Learnable Group Transform For Time-Series

Romain Cosentino [1] **Behnaam Aazhang** [1]

Abstract

We propose a novel approach to filter bank learning for time-series by considering spectral decompositions of signals defined as a Group Transform. This framework allows us to generalize classical time-frequency transformations such as the Wavelet Transform, and to efficiently learn the representation of signals. While the creation of the wavelet transform filter-bank relies on affine transformations of a mother filter, our approach allows for non-linear transformations. The transformations induced by such maps enable us to span a larger class of signal representations, from wavelet to chirplet-like filters. We propose a parameterization of such a non-linear map such that its sampling can be optimized for a specific task and signal. The Learnable Group Transform can be cast into a Deep Neural Network. The experiments on diverse time-series datasets demonstrate the expressivity of this framework, which competes with state-of-the-art performances.

1. Introduction

To this day, the front-end processing of time-series remains a keystone toward the improvement of a wealth of applications such as health-care (Saritha et al., 2008), environmental sound (Balestriero et al., 2018; Lelandais & Glotin, 2008), and seismic data analysis (Seydoux et al., 2016). The common denominator of the recorded signals in these fields is their undulatory behavior. While these signals share this common behavior, two significant factors imply the need of learning the representation: **1)** time-series are intrinsically different because of their physical nature, **2)** the machine learning task can be different even within the same type of data. Therefore, the representation should be induced by both the signal and the task at hand.

A common approach to performing inference on time-series consists of building a Deep Neural Network (DNN) that operates on a spectral decomposition of the time-series such as wavelet transform (WT) or Mel Frequency Spectral Coefficients (MFSC). These decompositions represent the signal. While the use of these decompositions is extensive, we show in Section 2 their inherent biases and motivate the development of a generalized framework. The selection of the judicious transform is either performed by an expert on the signal at hand, or by considering filter selection methods (Coifman & Wickerhauser, 1992; Mallat & Zhang, 1993; Gribonval & Bacry, 2003). However, an inherent drawback is that the selection of the filters decomposing the signals is often achieved with criteria that do not align with the task. For instance, a selection based on the sparsity of the representation while the task is the classification of the signals. Besides, these selection methods and transformations require substantial cross-validations of a large number of hyperparameters such as mother filter family, number of octaves, number of wavelets per octave, size of the window (Le & Argoul, 2004; Cosentino et al., 2017).

In this work, we alleviate these drawbacks by proposing a simple and efficient approach by considering the generalization of these spectral decompositions. They consist of taking the inner product between filters and the signals. From one decomposition to the other, only the filter bank differs. The filters of well-known spectral decompositions, such as the short-time Fourier transform (STFT) and the continuous wavelet transform (CWT) are built following a particular scheme. Each filter is the result of the action of a transformation map on a selected mother filter, e.g., a Gabor filter. If the transformation map is induced by a Group, the representation is called a Group Transform (GT), and both the group with the mother filter characterize the decomposition.

We propose to enable the learnability of such a scheme. More precisely, our contributions are: 1) we generalize common Group Transforms by proposing the utilization of strictly increasing and nonlinear transformations, 2) draw the connection between filters that can be learned by our framework and commonly observed filters in biological time-series 3) we show how the equivariance properties of the representation differs from traditional affine transforma-

[1]Department of Electrical and Computer Engineering, Rice University, USA. Correspondence to: Romain Cosentino <rom.cosentino@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

tions, Section 3.1, 4) we propose an efficient way of optimizing the sampling of such functional space, Section 3.2, and 5) apply our method to three datasets containing complementary challenges a) artificial data showing the limitation and drawbacks of well-known GTs, b) a large bird detection dataset ($\approx$ 20 hours of audio recording, $20\times$ larger than CIFAR10 in term of number of scalar values in the dataset) where optimal spectral decomposition are known and developed by expert, and c) a haptic dataset that does not benefit from expert knowledge regarding important features, Section 4.

We can summarize our approach to

- given a filter ψ with its analytical formula
- generate increasing and continuous maps using 1-Layer Relu Network (the number of increasing and continuous map will be the number of filters in the filter bank)
- compose the increasing and continuous maps with the filter ψ
- convolve the filters obtained with the signal to acquire the representation

2. Related Work and Background

One approach to represent the data consists of building equivariant-invariant representations. For instance, in (Mallat, 2012; Bruna, 2013) they propose a translation-invariant representation, the Scattering Transform, which is stable under the action of small diffeomorphisms. In (Oyallon et al., 2018; Cohen & Welling, 2016), they focus on equivariant-invariant representations for images, which reduces the sample complexity and endow DNN's layers with interpretability.

The closest work to ours consist of learning the filter bank in an end-to-end fashion. (Cakir et al., 2016; Ravanelli & Bengio, 2018; Balestriero et al., 2018; Zeghidour et al., 2018) investigated the learnability of a mother filter such that it can be jointly optimized with the DNN. In order to build the filter bank, this learnable mother filter is transformed by deterministic affine maps. The representation of the signal is obtained by convolving the filter bank elements with the signals. Recently, (Khan & Yener, 2018) investigated the learnability of the affine transformations, that is, the sampling of the dilation parameter of the affine group inducing the wavelet filter bank. Optimized jointly with the DNN, their method allows for an adaptive transformation of the mother filter. Our work generalizes this approach and provide its theoretical properties and building blocks.

One of the main drawbacks of these approaches using time-frequency representation is that the filter bank induces a bias that might not be adapted to the data. This bias can be understood by considering the time-frequency tiling of each GT. It is known that the spread of a filter and its Fourier transform are inversely proportional as per the Heisenberg uncertainty principle (Mallat, 1999).

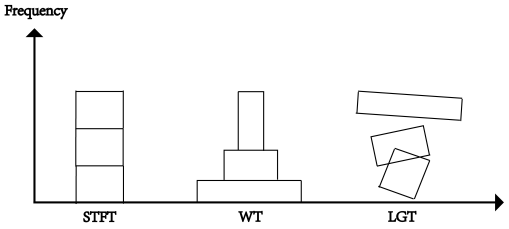

Figure 1. **Time-Frequency Tilings** at a given time τ: (*left*) short-time Fourier transform, i.e., constant bandwidth, (*middle*) wavelet transform, i.e., proportional bandwidth, (*right*) Learnable Group Transform, i.e, adaptive bandwidth, the "tiling" is induced by the learned non-linear transformation underlying the filter bank decomposition.

Following this principle, we can observe that in the case of STFT (respectively WT with a Gabor wavelet), at a given time τ, the signal is transformed by a window of constant bandwidth (respectively proportional bandwidth) modulated by complex exponential resulting in a uniform tiling (respectively proportional) on the frequency axis, Figure 1. This implies that, for instance, in the case of WT, the precision in frequency degrades as the frequency increases while its precision in time increases (Mallat, 1999). Thus, WT is not adapted for fast-varying frequency signals (Xu et al., 2016). In the case of STFT, the uniform tiling implies that the precision is constant along the frequency axis. In our proposed framework, the LGT allows for an adaptive tiling, as illustrated in Figure 1 such that the trade-off between time and frequency precision depends on the task and data.

3. Learnable Group Transform

Common time-frequency filter banks are built by transforming a mother filter that we denote by ψ. We consider the transformations of this mother filter defined as $\psi \circ g, \quad g \in \mathcal{F}$, where $\mathcal{F}$ defines the functional space of the transformation and $\psi \circ g$ denotes the function composition. Note that in signal processing, such a transformation is called warping (Goldenstein & Gomes, 1999; Kerkyacharian et al., 2004). Given a space $\mathcal{F}$, the filter bank with K filters is created by first, sampling K transformation maps from $\mathcal{F}$ and then, by transforming the mother filter such as

$$\{\psi \circ g_1, \dots, \psi \circ g_K | g_1, \dots, g_K \in \mathcal{F}\}.$$

Now, let's denote a signal by $s \in \mathbb{L}_2(\mathbb{R})$, we will consider the representation of the signal as the result of its convolu-

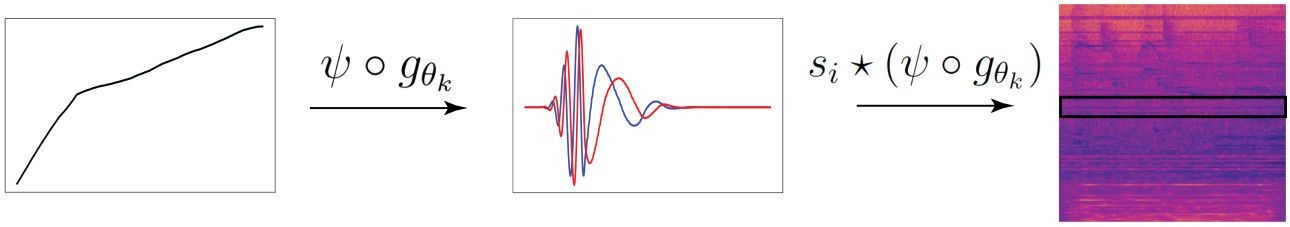

Figure 2. **Learnable Group Transform:** (left) generating the strictly increasing continuous functions g_{θ_k} with parameters θ_k, $\forall k \in \{1, \dots, K\}$, where K denotes the number of filters in the filter bank. The x-axis is the time variable and the y-axis the amplitude. (middle) The mother filter, ψ (presently a Morlet wavelet), is composed with each warping function g_{θ_k} , where the imaginary part is shown in red and the real part in blue. The x-axis represents the time and y-axis the amplitude of the filter. These transformations lead to the filter bank (only the k^{th} element is displayed). Then, the convolutions between the filter bank elements and the signal s_i lead to the LGT of the signal. The black box on the LGT representation (right) corresponds to the convolution of the k^{th} filter with the signal. In this figure, the horizontal axis corresponds to the time, each row corresponds to the convolution with a filter of the filter bank, and the color displays the amplitude of each inner product. Notice that a complex modulus has been applied to the LGT. The strictly increasing and continuous piecewise linear functions can be learned efficiently by back-propagating the error induced by the generated GT.

tion with the filter bank elements and denote it by

$$\mathcal{W}[s, \psi](\mathbf{g}, .) = [\mathcal{W}[s, \psi](g_1, .), \dots, \mathcal{W}[s, \psi](g_K, .)]^T,$$

where

$$\mathcal{W}[s, \psi](g, .) = s_i \star (\psi \circ g), \forall g \in \mathcal{F},$$

with $\star$ the convolution operator and $(.)$ corresponds to the time axis.

Therefore, the properties of the representation are carried by the mother filter ψ, and space $\mathcal{F}$. In this work, we focus on the warping that generalizes common time-frequency decompositions as well as the properties carried by the associated filter bank, in particular we consider nonlinear warping. We provide a parameterization of such a warping and show how one can efficiently learn these parameters. The decomposition of the signal by this learned filter bank defines a Group Transform. The overall building blocks of the LGT, and its application on a signal is depicted in Figure 2.

3.1. Time Warped Filters

We propose to transform the mother filter by means of a subset of invertible maps on $\mathbb{R}$. Instead of the affine warping used in WT, we propose the use of a more general transformation map space $\mathcal{F}$. In particular, we will use the space of strictly increasing and continuous functions defined as

$$C^0_{\text{inc}}(\mathbb{R}) = \left\{ g \in C^0(\mathbb{R}) | g \text{ is strictly increasing} \right\},$$

where $C^0(\mathbb{R})$ defines the space of continuous functions defined on $\mathbb{R}$. This set of functions is composed of invertible maps which is crucial in order to derive invariance properties as well as avoid artifacts in the transformed filters.

The transformation of a mother filter ψ is defined by the linear operator $\rho_{\text{inc}}(g)$ such as

$$\rho_{\text{inc}}(g)\psi = \psi \circ g, \;\; g \in C^0_{\text{inc}}(\mathbb{R}),$$

By construction, this space allows for non-linear transformations of a mother filter. An example of such a warping can be visualized in Figure 3.

In the next paragraph, we introduce some filters that can be recovered using this transformation map. For some of these filters, the estimation of their parameter has been investigated (Gribonval, 2001; Wang & Jiang, 2008; Xu et al., 2016), however, our method provides two benefits, first, the generalization which alleviates the need of selecting a specific type of filter bank, second, the scalability of our method leading to a learnable filter bank.

Recovering Standard Filter Banks: The space $C^0_{\text{inc}}(\mathbb{R})$ allows us to span well known transformations. In particular, a filter can inherit a particular chirpyness[1] from nonlinear transformations belonging to $C^0_{\text{inc}}(\mathbb{R})$.

This property is interesting for the decomposition of non-stationary and fast-varying signals. In fact, various signals include such an intricate feature, such as bird song, speech, sonar system (Flandrin, 2001). Among the possible transformations induced on a mother filter by the mapping $g \in C^0_{\text{inc}}(\mathbb{R})$, some of them correspond to well-known filters described in Table 1.

For instance, let's consider the case where $\mathcal{F}$ is the space of linear function with positive slope and defined as $\forall g \in \mathcal{F}, g(t) = \frac{t}{\lambda}$, where λ is positive. In this case, we recover the transformation leading to the dilation or contraction of a wavelet mother filter. The filter bank is then

[1] Chirpyness is defined as the rate of change of the instantaneous frequency of the filter (Mann & Haykin, 1995).

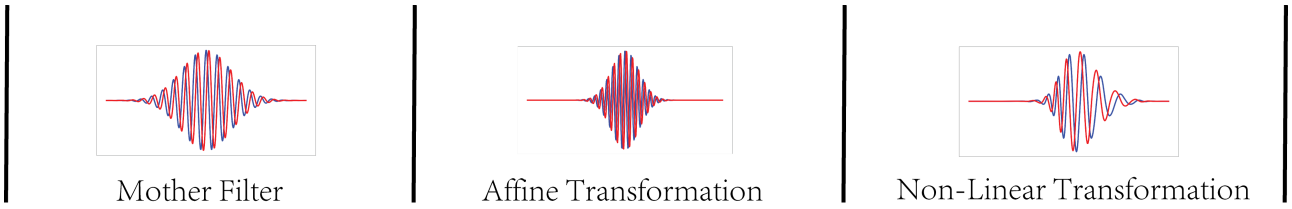

Figure 3. **Transformation of a Morlet Wavelet:** For all the filters, the real part is shown in blue and the imaginary in red. (*left*) Morlet wavelet mother filter. (*middle*) Transformation of the mother filter with respect to an affine transform: the dilation parameter $0 < a < 1$, i.e., contraction, and translation $b = 0$, i.e., no translation. (*right*) Increasing and continuous transformation of the mother filter for some randomly generated function $g \in C^0_{\text{inc}}(\mathbb{R})$ leading to chirplet-like filter.

Table 1. Recovering well-known filters

$g \in C^0_{inc}(\mathbb{R})$	$\psi \circ g$
Affine	Wavelet
Quadratic Convex	Increasing Quadratic Chirplet
Quadratic Concave	Decreasing Quadratic Chirplet
Logarithmic	Logarithmic Chirplet
Exponential	Exponential Chirplet

generated by sampling a few elements of the group. In the case of the dyadic wavelet transform, the dilation parameters follow a geometric progression of common ratio equals to 2, such as $\lambda_k = 2^{(k-1)/Q}, k = 1, ..., K$, where $K = J \times Q$, with J and Q are the number of octaves and wavelets per octave, respectively. The filter bank obtained is $\left\{\psi(\frac{t}{\lambda_1}), \ldots, \psi(\frac{t}{\lambda_K})\right\}$, and the representation of signal is obtained by convolutions between the filter bank elements and the signal. Equivalently, the space $\mathcal{F}$ can be defined as affine, and the WT is achieved by inner products between the filters and the signal.

While the WT filter bank can easily be recovered, our modelization of the filter bank does not allow for elements with a number of oscillations that differ from the mother filter. To enable such a transformation, another function h with a number of oscillations that differs from the mother filter could be multiplied with the mother filter, such that $h \times \psi \circ g$ provides the elements of the filter bank. Therefore, STFT is not part of the representations that such a framework encompasses.

In this work, we also consider the case where the representation of the signal is performed by convolutions. This representation has equivariance properties that are induced by the convolutional operator as well as the space $C^0_{\text{inc}}(\mathbb{R})$.

Equivariance Properties of The Filter Bank: The equivariance-invariance properties of signal representations play a crucial role in the efficiency of the algorithm at hand as they define how some variations in the signal may or may not be captured (Mallat, 2016). These properties can be intuitively explained and analyzed by considering the representation of the signal as a function of group elements. Details regarding the background of group theory and its link with wavelet analysis are provided in Appendix A. Considering the mapping $\rho_{\text{inc}} = \psi \circ g,\ g \in C^0_{\text{inc}}(\mathbb{R})$, as a group action on the space of the mother filter, i.e., $\mathbb{L}_2(\mathbb{R})$, or more precisely, a representation of a group on $\mathbb{L}_2(\mathbb{R})$, we can develop the equivariance properties of the LGT. The proof that ρ_{int} is in fact a representation is given in Appendix D.1. We can consider the set $C^0_{\text{inc}}(\mathbb{R})$ with the operation $\odot$ consisting of the composition of functions to form the group of strictly increasing and continuous maps denoted by $\mathbf{G}_{\text{inc}}$. This formulation eases the derivation of the equivariance properties of group transforms which can be defined for a group $\mathbf{G}$ for all $g, g' \in \mathbf{G}$ by

$$\mathcal{W}[\rho(g')s_i, \psi](g, .) = \mathcal{W}[s_i, \psi]((g')^{-1} \odot g, .).$$

Transforming the signal with respect to the group $\mathbf{G}$ and computing its representation is equal to computing the representation of the signal and then transforming the representation. If $\mathbf{G}$ corresponds to the affine group, the associated group transform is the WT, which is equivariant to scalings and translations. One can already notice that since $\mathcal{W}(.,.)$ employs convolution to decompose the signal, for any group $\mathbf{G}$, the LGT is translation equivariant. We now focus on more specific equivariance properties of the LGT by defining the local equivariance for all $g, g' \in \mathbf{G}$ by

$$\exists \tau \in \mathbb{R}, \mathcal{W}[\rho(g')s_i, \psi](g, \tau) = \mathcal{W}[s_i, \psi]((g')^{-1} \odot g, \tau).$$

That is, the representation of a local transformation of a signal in a window centered at τ is equal to the transformation of the representation at τ. The size of the window depends on the support of the filter. As a matter of fact, assuming that the representation of $\mathbf{G}_{\text{inc}}$ is unitary, we have the following proposition.

Proposition 1. *The LGT is locally equivariant with respect to the action of the group* $\mathbf{G}_{inc}$.

The proof is given in Appendix D.

As we mentioned, a filter bank of K filters is created by sampling the space $C^0_{\text{inc}}(\mathbb{R})$. We now show how this sampling

can be achieved efficiently by proposing a parametrization of functions belonging to such a space.

3.2. Learning the Time Warping

In this work, we are specifically interested in the learnability of such an increasing and continuous map. We provide a way to sample this space via its parameterization. We use piecewise affine functions constrained such that they belong to the class of strictly increasing and continuous functions, which can be efficiently performed by sorting the output of a 1-layer ReLU NN.

Adaptive Knot Implementation: To implement the nonlinear mapping induced by the representation of the piecewise affine group, we use the fact that a piecewise continuous function can be re-written as a 1-layer ReLU Neural Network (Arora et al., 2016; Yarotsky, 2017).

Besides the computational advantages of such relationships and the differentiable property of the weights of the NN, this model is a knot-free piecewise affine mapping, providing more flexibility regarding the warping function. The knot-free mapping implies that instead of having each affine piece of the function with uniform support, it can vary. As such, this flexibility induces better approximation property (Jupp, 1978). Then, the increasing constraint on the mapping is implemented by sorting the output of the NN. This operation has a $\mathcal{O}(n \log n)$ complexity and is applied on the warped time, which is usually of size $\approx 2^9$.

Objective Function and Learning: Let θ_k be the parameters of each increasing piecewise affine map computed by the NN and we denote by g_{θ_k} the sorted outputs of the NN. The LGT filter bank has the following form

$$\{\psi \circ g_{\theta_1}, \dots, \psi \circ g_{\theta_K}\}.$$

Given a set of signals $\{s_i \in \mathbb{L}_2(\mathbb{R})\}_{i=1}^N$ and given a task specific loss function L, we aim at solving the following optimization problem

$$\min_{\Theta} \sum_{i=1}^{N} L\big(F(\mathcal{W}[s_i, \psi](\mathbf{g}_\Theta, .))\big),$$

where $\Theta = (\theta_1, \dots, \theta_K)$, N denotes the number of signals, K the number of filters, F represents a DNN, and we recall that

$$\mathcal{W}[s_i, \psi](\mathbf{g}_\Theta, .) = [\mathcal{W}[s_i, \psi](g_{\theta_1}, .), \dots, \mathcal{W}[s_i, \psi](g_{\theta_K}, .)]^T.$$

Since, the g_{θ_k} are computed by sorting the output of the NN and the parameters can be learned by a gradient descent optimization jointly with the parameters of F.

Model Constraints to Reduce Aliasing: The nonlinearity of the transformation might reduce the localization of the filter in the frequency domain, and produce aliasing. For some applications, the localization of each filter in the frequency domain is crucial, e.g., the bird detection task in Section 4.2.

In order to limit the possible aliasing induced by the piecewise increasing mappings applied to the mother filter, we propose different settings. Besides, these constraints also impact the type of filter bank our method can reach.

First, we propose a normalization of the frequency of the transform filter (denoted in the result tables by nLGT). This normalization helps to reduce the aliasing induced by the filters. We propose to use $\hat{f}$, the normalized frequency f with respect to the maximum slope of the piecewise affine mapping. For instance, in the case of a Morlet wavelet, the normalization is as follows

$$(\psi \circ g_\theta)(t) = \pi^{-\frac{1}{4}} \exp\Big(2\pi j \hat{f} g_\theta(t)\Big) \exp\big(-\tfrac{1}{2}(g_\theta(t)/\sigma)^2\big),$$

where $\hat{f} = f / \max_{l \in \{1,\dots,n\}} a_l$, where n denotes the number of pieces of the piecewise map, and a_l the slope of each piece, j is the imaginary unit, and σ is the width parameter defining the localization of the wavelet in time and frequency. This normalization will be performed for each sample of the group, and thus for each generated filter $k \in \{1, \dots, K\}$ of the filter bank.

Second, we constrain the domain of the piecewise affine map (denoted in the result tables by cLGT). In the following experiments, we propose a dyadic constraint of the domain as in the WT. The support of the filter is close to the support of a wavelet filter bank. However, the envelope of the filter and the instantaneous frequency still has a learned chirpyness.

4. Experiments

For all the experiments and all the settings, i.e., LGT, nLGT, cLGT, cnLGT, the increasing and continuous piecewise affine map is initialized randomly, and the optimization is performed with Adam Optimizer, and the number of knots of each piecewise affine map is 256. The mother filter used for our setting is a Morlet wavelet filter. The code of the LGT framework is provided in the following repository https://github.com/Koldh/LearnableGroupTransform-TimeSeries.

4.1. Artificial Data: Classification of Chirp Signals

We present an artificial dataset that demonstrates how a specific time-frequency tiling might not be adapted or would require cross-validations for a given task and data. To build the dataset, we generate one high frequency ascending chirp and

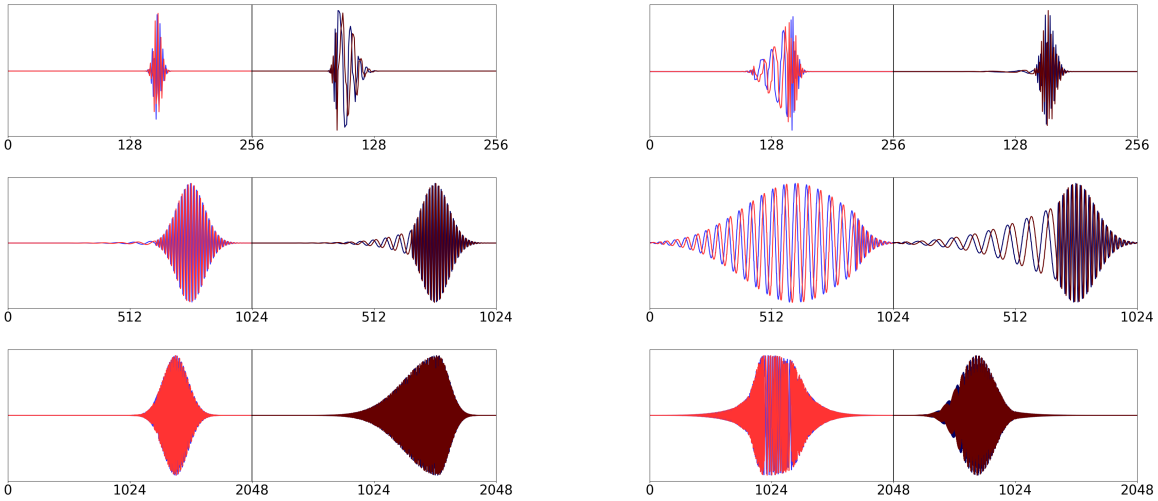

Figure 4. **Learnable Group Transform Filters** for the Artificial Data - Each row displays two selected filters (left and right sub-figure) for different settings: (*from top to bottom*) nLGT, cLGT, cnLGT. For each subfigure, the left part corresponds to the filter before training and the right part to the filter after training. The blue and red denote the real and imaginary parts of the filters, respectively.

Table 2. Testing Accuracy for the Chirp Signals Classification Task

Representation + Non-Linearity + Linear Classifier	*Accuracy*
Wavelet Transform (64 Filters)	53.01 ± 5.1
Short-Time Fourier Transform (64 Filters)	65.1 ± 11.9
Short-Time Fourier Transform (128 Filters)	86.6 ± 9.8
Short-Time Fourier Transform (512 Filters)	$\mathbf{100 \pm 0.0}$
LGT (64 Filters)	92.9 ± 4.0
nLGT (64 Filters)	95.7 ± 3.3
cLGT (64 Filters)	56.8 ± 1.6
cnLGT (64 Filters)	$\mathbf{100.0 \pm 0.0}$

one descending high-frequency chirp of size 8192 following the chirplet formula provided in (Baraniuk & Jones, 1996). Then for both chirp signals, we add Gaussian noise samples (100 times for each class), see Figures in Appendix C.1. The task aims at being able to detect whether the chirp is ascending or descending. Both the training and test sets are composed of 50 instances of each class. For all models, set the batch size to 10, the number of epochs to 50. Each experiment was repeated 5 times with randomly sampled train and test set, and the accuracy was the result of the average over these 5 runs. Each GT is composed with a complex modulus, and the inference is performed by a linear classifier. For the case of WT and LGT, the size of the filters is 512.

As we can observe in Table 2, the WT, as well as the STFT with few numbers of filters, perform poorly on this dataset. The chirp signals to be analyzed are localized close to the Nyquist frequency, and in the case of WT, as illustrated in Figure 1, the wavelet filter bank has a poor frequency resolution in high frequency while benefiting from a high time resolution. In this experiment, we can see that this characteristic the WT time-frequency tiling implies that through time, the small frequency variations of the chirp are not efficiently captured. In the case of STFT, as the number of filters decreases, the frequency resolution is altered. Thus, this frequency variation is not captured. Using a large window for the STFT increases the frequency resolution of the tiling and thus enables to capture the difference between the two classes. In the LGT setting, the tiling has adapted to the task and produces good performances except for the cLGT model. In fact, the domain of the piecewise linear map is constrained to be dyadic, and thus the adaptivity of the filter bank is reduced, which is not suitable for this specific task.

Some of the filters can be visualized in Figure 4, as well as the representations of the signals in Appendix C.1.2. This experiment shows an example of signals that are not easily classified by neither the proportional-bandwidth nor the constant-bandwidth without considering cross-validation of hyperparameters.

4.2. Supervised Bird Detection

Table 3. Testing AUC for the Bird Detection Task

Representation + Non-Linearity + Deep Network	*AUC*
MFSC (80 Filters)	77.83 ± 1.34
Conv. Filter init. random (80 Filters)	66.77 ± 1.04
Conv. Filter init. Gabor (80 Filters)	67.67 ± 0.98
Spline Conv. init. random (80 Filters) (Balestriero et al., 2018)	78.17 ± 1.48
Spline Conv. init. Gabor (80 Filters) (Balestriero et al., 2018)	79.32 ± 1.52
LGT (80 Filters)	78.41 ± 1.38
nLGT (80 Filters)	75.50 ± 1.39
cLGT (80 Filters)	79.14 ± 0.83
cnLGT (80 Filters)	$\mathbf{79.68 \pm 1.35}$

We now propose a large scale dataset to validate the suitabil-

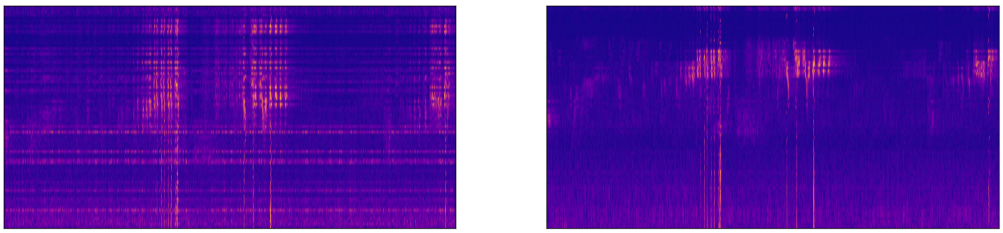

Figure 5. **Learnable Group Transform** - Visualization of a sample containing a bird song (cLGT), where (*left*) at the initialization and (*right*) after learning. For each subfigure, the x-axis corresponds to time and the y-axis to the different filters. Notice that the y-axis usually corresponds to the scale or the center-frequency of the filters. Other representations are displayed in Appendix C.2.2. We can observe that compared to the initialization, the learned representation is sparser and the SNR is increased. Besides, the representation is less redundant in the frequency axis.

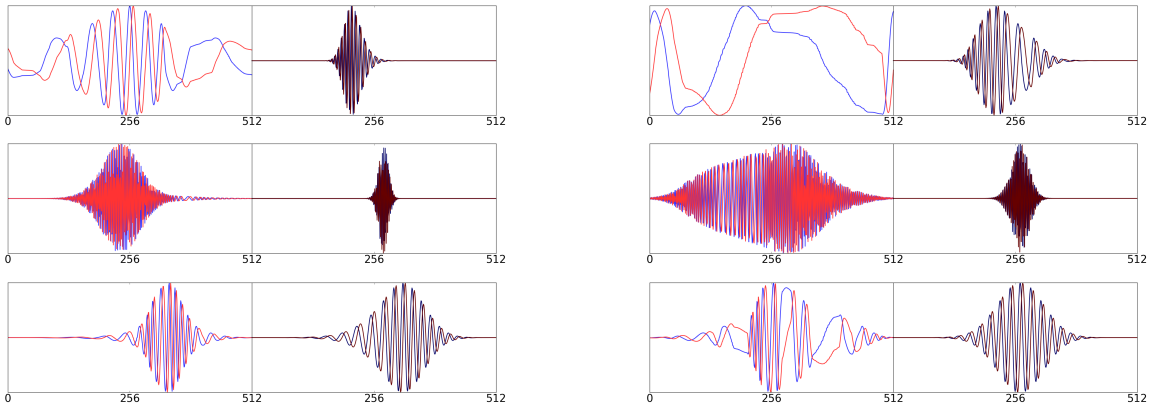

Figure 6. **Learnable Group Transform Filters** for the Bird Detection Data - Each row displays two selected filters (left and right sub-figure) for different settings: (*from top to bottom*) LGT, nLGT, cLGT. For each subfigure, the left part corresponds to the filter before training and the right part to the filter after training. The blue and red denote the real and imaginary parts of the filters, respectively.

ity of our model in a noisy and realistic setting. The dataset is extracted from the Freesound audio archive (Stowell & Plumbley, 2013). This dataset contains about $7,000$ field recording signals of 10 seconds sampled at 44 kHz, representing slightly less than 20 hours of audio signals. The content of these recordings varies from water sounds to city noises. Among these signals, some contain bird songs that are mixed with different background sounds having more energy than the bird song. A visualisation of a sample is shown in Appendix C.2.1. The given task is a binary classification where one should predict the presence or absence of a bird song. As the dataset is unbalanced, we use the Area Under Curve (AUC) metric. The results we propose for both the benchmarks and our models are evaluated on a test set consisting of 33% of the total dataset.

In order to compare with previously used methods, we use the same seeds to sample the train and test set, the batch size, i.e., 10, and the learning rate cross-validation grid as in (Balestriero et al., 2018). For each model, the best hyperparameters are selected, and we train and evaluated randomly 10-times the models with early stopping, the results are shown in Table 3. While the first layer of the architecture has a model-dependent representation (i.e., MFSC, LGT, Conv. filters,...), we use the state-of-the-art architecture (Grill & Schlüter, 2017) for the DNN architecture, described in Appendix B.2. Notice that this specific DNN architecture has been designed and optimized for MFSC representation.

As we can see in Table 3, the case without constraints (LGT) reaches better accuracy than the domain expert benchmark (MFSC). Besides, including more constraints on the model (cnLGT) reduces overfitting and further improve results to outperform the other benchmarks. One can also remark that both the LGT framework and learnable mother wavelet reach almost the same accuracy, while they both outperform the hand-crafted feature as well as the unconstrained convolutional filters. One can notice that all the learned filters in Figure 6 contain either an increasing chirp or a decreasing chirp, corresponding respectively to the convexity or concavity of the instantaneous phase of the filter and thus of the piecewise linear map. Such a feature is being used and is

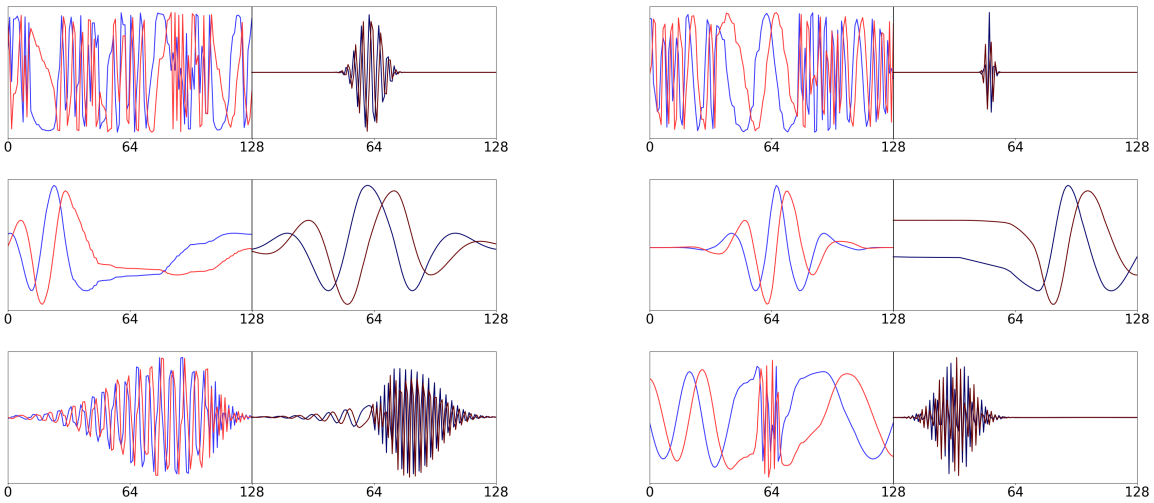

Figure 7. **Learnable Group Transform Filters** for the Haptics Data - Each row displays two selected filters (left and right sub-figure) for different settings: (*from top to bottom*) nLGT, cLGT, cnLGT. For each subfigure, the left part corresponds to the filter before training and the right part to the filter after training. The blue and red denote the real and imaginary parts of the filters, respectively.

crucial in the detection and analysis of bird song (Stowell & Plumbley, 2012).

4.3. Haptics Dataset Classification

Table 4. Testing Accuracy for the Haptics Classification Task

Model	*Accuracy*
DTW (Al-Naymat et al., 2009)	37.7
BOSS (Schäfer, 2015)	46.4
Residual NN (Wang et al., 2017)	50.5
COTE ((Bagnall et al., 2015)	51.2
Fully Convolutional NN (Wang et al., 2017)	55.1
WD + Convolutional NN (Khan & Yener, 2018)	57.5
LGT (96 Filters)+ Non-Linearity + Linear Classifier	53.5
nLGT (96 Filters)+ Non-Linearity + Linear Classifier	50.4
cLGT (96 Filters)+ Non-Linearity + Linear Classifier	**58.2**
cnLGT (96 Filters)+ Non-Linearity + Linear Classifier	54.3

The Haptics dataset is a classification problem with five classes and 155 training and 308 testing samples from the UCR Time Series Repository (Chen et al., 2015), where each time-series has 1092 time samples. As opposed to the bird dataset where features of interests are known, and competitive methods have been established, there is no expert knowledge regarding the specific signal features (see Table 4). One can see that our method outperforms other approaches in the cLGT setting while performing the classification with a linear classifier as opposed to other methods using DNN algorithms. This demonstrates the capability of our method to transform the data efficiently while not requiring a further change of basis as well as knowledge on the features of interests. Besides, even in a small dataset regime, our approach is capable of learning an efficient transformation of the data.

We provide in Figure 7 the visualization of some sampled filters before and after learning as well as representations in Appendix C.3.2. As opposed to the supervised bird dataset, we can see that the filters do not coincide with well-known filters that are commonly used in signal processing. This is an example of an application where the features of interest in the signals are unknown, and one requires a learnable representation.

5. Conclusion

In this work, we enable the learnability of Group Transform and generalize the wavelet transform by introducing non-linear transformations of a mother filter as well as an efficient way to sample this mapping. We establish the connections with well-known time-frequency filters that are common in diverse biological signals as well as the derivation of the equivariance properties of the LGT. Also, we have shown a tractable way to learn to sample these transformations using a 1-layer NN enabling an end-to-end approach. Our approach competes with state-of-the-art methods without a priori knowledge on the signal power spectrum and outperforms classical hand-crafted time-frequency representations. Interestingly, in the bird detection experiment, we recover chirplet filters that are known to be crucial to their detection, while in the case of the haptic dataset where important features to be captured to perform the classification of the signals are unknown, the filters learned are very dissimilar to classical time-frequency filters and allow to outperform state-of-the-art methods with a linear classifier.

Acknowledgment

The authors would like to thank Randall Balestriero, Yanis Bahroun, and Anirvan M. Sengupta for their qualitative comments and reviews. The authors were supported by NSF grant SCH-1838873 and NIH grant R01HL144683-CFDA.

References

Al-Naymat, G., Chawla, S., and Taheri, J. Sparsedtw: A novel approach to speed up dynamic time warping. In *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, pp. 117–127. Australian Computer Society, Inc., 2009.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Bagnall, A., Lines, J., Hills, J., and Bostrom, A. Time-series classification with cote: the collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, 2015.

Balestriero, R., Cosentino, R., Glotin, H., and Baraniuk, R. Spline filters for end-to-end deep learning. In *International Conference on Machine Learning*, pp. 373–382, 2018.

Baraniuk, R. G. Shear madness: signal-dependent and metaplectic time-frequency representations. 1993.

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

Bruna, J. *Scattering representations for recognition*. PhD thesis, 2013.

Cakir, E., Ozan, E. C., and Virtanen, T. Filterbank learning for deep neural network based polyphonic sound event detection. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 3399–3406. IEEE, 2016.

Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999, 2016.

Coifman, R. R. and Wickerhauser, M. V. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2):713–718, 1992.

Cosentino, R., Balestriero, R., Baraniuk, R. G., and Patel, A. Overcomplete frame thresholding for acoustic scene analysis. *arXiv preprint arXiv:1712.09117*, 2017.

Daubechies, I. *Ten Lectures on Wavelets*, volume 61. Siam, 1992.

Feichtinger, H. G., Kozek, W., and Luef, F. Gabor analysis over finite abelian groups. *Applied and Computational Harmonic Analysis*, 26(2):230–248, 2009.

Flandrin, P. Time frequency and chirps. In *Wavelet Applications VIII*, volume 4391, pp. 161–175. International Society for Optics and Photonics, 2001.

Goldenstein, S. and Gomes, J. Time warping of audio signals. In *cgi*, pp. 52. IEEE, 1999.

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

Gribonval, R. and Bacry, E. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, 2003.

Grill, T. and Schlüter, J. Two convolutional neural networks for bird detection in audio signals. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, August 2017. URL http://ofai.at/~jan.schlueter/pubs/2017_eusipco.pdf.

Jupp, D. L. B. Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, 1978.

Kerkyacharian, G., Picard, D., et al. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.

Khan, H. and Yener, B. Learning filter widths of spectral decompositions with wavelets. In *Advances in Neural Information Processing Systems*, pp. 4601–4612, 2018.

Korda, M. and Mezić, I. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93:149–160, 2018.

Le, T.-P. and Argoul, P. Continuous wavelet transform for modal identification using free decay response. *Journal of sound and vibration*, 277(1-2):73–100, 2004.

Lelandais, F. and Glotin, H. Mallat's matching pursuit of sperm whale clicks in real-time using daubechies 15 wavelets. In *New Trends for Environmental Monitoring Using Passive Systems, 2008*, pp. 1–5. IEEE, 2008.

Mallat, S. *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

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

Mallat, S. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

Mann, S. and Haykin, S. The chirplet transform: Physical considerations. *IEEE Transactions on Signal Processing*, 43(11):2745–2761, 1995.

Oyallon, E., Zagoruyko, S., Huang, G., Komodakis, N., Lacoste-Julien, S., Blaschko, M. B., and Belilovsky, E. Scattering networks for hybrid representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

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

Saritha, C., Sukanya, V., and Murthy, Y. N. Ecg signal analysis using wavelet transforms. *Bulg. J. Phys*, 35(1): 68–77, 2008.

Schäfer, P. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.

Seydoux, L., Shapiro, N. M., de Rosny, J., Brenguier, F., and Landès, M. Detecting seismic activity with a covariance matrix analysis of data recorded on seismic arrays. *Geophysical Journal International*, 204(3):1430–1442, 2016.

Stowell, D. and Plumbley, M. D. Framewise heterodyne chirp analysis of birdsong. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2694–2698. IEEE, 2012.

Stowell, D. and Plumbley, M. D. An open dataset for research on audio field recording archives: freefield1010. *CoRR*, abs/1309.5275, 2013. URL http://arxiv.org/abs/1309.5275.

Torrésani, B. Wavelets associated with representations of the affine weyl–heisenberg group. *Journal of Mathematical Physics*, 32(5):1273–1279, 1991.

Vilenkin, N. Y. *Special Functions and the Theory of Group Representations*, volume 22. American Mathematical Soc., 1978.

Wang, Y. and Jiang, Y.-C. Modified adaptive chirplet decomposition with application in isar imaging of maneuvering targets. *EURASIP Journal on Advances in Signal Processing*, 2008(1):456598, 2008.

Wang, Z., Yan, W., and Oates, T. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 international joint conference on neural networks (IJCNN)*, pp. 1578–1585. IEEE, 2017.

Xu, C., Wang, C., and Gao, J. Instantaneous frequency identification using adaptive linear chirplet transform and matching pursuit. *Shock and Vibration*, 2016, 2016.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Zeghidour, N., Usunier, N., Kokkinos, I., Schaiz, T., Synnaeve, G., and Dupoux, E. Learning filterbanks from raw speech for phone recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5509–5513. IEEE, 2018.