matrix of the kernels, it will range from quadratic when the kernels tend toward a Delta function (as the representation computation bottleneck becomes the computation of the exact WV) to linear for large covariance matrices. In such case, the computation is done with spectrograms of small windows (and with almost no spectral frequency correlation) for the smoothed pseudo WV. In addition to those cases, the use of translation (in time and/or frequency) also greatly reduce computational costs as it allows for the use of convolutions into the K-transform computation. This allows to put all the computation bottleneck into the computation of the smoothed pseudo WV.

Appendix H. Data sets description

Audio-MNIST. This dataset consists of multiple (60) speakers of various characteristics enunciation digits from 0 to 9 inclusive Becker et al. (2018). The classification task consists of classifying the spoken digit, $30,000$ recordings are given in the dataset. Each recording is from a controlled environment without external noise except breathing and standard recording device artifacts. **BirdVox-70k.** This dataset consists of avian recording obtained during the fall migration season of 2015. The recording were obtained in North American and is made of a balanced binary classification of predicting the presence or not of a bird in a given audio clip Lostanlen et al. (2018). In total $70,000$ clips are contaiend in the dataset. The outside recordings present various types of non stationnary noises and sources. **Google Speech Commands.** This dataset consists of $65,000$ one-second long utterances of 30 short words such as “yes”, “no”, “right” or “left” Warden (2017). The recordings are obtained from thousands of different people who contributed through the AIY website [1]. The task is thus to classify the spoken command. **FreeSound DCASE.** This task addresses the problem of general-purpose automatic audio tagging Fonseca et al. (2018). The dataset

[1] https://aiyprojects.withgoogle.com/open_speech_recording

provided for this task is a reduced subset from FreeSound and is a large-scale, general-purpose audio dataset annotated with labels from the AudioSet Ontology. There are 41 classes and ≈95,000 training samples, classes are diverse such as Bus, Computer keyboard, Flute, Laughter or Bark. The main challenge of the task is the noise present in the training set labels reflecting the expensiveness of having high quality annotations.

DOCC10. The goal of this dataset is to classify biosonar waveform from cetaceans recorded from various subsea acoustic stations or autonomous surface vehicles. This task (Ferrari et al., 2020) has been part of ENS Paris College de France challenge. The recordings of the clicks are from a post-processed subset of DCLDE 2018 challenge (9 species, http://sabiod.org/DCLDE) from Scripps Institute, plus recordings of Cachalot (Physteter macrocephalus) from ASV Sphyrna (http://sabiod.org/pub/SO1.pdf) near Toulon, France. Each input signal is 8192 bins, at a sampling rate of 200 kHz. They each includes a click centered in the middle of the window in the case of the test set. The clicks in the training set are in various positions and background noises.

Challenge goal is to classify each click according to the corresponding emitting species. The 10 species are : (0) Gg: Grampus griseus- Risso's dolphin (1) Gma: Globicephala macrorhynchus- Short-finned pilot whale (2) La: Lagenorhynchus acutus- Atlantic white-sided dolphin (3) Mb: Mesoplodon bidens- Sowerby's beaked whale (4) Me: Mesoplodon europaeus- Gervais beaked whale (5) Pm: Physeter macrocephalus - Sperm whale (6) Ssp: Stenella sp.Stenellid dolphin (7) UDA: Delphinid type A - a group of dolphins (species not yet determined) (8) UDB: Delphinid type B - another group of dolphins (species not yet determined) (9) Zc: Ziphius cavirostris- Cuvier's beaked whale The metric is the MAP. Most of the recording consist of acoustic recordings from multiple deployments of high-frequency acoustic recording packages (Wiggins and Hildebrand, 2007) deployed in the Western North Atlantic (US EEZ) and Gulf of Mexico. It has a 100 kHz of bandwidth (200 kHz sample rate) http://sabiod.org/DCLDE/challenge.html#highFreqData. The initial data set is 3 To with weak labeled. (Ferrari et al., 2020) filtered the labels by several detector and clustering, yielding to 90 000 samples (6 Go). Another part (1 class) is from Mediterranean sea, ASV Sphyrna, at 384 kHz SR downsampled at 200 kHz SR. In sum, it yields around 11312 samples per class in the train set, and 2096 samples per class in the test set.

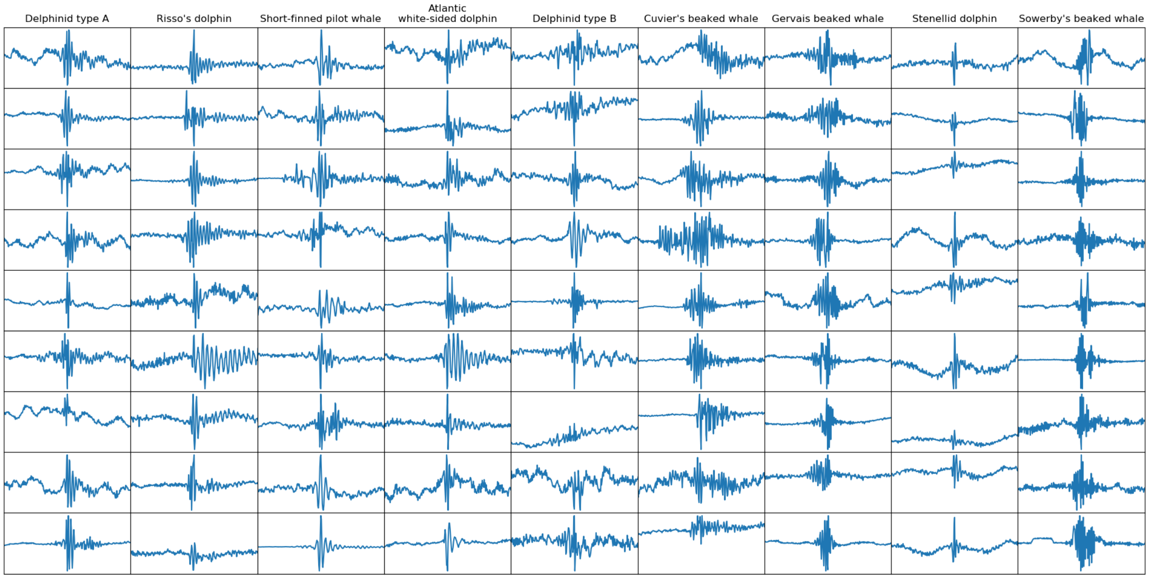

Figure 8: DOCC10 dataset samples, one class per column. The click is centered. Details in ENS DATA CHALLENGE WEB SITE.