Extraction of Rhythmical Features with the Gabor Scattering Transform

Daniel Haider, Peter Balazs

Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Vienna, Austria.

Abstract. In this paper we use the scattering transform to extract wide-scale information of musical pieces in terms of rhythmical features. This transform computes a layered structure, similar to a convolutional neural network (CNN) but with no learning involved. Applied to audio it is able to capture temporal dependencies beyond those possible for common time-frequency representations. This is already demonstrated by experiments for modulations of single tones. Here we provide a setup to include real world music signals, which extends the temporal range to the scale where rhythm and tempo live, allowing very intuitive explanations of how these scales are reached. In this way we also get an intuition of the mechanics inside a neural network when "listening" to rhythm.

Keywords: Gabor Transform, Scattering Transform, Rhythm, Tempo, Time-Frequency Representations, Convolutional Neural Networks

1 Introduction

A great goal in computer sciences these days seems to be making a computer to "listen" like humans do [1]. Among many machine learning approaches which have been aiming towards this, *Convolutional Neural Networks* (CNNs) perform particularly well, archieving impressive results in various audio-related learning tasks [2]. When such a network is tasked to learn from audio, it needs to identify and interpret patterns on several temporal scales: Pitch and timbre live within the scale of milliseconds, tempo and rhythmical structures are spread over periods of seconds and general progressions pass minutes and hours. To get a feeling of how this can be archieved, we briefly introduce CNNs and discuss the responsible mechanics. This motivates a special transform of audio signals, called *Scattering Transform*, which is computed by a cascade of time-frequency transforms setting up a layered network, similar to the structure of a CNN [5]. We emphasize on its ability to capture wide range dependencies of an audio signal and setup a framework to extract rhytmical structures of a musical piece in terms of its tempo. For this we use a scattering procedure based on a sampled Short-Time Fourier Transform (STFT), also called *Gabor Scattering* [9], which allows to illustrate particularly well, how coarser structures are captured and depicted in a very intuitive way. We may paste these insights back to the CNN-case as a deterministic analogue.

2 Computers Listening - Convolutional Neural Networks

The idea of a neural network is to setup a function, that can theoretically approximate any other function via an optimization procedure. Such a network has a layered structure, each consisting of single neurons which filter the importance of the information arriving and a non-linear function, called *activation function* that controls the "significance" of the neuron to the network. *Convolutional Neural Networks* are a specialized form of DNNs to deal with grid-like data, originally introduced for image processing problems, [2]. The idea is to convolve the input matrix with 2D filters that are much smaller than the input dimension, which can be interpreted as localization of certain properties of the data.
To obtain a feature extraction procedure, the dimensionality of the input data has to be reduced. This can be archieved by *pooling*. Pooling computes a "summary" of nearby elements, so it decreases the dimensionality and generates invariances to specific deformations and variations in the data. Furthermore, this expands the range of the filters in the subsequent layers since the filters are applied to the pooled "summary"-elements, that are representative for a whole neighborhood of elements of the previous layer. Thus, the deeper the network gets, the wider dependencies are captured.
CNNs have led to an immense progress in image-related learning tasks. Clearly it makes totally sense to apply it also on the images obtained by time-frequency decompositions of audio signals [3]. Those decompositions already provide a representation of basic features of a signal, namely its frequency content, which is referred to as pitch, i.e. small-scale information. In the next section we will have a closer look at the scattering transform, a similar but non-learning construction based on cascades of time-frequency representations.

3 Computers Listening Revisited - The Scattering Transform

Most of the time-frequency representations used are set up via a filterbank construction, i.e. a collection of filters, that decompose the signal into different frequency bins [4]. The filtering can be realized via the convolution operation; we write the discrete version as $(x * w_k)[m] = \sum_n x[n] w_k[m-n]$, where x denotes the signal and w_k the k-th filter. Usually, a modulus $|\,.\,|$ or a modulus squared $|\,.\,|^2$ is applied on the computed coefficients. Furthermore, some transformations also use dimensionality reduction in time, e.g. subsampling or scaling. This construction, i.e. filtering, taking the modulus and dimensionality redution points out the fundamental link between filterbank decompositions and the principle structure of a CNN.
The Scattering Transform extends this link by computing a layered network structure, based on time-frequency decompositions with a modulus. In other words, time-frequency magnitude decompositions are applied on the single filter outputs of the previously decomposed filter outputs, see Figure 1.

Definition 1 (Scattering). *We compute a member of the ℓ-th layer L_ℓ of the scattering network by following a path $p = (p_1, \ldots, p_\ell)$ through its tree-like structure. Denoting the k-th filter in the ℓ-th layer as w_k^ℓ, then p denotes the indices of the used filters per layer, i.e. $w_{p_1}^1, \ldots, w_{p_\ell}^2$. So we can define,*

$$L_\ell[p] = ||\ldots|x * w_{p_1}^1| * \ldots| * w_{p_\ell}^\ell| \tag{1}$$

Originally, the transform was introduced by Mallat in [5] and was based on the wavelet transform. It came with a rigorous mathematical analysis, enabling it to show some translation invariance and stability w.r.t. time-deformations. Later the approach was generalized to semi-discrete frames, which include common filterbank constructions [6]. In [7] it was used as a feature extractor for audio and it turned out that it is able to represent temporal features beyond those possible for common time-frequency representations. Based on the pitch and filter structures contained in the first layer, the second layer reveals transient phenomena, such as note attacks and modulation of the amplitude and frequency. The examples there show that indeed, as we look at deeper layers of the scattering network, wider structures of the signal are represented. Applied on an amplitude modulated tone it is the envelope that appeared, i.e. the timbre of this particular sound and in a vibrato modulated tone it is the frequency of the vibrato pulse. In the next chapter we set up a framework to analyze rhythmical aspects of a musical signal and show that also these coarser temporal structures can be captured in the second layer of a scattering network.

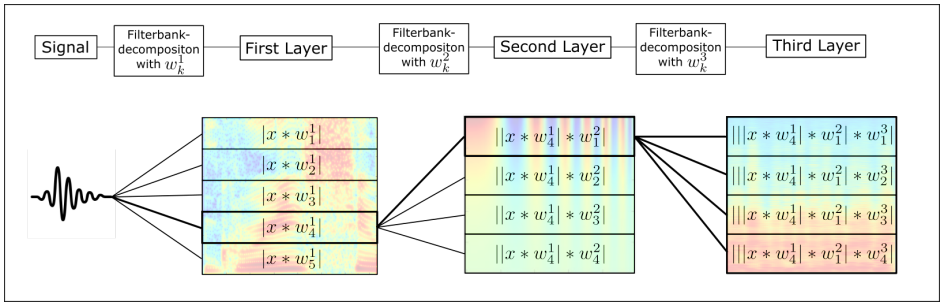

Fig. 1. The figure illustrates the scattering procedure by computing a third layer w.r.t. the path $[4, 1]$.

4 Rhythmical Feature Analysis

Rhythm refers to the timing of events within a musical piece and has different levels of periodicity. In the notation of western music, a hierachial metrical structure is used to distinguish between different time scales. The *Tatum* (temporal atom) is the smallest and is related to the shortest durational value encountered between two events within the musical piece. The *Tactus* (beat) is the perceptually most prominent and refers to the rate, most people would "tap" their feet to. Finally the widest, the *Bar* (measure) is related to the length of a rhythmical

pattern [8]. In a classical drumbeat in 4/4 with a bass drum on the 1 & 3, a snare on the 2 & 4 and a eighth note hihatpattern, Tatum would correspond to the hihat, Tactum to the snare and bass drum and the Bar to the whole pattern. The *tempo* of a musical piece is referred to the speed of the most prominent rhythm pattern, which is usually the Tactus. Tempo is measured in bpm (beats per minute) and reaches among different genres and styles of music from 30-300bpm. Embodying the tempo as a periodic pattern of events in time we could also assign a frequency to it; here in the range of 0.5-5Hz. As a frequency, this is clearly not audible, but indeed perceivable as a rhythmical pattern. The scattering transform is also capable of "perceiveing" a rhythmical pattern by depicting its (subsampled) frequency in the second layer. We explain this in particular, demonstrated on the most simple embodiment of tempo, a *metronome*.

4.1 Gabor-Scattering of a Metronome

We use a scattering transform based on a sampled Short-Time Fourier Transform (STFT) with a time-hop size parameter α, a window ϕ and a subsequent modulus operation applied. This is also known as *Gabor-Scattering* [9]. The following equation defines the sampled STFT applied on a signal x and shows, how it can be interpreted as filterbank decomposition w.r.t. filters w_k.

$$X_k[n] := \sum_m x[m] \underbrace{e^{-i\omega_k m}\phi[m-\alpha n]}_{=:w_k[\alpha n - m]} = (x * w_k)[\alpha n]. \tag{2}$$

We can model a metronome simply by an impulse train with periodicity T, i.e. a frequency at $1/T$,

$$e[n] = \sum_m \delta(n - mT),\ T \geq 0, \tag{3}$$

see Figure 2(a). We perform a STFT on e using a window, that is smaller than T to avoid overlapping. After applying a pointwise modulus the first layer can be written as,

$$L_1[k][n] = \sum_m |\phi(n\alpha - mT)| \tag{4}$$

for all w_k. This can be explained by viewing the convolution in Eqn. (2) as moving the filters w_k along single peaks of ones, which would simply give $\sum_m |w_k(n\alpha - mT)|$. The (pointwise!) modulus then removes the modulation part in w_k and only a sum of shifted windows remains. Eqn. (4) shows a smoothed and subsampled version of the metronome with a periodicity of T/α instead of T, i.e. a frequency of α/T, instead of $1/T$, see Figure 2(b). The idea is now to choose α sufficiently large, such that α/T can be captured and depicted by another STFT, the "conventional" way. The second layer will thus show a constant frequency of α/T, see Figure 2(d). By sampling the time-frequency representation in wide time-steps we obtain downscaled signals in the filterbins, which will be interpreted as having a higher frequency by subsequent filters. This enables to capture wider scales; the larger α is chosen, the wider the scale in focus.

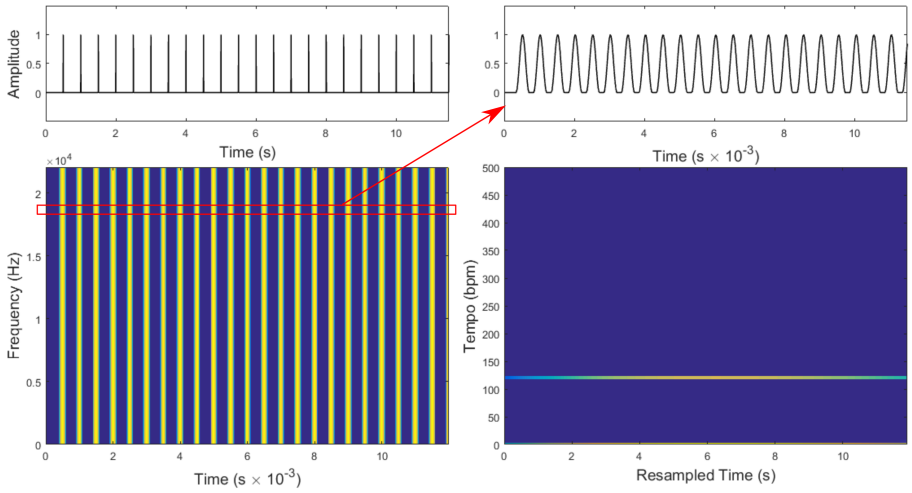

Fig. 2. Left: impulse train in the tempo of 120bpm, i.e. 2Hz and below its STFT, using a hop size of $\alpha_1 = 10^3$. Note that this is plotted here using a scaled timeaxis in seconds$\times 10^{-3}$! Right: the single filter outputs: the impulse train is smoothed by the window and scaled in time by the factor α, i.e. it is shorter and has a higher frequency, 2000Hz. Below is then its STFT with $\alpha_2 = 1$ on a resampled timescale and a "tempo"scale measured in bpm. It clearly depicts the tempo of the original signal train, 120bpm.

If we want to consider more complex musical signals, we may see them as consisting of tonal, transient and stochastic components. Truly, the transient parts of a signal indicate its rhytmical structure, e.g. by percussive elements like drums, note-onsets, etc. In that manner, the scattering transform will detect periodic patterns among the transient arrangements and extract the temporal information with respect to those. The measured level will depend on the most salient transient patterns in the signal. We set up three musical situations, where different levels of tempo are captured:

(a) A recording of a melody, played on a guitar in fingerpicking style. The melody consists of consecutive eighth notes, played rather monotonically, therefore the rhythmical structure, indicated by the onset-transients of the single notes yields a periodic pattern. Using a Tactus of 120bpm (referring to a quarter note beat), the Tatum of the melody is 240bpm. As the pattern is played rather monotonically, the most salient pattern is the one, induced by every note onset, i.e. 240bpm. Thus, this is the tempo, the second layer computes the highest values for, Figure 4(a) Right.
(b) A recording of chords, played on a guitar with hard palm-mute strokes. The chords are played in consecutive quarter notes in 150bpm (Tactus) with a strong accentuation on every fifth stroke. The rhythmical structure, indicated by the strokes has two levels now: on one side the Tactus of the quarter notes

in 150bpm and on the other side the Bar from the accentuations in 30bpm. The strong accentuation makes both tempo levels accessible in the second layer of the transform, Figure 4(b) Right.

(c) An excerpt (2:45-3:20) of the song "Money" by Pink Floyd. It covers the transition from the saxophone solo (until 3:03) in around 120bpm into David Gilmour's guitar solo (from 3:07) in around 126bpm. The rhytmical structure is indicated by the drums, see Figure 3 below for the transcription. The hihat patterns usually defines the Tatum. In the saxophone solo, it consists of quarter notes over a 7/4 meter and since no finer scaled rhythm elements are present, the Tatum is 120bpm. In the guitar solo the pattern changes to triplets over a 4/4, i.e. the Tatum paces up to 378bpm. These values are clearly depicted in Figure 4(c) Right. The perceived tempo, the Tactus, is usually defined by the bassdrum/snare pattern. In the saxophone solo, Nick Mason plays the snare on the two, four and six with bassdrums inbetween, which yields an aperiodic pattern Bar-wise. Thus, the transform struggles to detect something meaningful for this level. In the guitar solo the pattern becomes periodic with 126bpm.

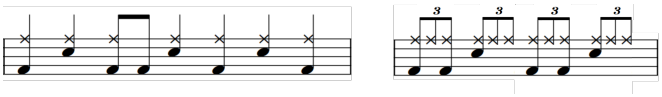

Fig. 3. Left: the drumpattern in the saxophone solo. Right: the drumpattern in the guitar solo. Crosses represent a closed hihat, F-notes the bass drum and C'-notes the snare.

We used time-hop sizes of 1000 samples for computing the first layers in (a),(b) and 2000 for (c). Different values amplify the access to different tempo levels, which makes it possible to isolate certain levels as well by choosing the parameter appropiately. For the second layer, picking a single channel output can be problematic since we may catch one with a lot of tonal material, disturbing the detection of the transient pattern. As a preliminary solution we here computed the second layer using an average over channels corresponding to frequency regions where transients are most dominant. Of course, this also has an impact on the strenghts of certain tempo levels depending on the frequency distribution of the single transient sound. We used the channels corresponding to $1800-10300$Hz in (a),(b) and $1300-3500$Hz in (c).

The proposed procedure was not intended to be a tempo estimator as such, like [10] or [11]. The idea of this simple scattering procedure is to extract temporal structures on several levels in their natural appearance. Other than common tempo estimators, salience is incorporated in the output as well, based on the energy of the transient sounds, which can be beneficial for further processing e.g. rhythm related tasks. Furthermore, it does not depend on external processing like beat tracking algorithms or a learning procedure, which makes it a simple, powerful tool.

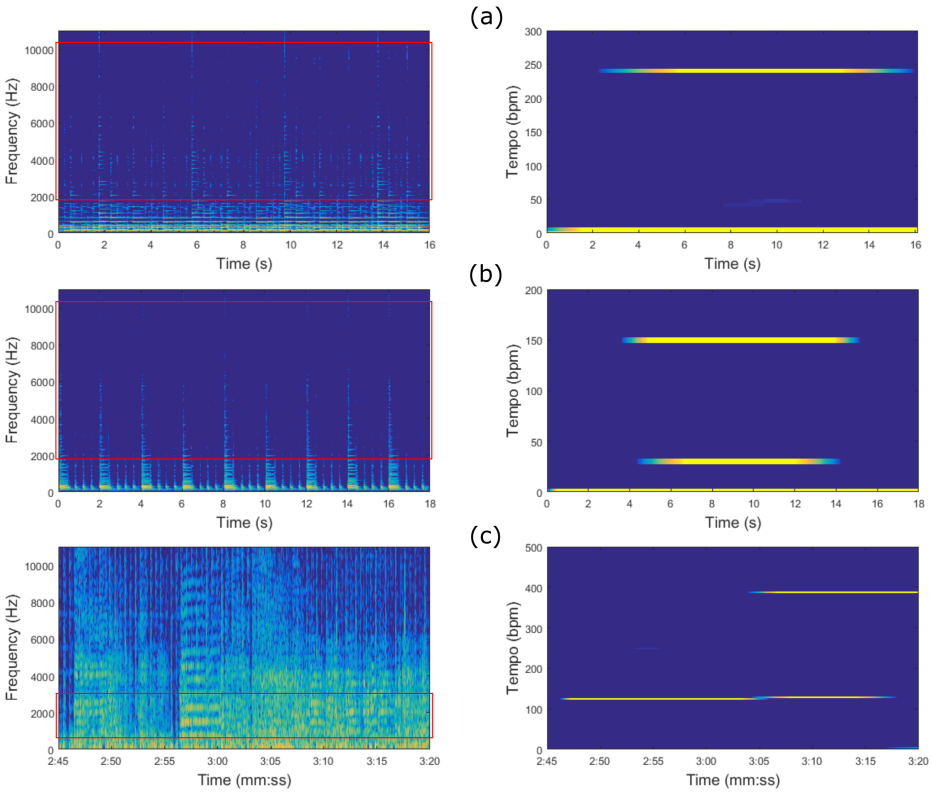

Fig. 4. Left: subsampled STFTs of the signals w.r.t. different (heuristically chosen) hop sizes α_1 and a modulus, plotted in seconds and Hz. The red areas indicate the channels used to compute the second layer. Right: common STFTs ($\alpha_2 = 1$) of the channel averages with modulus and averaging in time (second layer), plotted in seconds (referring to the original signal) and bpm.

5 Conclusion

We presented an intuitive motivation of the scattering transform by extending the concept of time-frequency decompositions in a natural way. With a scattering procedure based on a sampled STFT we ilucidated the mechanics behind the expansion of the captured temporal scales, explained on a simple impulse train. Examples in Figure 4 covered different musical situations with several tempo levels present, depicted in an intuitive way. As it is presented here, it is very preliminary but with potential to be expandable, fine-tuned and easily incorporated in other approaches. What the insights into the network-like structure of the scattering transform provides, can also give an intuition on the mechanics inside a CNN, when trying to learn the rhythmical structures of a musical piece.

Acknowledgements

The work on this paper was partially supported by the Austrian Science Fund (FWF) START-project FLAME (Frames and Linear Operators for Acoustical Modeling and Parameter Estimation; Y 551-N13). The authors thank Nicki Holighaus and Andrés Marafioti for fruitful discussions.

References

1. R. F. Lyon. Machine Hearing: An Emerging Field. *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131-139, 2010.
2. I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*, MIT Press, `http://www.deeplearningbook.org`, 2016.
3. J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann and X. Serra. End-To-End Learning for Music Audio Tagging at Scale. *Proc. of the 19th ISMIR Conference*, Paris, France. September 23-27, 2018.
4. P. Balazs, N. Holighaus, T. Necciari, D. T. Stoeva. Frame Theory for Signal Prcoessing in Psychoacoustics in: R. Balan, J. J. Benedetto, W. Czaja, M. Dellatorre, K. A. Okoudjou (eds.), *Excursions in Harmonic Analysis Vol. 5.*. Basel (Springer), pp. 225-268, 2017.
5. S. Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331-1398, 2012.
6. T. Wiatowski and H. Bölcskei. Deep Neural Networks Based on Semi-Discrete Frames. *Proc. of IEEE International Symposium on Information Theory(ISIT)*, Hong Kong, China. June 14-19, 2015
7. J. Andén and S. Mallat. Scattering Representation of Modulated Sounds. *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-12)*, York, UK. September 17-21, 2012.
8. A. Klapuri, M. Davy. *Signal Processing Methods for Music Transcription*, Springer, 2007.
9. R. Bammer and M. Dörfler. Invariance and Stability of Gabor Scattering for Music Signals. *Proc. of Sampling Theory and Applications (Sampta)*, Tallin, Estonia. July 3-7, 2017.
10. H. Schreiber and M. Müller. A Single-Step Approach to Musical Tempo Estimation Using a Convolutional Neural Network. *Proc. of the 19th ISMIR Conference*, Paris, France. September 23-27, 2018.
11. S. Dixon. Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research*, vol. 30, no. 1, pp. 39-58, 2001.
12. P. Søndergaard, B. Torrésani, P. Balazs. The Linear Time Frequency Analysis Toolbox. *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 4, 1250032, 2012.