1 Abstract

In this paper a feature extractor based on Gabor frames and Mallat’s scattering transform, called Gabor scattering, is introduced. This feature extractor is applied to a simple signal model for audio signals, i.e. a class of tones consisting of fundamental frequency and its multiples and an according envelope. Within different layers, different invariances to certain signal features occur. In this paper we give a mathematical explanation for the first and the second layer which are illustrated by numerical examples. Deformation stability of this feature extractor will be shown by using a decoupling technique, previously suggested for the scattering transform of Cartoon functions. Here it is used to see if the feature extractor is robust to changes in spectral shape and frequency modulation.

2 Introduction

Popular machine learning techniques such as (deep) convolutional networks (CNN) have led to impressive results in several applications, e.g., classification tasks. In audio, data usually undergo some pre-processing, often called feature extraction, before being fed into the trainable machine. CNNs as feature extractors have been given a rigorous mathematical analysis by Mallat [6], leading to the so-called scattering transform, based on a wavelet transform in each network layer. Invariance and deformation stability properties of the resulting feature extractors were investigated in [6, 2, 1]. Wiatowski and Bölcskei extended the scattering transform by allowing different semi-discrete frames in each network layer, general Lipschitz-continuous non-linearities and pooling operators, see [9, 8]. In this general network, the authors proved vertical translation invariance and deformation stability for band-limited functions. In the current contribution we introduce a new feature extractor called Gabor scattering previously introduced in the conference contribution [3], which is a scattering
transform based on Gabor frames in each layer. Gabor frames are closely related to the short-time Fourier transform (STFT), see Section 2.2, which is commonly used in practice to analyse audio data. We then study invariance properties of Gabor scattering when applied to a common signal model for audio signals, Section 3 and use the same signal model to establish deformation bounds for a feature extractor based on Gabor scattering, Section 4. This approach is motivated by the observation, that assuming more specific restrictions than band-limitation will lead to more precise deformation estimates, when dealing with signals with certain known properties. Thereby, we follow the idea presented in Grohs et al. [5] and use a decoupling technique which proves deformation stability by exploiting stability of the signal class introduced in Section 2.3 in combination with general properties, namely contractivity, which is entirely due to the network architecture.

2.1 Deep Convolutional Neural Networks (DCNNs) and Invariance

In order to understand the mathematical construction used within this paper, we briefly introduce the principal idea and structure of a DCNN. DCNNs enable the computer to learn from observation data. Their structure is inspired by neurones and the human process of learning. Usually a network consists of several layers, namely an input, several hidden (since we consider the case of deep CNN the number of hidden layers is supposed to be $\geq 2$) and one output layer. As input we use the data, that should be classified or more generally whose properties should be learnt. A hidden layer consists of several ingredients: first the convolution of the data with a small weighting matrix, which can be interpreted as localization of certain properties of the input data. Similarly neurones in our brain act only in a local environment, when certain stimulation appears. The next ingredient of the hidden layer is the application of a non-linearity function, also called activation function, which signals if information of this neurone is relevant to be transmitted. Furthermore, in order to reduce redundancy and increase invariance, pooling is applied. Due to the ingredients of the hidden layer, i.e. convolution with a weighting matrix, non-linearity function and pooling, certain invariances are generated [7]. In order to make concrete statements about invariances generated by certain layers for a certain class input data, one needs to develop a signal model, as we do in Section 2.3. Within our simple model for tones, convolution is performed with a low-pass filter and thus in combination with pooling, temporal fine structure is averaged out, while frequency information is maintained. Hence in our case we have some invariance w.r.t. the envelope in the first layer. Moreover, after further convolutions, more temporal fine structure is removed and information on large scales is captured in higher layers. In our model the second layer is invariant w.r.t. the pitch and information about the envelopes is visible, cp. Section 3 and 5. Note that in a neural network, in particular in CNNs, the output, e.g. classification labels, is obtained after several concatenated hidden layers. In the case of scattering network the outputs of each layer are stacked together into a feature vector and further processing is necessary to obtain the desired result. Usually, after some kind of dimensionality reduction, cf. [10], this vector can be fed into a support vector machine or a general NN, which performs the classification task.
2.2 Gabor Scattering

Since Wiatowski and Bölcskei used general semi-discrete frames to obtain a wider class of window functions for the scattering transform (cp. [9, 8]), it seems natural to consider a specific model used for audio data and analyse Gabor frames for the scattering transform and study corresponding properties. We next introduce the basics of Gabor frames and refer to [4] for more details. A sequence \((g_k)_{k=1}^\infty\) of elements in a Hilbert space \(H\) is called frame if there exist positive frame bounds \(A, B > 0\) such that for all \(f \in H\)

\[
A \|f\|^2 \leq \sum_{k=1}^\infty |\langle f, g_k \rangle|^2 \leq B \|f\|^2.
\]

(1)

If \(A = B\), then we call \((g_k)_{k=1}^\infty\) a tight frame. In order to define Gabor frames we need to introduce two operators, i.e. the translation and modulation operator of a function \(f \in L^2(\mathbb{R})\).

- the translation (time shift) operator is defined as \(T_x f(t) = f(t - x)\) for all \(t \in \mathbb{R}\).
- the modulation (frequency shift) operator is defined as \(M_\omega f(t) = e^{2\pi i \omega t} f(t)\) for all \(\omega \in \mathbb{R}\).

Moreover, we can use these operators to express the short-time Fourier transform (STFT) of a function \(f \in L^2(\mathbb{R})\) with respect to a given window function \(g \in L^2(\mathbb{R})\) as \(V_g f(x, \omega) = \langle f, M_\omega T_x g \rangle\). In order to reduce redundancy, we sample \(V_g f\) on a separable lattice \(\Lambda = \alpha Z \times \beta Z\), in time by \(\alpha > 0\) and in frequency by \(\beta > 0\). The resulting samples correspond to the coefficients of \(f\) w.r.t. a Gabor system.

Definition 1. (Gabor System)

Given a window function \(0 \neq g \in H\), where the Hilbert space \(H\) is either \(L^2(\mathbb{R})\) or \(\ell^2(\mathbb{Z})\), and lattice parameters \(\alpha, \beta > 0\), the set of time-frequency shifted versions of \(g\)

\[
G(g, \alpha, \beta) = \{M_\beta j T_\alpha k g : j, k \in \mathbb{Z}\}
\]

is called a Gabor system.

We proceed to introduce a scattering transform for Gabor frames. We base our considerations on [9] by using a triplet-sequence \(\Omega = ((\Psi_\ell, \sigma_\ell, S_\ell))_{\ell \in \mathbb{N}}\) here \(\ell\) is associated to the \(\ell\)-th layer of the network. Note that in this contribution we will deal with Hilbert spaces \(L^2(\mathbb{R})\) or \(\ell^2(\mathbb{Z})\); more precisely in the input layer, i.e. the 0-th layer, we have \(H_0 = L^2(\mathbb{R})\) and due to the discretization inherent in the Gabor transform, \(H_\ell = \ell^2(\mathbb{Z})\) \(\forall \ell > 0\).

We briefly review the elements of the triplet:

- \(\Psi_\ell := \{g_{\lambda_t} \}_{\lambda_t \in \Lambda_\ell}\) with \(g_{\lambda_t} = M_\beta_{\ell j} T_\alpha_{\ell k} g_\ell, \lambda_t = (\alpha_{\ell k}, \beta_{\ell j})\), is a Gabor frame indexed by a lattice \(\Lambda_\ell = \alpha_{\ell} Z \times \beta_{\ell} Z\).
In order to explain the interpretation of Gabor scattering as CNN, we write $g: \mathbb{C} \rightarrow \mathbb{C}$, is applied pointwise and is chosen to be Lipschitz-continuous, i.e. $\|\sigma f - \sigma h\|_2 \leq L_\ell \|f - h\|_2$ for all $f, h \in L^2(\mathbb{R})$. In this paper we only use the modulus function with Lipschitz constant $L_\ell = 1$ for all $\ell \in \mathbb{N}$.

- Pooling depends on a pooling factor $S_\ell > 0$, which leads to dimensionality reduction. Mostly used are max- or average-pooling, some more examples can be found in [9]. In our context, pooling is covered by choosing specific lattices $\Lambda_\ell$ in each layer.

In order to explain the interpretation of Gabor scattering as CNN, we write $T(g)(t) = g(-t)$ and have $|\langle f, M_{\beta j} g \rangle| = |f \ast (T(M_{\beta j}(g)))(\alpha k)|$. Thus the Gabor coefficients can be interpreted as the samples of a convolution.

**Definition 2. (Gabor Scattering)**

Let $\Omega = (\{\Psi_\ell, \sigma_\ell, \Lambda_\ell\})_{\ell \in \mathbb{N}}$ be a triplet-sequence, with ingredients explained above. Then the $\ell$-th layer of the Gabor scattering transform is defined as the output of the operator $U_\ell : \beta_\ell \mathbb{Z} \times \mathbb{H}_{\ell-1} \rightarrow \mathbb{H}_\ell$:

$$f_\ell := U_\ell[\beta_\ell] f_{\ell-1}(k) := \sigma_\ell((f_{\ell-1}, M_{\beta_\ell} T_{\alpha_\ell} g_\ell)_{\mathbb{H}_{\ell-1}}),$$

where $f_{\ell-1}$ is the output-vector of the previous layer and $f_\ell \in \mathbb{H}_\ell \forall \ell$.

Taking the calculation steps of the previous layers into account, we can extend (2) to paths on index sets $q := (q_1, ..., q_\ell) \in \beta_1 \mathbb{Z} \times ... \times \beta_\ell \mathbb{Z} := B^\ell, \ell \in \mathbb{N}$ and obtain

$$U[q]f = U[(q_1, ..., q_\ell)]f := U_\ell[q_\ell] \cdots U_1[q_1]f.$$

Similar to [9] for each layer, we use one atom of the Gabor frame in the subsequent layer as output-generating atom, i.e. $\phi_{\ell-1} := g_{\lambda_\ell}; \lambda_\ell \in \Lambda_\ell$. Since this element is the $\ell$-th convolution, it is an element of the $\ell$-th frame, but because it belongs to the $(\ell-1)$-th layer, its index is $(\ell-1)$. We also want to introduce a countable set $Q := \bigcup_{\ell=0}^{\infty} B^\ell$ and the space $(\ell^2(\mathbb{Z}))^Q$ of sets $s := \{s_q\}_{q \in Q}, s_q \in \ell^2(\mathbb{Z})$ for all $q \in Q$. Now we can define the feature extractor $\Phi_\Omega(f)$ of a signal $f \in L^2(\mathbb{R})$ as in [9] Def. 3.

**Definition 3. (Feature Extractor)**

Let $\Omega = (\{\Psi_\ell, \sigma_\ell, \Lambda_\ell\})_{\ell \in \mathbb{N}}$ be a triplet-sequence and $\phi_\ell$ the output generating atom for each layer. Then the feature extractor $\Phi_\Omega : L^2(\mathbb{R}) \rightarrow (\ell^2(\mathbb{Z}))^Q$ is defined as

$$\Phi_\Omega(f) := \bigcup_{\ell=0}^{\infty} \{U[q]f \ast \phi_\ell\}_{q \in B^\ell}.$$  

In the following section we are going to introduce the signal model which we consider in this paper.
2.3 Musical Signal Model

Tones are one of the smallest units and simple models of an audio signal, consisting of one fundamental frequency $\xi_0$, corresponding harmonics $n\xi_0$ and a shaping envelope $A_n$ for each harmonic, providing specific timbre. Further, since our ears are limited to frequencies below $20kHz$, we develop our model over finitely many harmonics, i.e. $\{1, ..., N\} \subset \mathbb{N}$.

The general model has the following form

$$f(t) = \sum_{n=1}^{N} A_n(t)e^{2\pi i n t},$$  \hspace{1cm} (4)

where $A_n(t) \geq 0 \forall n \in \{1, ..., N\}$ and $\forall t$. For one single tone we choose $\eta_n(t) = n\xi_0 t$.

Moreover we can create a space of tones $T = \{ \sum_{n=1}^{N} A_n(t)e^{2\pi i n \xi_0 t} | A_n \in C^\infty(\mathbb{R}) \}$ and assume $\|A_n\|_\infty \leq 1$.

3 Gabor Scattering of Music Signals

In [2] it was already stated that due to the structure of the scattering transform the energy of the signal is pushed towards low frequencies, where it then is captured by a low-pass filter as output generating atom. In the current section we explain how Gabor scattering separates relevant structures of signals described by our signal model $T$. Due to the smoothing action of the output generating atom, each layer expresses certain invariances, which will be illustrated by numerical examples in Section 5. In Proposition 1, inspired by [2], we add some assumptions on the analysis window in the first layer $g$:

$$|\hat{g}(\omega)| \leq C\hat{g}(1 + |\omega|^s)^{-1}$$

for some $s > 1$ and $\|tg(t)\|_1 = C_g < \infty$.

**Proposition 1.** Let $f(t) \in T$ with $\|A'_n\|_\infty \leq C_n < \infty \forall n \in \{1, ..., N\}$. For fixed $j$, $n_0$ is chosen such that $n_0 = \arg\min n |\beta_j - \xi_0 n|$. Then we obtain

$$U[\beta_j](f)(k) = |\langle f, M_{\beta_j}T_{\alpha k}g \rangle| = A_{n_0}(\alpha k)\hat{g}(\beta_j - n_0 \xi_0) + E_1(k)$$

$$E_1(k) \leq C_g \cdot \sum_{n=1}^{N} \|A_n' \cdot T_k \chi[-\alpha; \alpha]\|_\infty + 2C_g \sum_{n>0} (1 + |\xi_0|^s n - \frac{1}{2})^s, \hspace{1cm} (5)$$

where $\chi$ is the indicator function.

Equation 5 shows that for slowly varying amplitude functions $A_n$, the first layer mainly captures the contributions near the frequencies of the tone’s harmonics. Obviously, for time-sections during which the envelopes $A_n$ undergo faster changes, such as during a tone’s onset, energy will also be found outside a small interval around the harmonics’ frequencies and thus the error estimate 6 becomes less stringent.

**Proof.** Step1 – Using the signal model for tones as input, interchanging the finite sum
with the integral and performing a substitution \( t' = t - \alpha k \), we obtain

\[
\langle f, M_{j\beta} T_{\alpha k} g \rangle = \langle \sum_{n=1}^{N} A_n M_{\beta j-n \xi_0} T_{\alpha k} g \rangle
\]

\[
= \sum_{n=1}^{N} \langle A_n, M_{\beta j-n \xi_0} T_{\alpha k} g \rangle
\]

\[
= \sum_{n=1}^{N} \int_{\mathbb{R}} A_n(t' + \alpha k) g(t') e^{2\pi i (n \xi_0 - \beta j)(t'+\alpha k)} dt'.
\]

After performing a Taylor series expansion for \( A_n(t' + \alpha k) = A_n(\alpha k) + t'R_n(\alpha k, t') \), where \( |R_n(\alpha k, t')| \leq \|A_n' \cdot T_k \chi[-\alpha; \alpha]\|_\infty \) we have

\[
\langle f, M_{j\beta} T_{\alpha k} g \rangle = \sum_{n=1}^{N} \left[ e^{2\pi i (n \xi_0 - \beta j) \alpha k} A_n(\alpha k) \int_{\mathbb{R}} g(t') e^{2\pi i (n \xi_0 - \beta j)(t'+\alpha k)} dt' \right. \\
+ \left. \int_{\mathbb{R}} t'R_n(\alpha k, t') g(t') e^{2\pi i (n \xi_0 - \beta j)(t'+\alpha k)} dt' \right].
\]

Hence we choose \( n_0 = \text{argmin}_n |\beta j - \xi_0 n| \), set

\[
E_n(k) = \int_{\mathbb{R}} t'R_n(\alpha k, t') g(t') e^{2\pi i (n \xi_0 - \beta j)(t'+\alpha k)} dt'
\]

(7)

\[
\tilde{E}(k) = \sum_{n \neq n_0}^{N} e^{2\pi i (n \xi_0 - \beta j) \alpha k} A_n(\alpha k) \hat{g}(\beta j - n \xi_0)
\]

(8)

and split the sum to obtain

\[
\langle f, M_{j\beta} T_{\alpha k} g \rangle = A_{n_0}(\alpha k) e^{-2\pi i (\beta j - n_0 \xi_0) \alpha k} \hat{g}(\beta j - n_0 \xi_0) + \tilde{E}(k) + \sum_{n=1}^{N} E_n(k).
\]

Step 2 – We now bound (7):

\[
|\sum_{n=1}^{N} E_n(k)| = |\sum_{n=1}^{N} \int_{\mathbb{R}} t'R_n(\alpha k, t') g(t') e^{2\pi i (n \xi_0 - \beta j)(t'+\alpha k)} dt'| \\
\leq \sum_{n=1}^{N} \|A_n' \cdot T_k \chi[-\alpha; \alpha]\|_\infty \int_{\mathbb{R}} |t' g(t')| dt' \\
\leq C_g \sum_{n=1}^{N} \|A_n' \cdot T_k \chi[-\alpha; \alpha]\|_\infty.
\]
For the second bound, i.e. the bound of Equation (8), we use the decay condition on \( \hat{g} \), thus

\[
|\tilde{E}(k)| \leq C_{\hat{g}} \sum_{n=1}^{N} |A_n(ak)|(1 + |\beta j - \xi_0 n|^s)^{-1}.
\]

Next we split the sum into \( n > n_0 \) and \( n < n_0 \). We estimate the error term for \( n > n_0 \):

\[
\sum_{n=n_0+1}^{N} |A_n(ak)|(1 + |\beta j - \xi_0 n|^s)^{-1} = \sum_{n=n_0+1}^{N-n_0} |A_{n_0+n}(ak)|(1 + |\beta j - \xi_0 n_0 - \xi_0 n|^s)^{-1}.
\]

(9)

Since \( n_0 = \arg \min_n |\beta j - \xi_0 n| \), we have \( |\beta j - \xi_0 n_0| < \frac{\xi_0}{2} \) and also using \( \|A_n\|_{\infty} \leq \frac{1}{n} \), we obtain

\[
\sum_{n=n_0+1}^{N-n_0} |A_{n_0+n}(ak)|(1 + |\xi_0 n_0 - \xi_0 n|^s)^{-1} \leq \sum_{n=n_0+1}^{N-n_0} \frac{1}{n+n_0} (1 + |\xi_0|^s n - \frac{1}{2}|s)^{-1}.
\]

Due to symmetry we get

\[
|\tilde{E}(k)| \leq 2C_{\hat{g}} \sum_{n=n_0+1}^{N-n_0} (1 + |\xi_0|^s n - \frac{1}{2}|s)^{-1}.
\]

(10)

Summing up the error terms, we obtain (8).

\[ \Box \]

**Remark 1.** The error bound of Equation (10) gets bigger for lower frequencies. This makes sense, since the separation of the fundamental frequency and corresponding harmonics by the analysis window deteriorates. For higher frequencies separation improves and hence the error term gets smaller.

We now introduce two more operators, first the sampling operator \( S_\alpha(f(x)) = f(\alpha x) \) \( \forall x \in \mathbb{R} \) and second the periodization operator \( P_\frac{1}{\alpha}(h) = \sum_{k \in \mathbb{Z}} h(\omega - \frac{k}{\alpha}) \) \( \forall \omega \in \mathbb{R} \).

These operators have the following relation \( F(S_\alpha(f))(\omega) = P_\frac{1}{\alpha}(f(\omega)). \) In order to see how the second layer captures relevant signal structures, depending on the first layer, we propose the following Corollary. Recall that \( g_\ell \in \mathcal{H}_\ell \) \( \forall \ell \in \mathbb{N} \).

**Corollary 1.** Let \( f(|t|) \in \mathcal{T} \), \( \sum_{k \neq 0} |A_{\ell k}(-\frac{1}{\alpha})| \leq \varepsilon_\alpha \) and \( |\hat{g}_2(h)| \leq C_{\hat{g}_2}(1 + |h|)\). Then the elements of the second layer can be expressed as

\[
U_2[\beta_2 h][U_1[\beta_1 j]f(m)] = |\hat{g}_1(\beta_1 j - \xi_0 n_0)| \langle M_{-\beta_2 h} A_{\ell 0}, T_{\ell 2} g_2 \rangle + E_2(m),
\]

where

\[
E_2(m) \leq \varepsilon_\alpha C_{\hat{g}_2} |\hat{g}_1(\beta_1 j - \xi_0 n_0)| \sum_r (1 + |\beta_2 h - r|^s)^{-1} + \|E_1\|_{\infty} \cdot \|g\|_{1}.
\]
Proof. Using the outcome of Proposition 1 we obtain
\[ U_2[\beta_2 h]U_1[\beta_1 j]f(m) = \]
\[ |\langle S_{a_1}(A_{n_0})|\tilde{g}_1(\beta_1 j - \xi_0 n_0)\rangle + E_1, M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle|_{\mathcal{L}(Z)} \leq \]
\[ |\langle S_{a_1}(A_{n_0})|\tilde{g}_1(\beta_1 j - \xi_0 n_0)\rangle, M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle|_{\mathcal{L}(Z)}| + |\langle E_1, M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle|_{\mathcal{L}(Z)}. \]

For the error \( E_1(k) \) we use the global estimate \( |\langle E_1, M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle|_{\mathcal{L}(Z)} \leq \| E_1 \| \cdot \| g \| \) and, using the notation above we proceed as follows:
\[ \langle S_{a_1}(A_{n_0}), M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle|_{\mathcal{L}(Z)} = \sum_{k \in \mathbb{Z}} A_{n_0}(\alpha_1 k)T_{\alpha_2 m}g_2(k)e^{-2\pi i \beta_2 h k} = \]
\[ F(S_{a_1}(A_{n_0}) \cdot T_{\alpha_2 m}g_2)(\beta_2 h) = F(S_{a_1}(A_{n_0})) \ast F(T_{\alpha_2 m}g_2)(\beta_2 h) = \]
\[ P_{\beta_1}(\hat{A}_{n_0}) \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h) = \left( \sum_{k \in \mathbb{Z}} \hat{A}_{n_0}(\cdot - \frac{k}{\alpha_1}) \right) \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h). \]  

Since the values are maximal in a neighborhood of the center frequency \( \beta_2 h \) we consider the case \( k = 0 \) separately and obtain
\[ \langle S_{a_1}(A_{n_0}), M_{\beta_2 h}T_{\alpha_2 m}g_2\rangle = \left( \hat{A}_{n_0} \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h) \right) + \left( \sum_{k \in \mathbb{Z}} \hat{A}_{n_0}(\cdot - \frac{k}{\alpha_1}) \right) \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h). \]  

It remains to bound the sum, i.e. the second term of Equation (12):
\[ \left| \left( \sum_{k \in \mathbb{Z}\setminus\{0\}} \hat{A}_{n_0}(\cdot - \frac{k}{\alpha_1}) \right) \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h) \right| = \]
\[ \left| \sum_r \left( \sum_{k \in \mathbb{Z}\setminus\{0\}} \hat{A}_{n_0}(r - \frac{k}{\alpha_1}) \right) \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h - r) \right| \leq \]
\[ \sum_{r} \sum_{k \in \mathbb{Z}\setminus\{0\}} \left| \hat{A}_{n_0}(r - \frac{k}{\alpha_1}) \right| \cdot \left| \hat{g}_2(\beta_2 h - r) \right| \]  

In the last Equation (13) we applied the triangle inequality twice and the modulation term can be ignored because of the modulus. Now we can use our assumption \( \sum_{k \in \mathbb{Z}\setminus\{0\}} \left| \hat{A}_{n_0}(\cdot - \frac{k}{\alpha_1}) \right| \leq \varepsilon_{a_1} \) and also the assumption on the Fourier transform of the analysis window \( g_2 \):
\[ \sum_{r} \sum_{k \in \mathbb{Z}\setminus\{0\}} \left| \hat{A}_{n_0}(r - \frac{k}{\alpha_1}) \right| \cdot \left| \hat{g}_2(\beta_2 h - r) \right| \leq \varepsilon_{a_1} \sum_{r} \left| \hat{g}_2(\beta_2 h - r) \right| \leq \varepsilon_{a_1} C_{\beta_2 h} \sum_{r} (1 + |\beta_2 h - r|^s)^{-1}. \]  

We rewrite the first term in (12):
\[ \left( \hat{A}_{n_0} \ast (M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h) \right) = \sum_r \hat{A}_{n_0}(r)(M_{-\alpha_2 m}\hat{g}_2)(\beta_2 h - r) - \langle \hat{A}_{n_0}, T_{\beta_2 h}LM_{-\alpha_2 m}\hat{g}_2 \rangle = -\langle A_{n_0}, M_{\beta_2 h}T_{\alpha_2 m}g_2 \rangle. \]  

(15)
The last Equation (15) uses Plancherel’s theorem. Rewriting the last term we obtain
\[-(A_{n_0}, M_{\beta h} T_{\alpha m} g_2) = -(M_{-\beta h} A_{n_0}, T_{\alpha m} g_2).\]

\[\textbf{Remark 2.}\] The sum \(\sum_r (1 + |\beta_2 h - r|^s)^{-1}\) decreases very fast, i.e. taking \(s = 3\) the summand is already smaller as \(10^{-5}\) for \(r = 48\).

\[\textbf{Remark 3.}\] Note that, since the envelopes \(A_n\) are expected to change slowly except around transients, their Fourier transforms concentrate their energy in the low frequency range. In Section 5 it will be shown by means of the analysis of example signals, how the second layer distinguishes tones which have a smooth onset (transient) from those which have a sharp attack, which leads to broadband characteristics of \(A_n\) around this attack. Similarly, if \(A_n\) undergoes an amplitude modulation, the frequency of this modulation can be clearly discerned, cf. Figure 7 and the corresponding example. This observation is clearly reflected in expression (12). Since \(\hat{g}_2\) decays fast, the aliasing terms in (12), for \(k > 0\) are sufficiently small.

To obtain the Gabor scattering coefficients, we need to apply the output generating atom as in (3).

\[\textbf{Corollary 2.}\] Let \(\phi_1 \in \Psi_2\), then the output of the first layer is
\[U_1[\beta_1 j] f * \phi_1(k) = |\hat{g}(\beta j - n_0 \xi)| (A_{n_0} * \phi_1)(k) + \epsilon_1(k),\]
where
\[\epsilon_1(k) \leq C'\sum_{n=1}^{N} \|A'_{n} \cdot T_{k} \chi[-\alpha; \alpha]\|_{\infty} + 2C'\sum_{n=0}^{N-n_0} (1 + |\xi_0|^s |n - \frac{1}{2}|)^{-1}\]
and with \(\phi_2 \in \Psi_3\) the second layer output is
\[U_2[\beta_2 h] U_1[\beta_1 j] f * \phi_2(m) = |\hat{g}_1(\beta_1 j - \xi_0 n_0)| (M_{-\beta_2 h} A_{n_0}, T_{\alpha m} g_2) * \phi_2 + \epsilon_2(m)\]
with
\[\epsilon_2(m) \leq \epsilon_1 C'\|\hat{g}_1(\beta_1 j - \xi_0 n_0)\| \sum_r (1 + |\beta_2 h - r|^s)^{-1} + \|E_1\|_{\infty} \cdot \|g\|_1 \|\phi_2\|_1.\]

\[\textbf{Remark 4.}\] Note that the convolution is a low-pass filter for sufficient smoothness of \(\phi_1\). Hence, in dependence on the pooling factor \(\alpha_1\), the temporal fine-structure of \(A_{n_0}\) is averaged out. In the second layer, applying the output generating atom \(\phi_2 \in \Psi_3\) removes the fine temporal structure and thus, the second layer reveals information contained in the envelopes \(A_n\).
Proof. We show the calculations for the first layer, for the subsequent layer it is the same:

\[
\frac{|\langle f, M_{\beta j} T_{\alpha k} g \rangle | * \phi - |\hat{g}(\beta j - \xi_0 n_0) A_{\alpha_0} (\alpha k) * \phi |}{\phi - |\hat{g}(\beta j - \xi_0 n_0) A_{\alpha_0} (\alpha k) * \phi |} \\
\leq |\int E(x) |* \phi (k - x) dx |^2 \\
\leq |\int K \phi (k - x) dx |^2 = K^2 |\int \phi (k - x) dx |^2 \leq K^2 \cdot \| \phi \|_1^2,
\]

where \( K = C_\varphi \sum_{n=1}^{N} \| A'_n \cdot T_k \chi [-\alpha; \alpha] \|_\infty + 2 C_\varphi \sum_{n>0} (1 + |\xi_0|^n n - \frac{1}{2}|s|)^{-1} \). The factor \( \| \phi \|_1 \) will be absorbed by the constants in the error terms, i.e.

\[
\epsilon_1(k) \leq K \cdot \| \phi \|_1 = C_\varphi \sum_{n=1}^{N} \| A'_n \cdot T_k \chi [-\alpha; \alpha] \|_\infty + 2 C_\varphi \sum_{n>0} (1 + |\xi_0|^n n - \frac{1}{2}|s|)^{-1}.
\]

Calculations are similar for the second layer. \( \square \)

4 Deformation stability

In this section we study to which extent Gabor scattering is stable with respect to certain deformations. We consider changes in spectral shape as well as frequency modulations. The method we apply is inspired by [5] and uses the decoupling technique, i.e. in order to prove stability of the feature extractor we first take the structural properties of the signal class into account and search for an error bound of deformations of the signals in \( T \). In combination with the contractivity property \( \| \Phi_\Omega (f) - \Phi_\Omega (h) \|_2 \leq \| f - h \|_2 \) of \( \Phi_\Omega \), see [9, Prop. 4], which follows from \( B_{\ell} \leq 1 \) \( \forall \ell \in \mathbb{N} \), where \( B_{\ell} \) is the upper frame bound of the Gabor frame \( G(g_\ell, \alpha_\ell, \beta_\ell) \), this yields deformation stability of the feature extractor.

4.1 Envelope Changes

Simply deforming a tone would correspond to deformations of the envelope \( A_n, \ n = 1, \ldots, N \). This corresponds to a change in timbre, for example by playing a note on a different instrument. Mathematically this can be expressed as:

\[
\tilde{f}_{A_n}(f)(t) = \sum_{n=1}^{N} A_n (t + \tau (t)) e^{2\pi i n \xi_0 t}.
\]

Lemma 1. Let \( f(t) \in T \) and \( |A'_n (t)| \leq C_n (1 + |t|^s)^{-1} \), for constants \( C_n > 0, \ n = 1, \ldots, N \) and \( s > 1 \). Moreover let \( \| \tau \|_\infty \ll 1 \). Then

\[
\| f - \tilde{f}_{A_n}(f) \|_2 \leq D \| \tau \|_\infty \sum_{n=1}^{N} C_n,
\]

for \( D > 0 \) depending only on \( \| \tau \|_\infty \).

10
Proof. Setting $h_n(t) = A_n(t) - \mathcal{F}_{A_n}(A_n(t))$ we obtain

$$
\|f - \mathcal{F}_{A_n}(f)\|_2 \leq \sum_{n=1}^{N} \|h_n(t)\|_2.
$$

We apply the mean value theorem for a continuous function $A_n(t)$ and get

$$
|h_n(t)| \leq \|\tau\|_\infty \sup_{y \in B_{r=\|\tau\|_\infty}(t)} |A_n'(y)|.
$$

Applying the $2$–norm on $h_n(t)$ and the assumption on $A_n'(t)$, we obtain:

$$
\int_{\mathbb{R}} |h_n(t)|^2 dt \leq \int_{\mathbb{R}} \|\tau\|_\infty^2 (\sup_{y \in B_{r=\|\tau\|_\infty}(t)} |A_n'(y)|)^2 dt
$$

$$
\leq C_n^2 \|\tau\|_\infty^2 \int_{\mathbb{R}} \sup_{y \in B_{r=\|\tau\|_\infty}(t)} (1 + |y|)^{-2} dy.
$$

Splitting the integral into $B_1(0)$ and $\mathbb{R} \setminus B_1(0)$ and using the monotonicity of $(1 + |t|^s)^{-1}$ we have

$$
\|h_n(t)\|_2^2 \leq C_n^2 \|\tau\|_\infty^2 \left( \int_{B_1(0)} 1 dy + \int_{\mathbb{R} \setminus B_1(0)} (1 + |y| - \|\tau\|_\infty)^{-2} dy \right)
$$

Moreover for $y \notin B_1(0)$ we have $|(1 - \|\tau\|_\infty)y|^s \leq |(1 - \|\tau\|_\infty)|y|^s$. This leads to

$$
\|h_n(t)\|_2^2 \leq C_n^2 \|\tau\|_\infty^2 \left( \text{vol}(B_1(0)) + \int_{\mathbb{R} \setminus B_1(0)} (1 + |(1 - \|\tau\|_\infty)y|^s)^{-2} dy \right)
$$

(16)

$$
= C_n^2 \|\tau\|_\infty^2 \left( \text{vol}(B_1(0)) + \frac{1}{1 - \|\tau\|_\infty} \int_{\mathbb{R}} (1 + |x|^s)^{-2} dx \right)
$$

$$
= C_n^2 \|\tau\|_\infty^2 \left( \text{vol}(B_1(0)) + \frac{1}{1 - \|\tau\|_\infty} \frac{1}{1 + |x|^s_2} \right).
$$

In Equation (16) we performed a change of variables, i.e. $x = (1 - \|\tau\|_\infty)y$. Setting $D^2 := (\text{vol}(B_1(0)) + \frac{1}{1 - \|\tau\|_\infty} \frac{1}{1 + |x|^s_2})$ and summing up we obtain

$$
\|f - \mathcal{F}_{A_n}(f)\|_2 \leq D\|\tau\|_\infty \sum_{n=1}^{N} C_n.
$$

\[\square\]

Remark 5. Harmonics’ energy decreases with increasing frequency, hence $C_n \ll C_{n-1}$. 

11
4.2 Frequency modulation

Another different kind of sound deformation results from frequency modulation of \( f(t) \in \mathcal{T} \). This corresponds to, for example, playing higher or lower pitch, or producing a vibrato. This can be formulated as:

\[
\tilde{g}_\tau f(t) \mapsto \sum_{n=1}^{N} A_n(t)e^{2\pi i(n\xi_0 t + \tau_n(t))}.
\]

Lemma 2. Let \( f(t) \in \mathcal{T} \). Moreover let \( \|\tau_n\|_{\infty} < \frac{\arccos(1 - \frac{\varepsilon^2}{2\pi})}{2\pi} \). Then

\[
\|f - \tilde{g}_\tau(f)\|_2 \leq \varepsilon \sum_{n=1}^{N} \frac{1}{n}.
\]

Proof. We have

\[
\|f - \tilde{g}_\tau(f)\|_2 \leq \sum_{n=1}^{N} \|h_n(t)\|_2,
\]

with \( h_n(t) = A_n(t)(1 - e^{2\pi i\tau_n(t)}) \). Now we compute the 2–norm of \( h_n(t) \):

\[
\int_{\mathbb{R}} |h_n(t)|^2 dt = \int_{\mathbb{R}} |A_n(t)(1 - e^{2\pi i\tau_n(t)})|^2 dt \\
\leq \|1 - e^{2\pi i\tau_n(t)}\|_2^2 \|A_n(t)\|_2^2.
\]

Now we rewrite

\[
|1 - e^{2\pi i\tau(t)}|^2 = (1 - \cos(2\pi \tau_n(t)))^2 + \sin^2(2\pi \tau_n(t)) = 2(1 - \cos(2\pi \tau_n(t))).
\]

Setting \( \|1 - e^{2\pi i\tau_n(t)}\|_2^2 \leq \varepsilon^2 \), this term gets small if \( \|\tau_n(t)\|_{\infty} \leq \frac{\arccos(1 - \frac{\varepsilon^2}{2\pi})}{2\pi} \). Using the assumptions of our signal model on the envelopes, i.e. \( \|A_n\|_{\infty} < \frac{1}{n} \), we obtain

\[
\|f - \tilde{g}_\tau(f)\|_2 \leq \varepsilon \sum_{n=1}^{N} \frac{1}{n}.
\]

\[\square\]

5 Numerical Examples

In this section we are going to show some in MATLAB performed numerical examples of the developed mathematical theory of the Gabor scattering transform. As input we produced a single tone following the signal model from Section 2.3. The first example, Fig. 1 shows two tones, played sequentially, having the same fundamental frequency \( \xi_0 = 800 \) Hz and 15 harmonics, but different envelopes, i.e. the first tone has a sharp attack, maintains and goes softly to zero, the second starts with a soft attack and has some amplitude modulation. An amplitude modulated signal would for example correspond to \( f(t) = \sum_{n=1}^{N} \sin(2\pi 200t)e^{2\pi i n\xi_0 t} \), here the signal is modulated by 20Hz. We explain now, what is visible within Figure 1.
• **Gabor transform**: In the spectogram showing the Gabor transform, we clearly see the difference between the envelopes and we see that the signals have the same pitch and the same harmonics.

• **First layer**: The first layer is invariant w.r.t. the envelope of the signals. This is due to the output generating atom and the subsampling, which removes temporal information of the envelope. In the spectogram corresponding to the first layer, there is no information about the envelope visible, hence the spectogram of the different signals look almost the same.

• **Second layer**: For the second layer we took as input a time vector at fixed frequency of the first layer. Here we fixed the fundamental frequency. The second layer is invariant w.r.t. the pitch, but differences on a larger scale are captured. Within this layer we are able to distinguish the different envelopes of the signals. We first see the sharp attack of the first tone and then we can distinguish the modulation, where a second frequency is visible.

The second example, Fig. 2 and 3 shows two tones, both having a smooth envelope, but different fundamental frequencies and number of harmonics. The first tone has fundamental frequency $\xi_0 = 800\text{Hz}$ and 15 harmonics and the second tone has fundamental frequency $\xi_0 = 1060\text{Hz}$ and 10 harmonics.

In the following we explain, what is displayed in the Figures 2 and 3.
Figure 2: Gabor transform and first layer of two tones having different fundamental frequencies.

Figure 3: Second layer of two tones having different fundamental frequencies, at different fixed frequencies of the first layer.
• **Gabor transform:** The first spectogram of Figure 2 shows the Gabor transform. We see the different fundamental frequencies of the two tones and also that tone one has more harmonics than tone two.

• **First layer:** The second spectogram of Figure 2 shows the first layer, which is the layer invariant w.r.t. the envelope of the signals. Since both tones have the same envelope, there is nothing to say about this. But we can compare this layer with the first layer of Figure 1 and see, that they look similar. The difference in the fundamental frequency is still visible.

• **Second layer:** For the second layer we prepared several outputs, displayed in Figure 3. As input for the first spectogram, we took a time vector at fixed fundamental frequency of the first tone, i.e. $\xi_0 = 800\text{Hz}$. Since the second tone does not contribute to this frequency, we do not see anything of the second tone. As input for the second spectogram, we took a time vector at fixed fundamental frequency of the second tone, i.e. $\xi_0 = 1060\text{Hz}$. We can see that the first tone is not contributing at this frequency. As input for the third spectogram, we took a frequency that both share within their harmonics, i.e. about $\xi = 9550\text{Hz}$. Here we can see that both tones contribute to the second layer output. Although we obtain different outputs for different fixed frequency vectors from the first layer, we cannot see a difference in using different frequencies, since the output bumps look all the same. Hence this layer is invariant to frequency.

The source code, and further examples are provided on [http://homepage.univie.ac.at/monika.doerfler/GaborScattering.html](http://homepage.univie.ac.at/monika.doerfler/GaborScattering.html).

6 Conclusion and Perspectives

Within this paper we introduced a method that combines Gabor frames and the scattering transform of Mallat, called Gabor scattering. This method has been applied to a simple signal model, consisting of the class of tones. We investigated that different layers of the Gabor scattering, especially here the first and the second, are invariant w.r.t. certain signal features, like pitch and envelope of an audio signal. Moreover we proved that this feature extractor is robust to changes in the spectral shape and frequency modulation. As a next step we will use the output of the Gabor scattering transform as input to various classifiers, e.g. SVM (support vector machine), but also simple (C)NN and see whether it improves the performance of the classification or not. Moreover an extension of this work will be the investigation of the impact of 2-dimensional window functions in higher layers, as it has already been introduced by [1]. In audio tasks this models standard CNNs more closely and is expected to capture complex time-frequency structures.
Acknowledgment

This work was supported by the Uni:docs Fellowship Programme for Doctoral Candidates in Vienna and the Vienna Science and Technology Fund (WWTF) project SALSA (MA14-018).

References

[1] J. Andén, V. Lostanlen, and S. Mallat. Joint time-frequency scattering for audio classification. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6, 2015.

[2] J. Andén and S. Mallat. Deep scattering spectrum. IEEE Transactions on Signal Processing, 62(16):4114–4128, 2014.

[3] R. Bammer and M. Dörfler. Invariance and Stability of Gabor Scattering for Music Signals. In Proc. of Sampling Theory and Application (Sampta), July 2017.

[4] K. Gröchenig. Foundations of time-frequency analysis. Applied and numerical harmonic analysis. Birkhäuser, Boston, Basel, Berlin, 2001.

[5] P. Grohs, T. Wiatowski, and H. Bölcskei. Deep convolutional neural networks on cartoon functions. In IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016, pages 1163–1167, 2016.

[6] S. Mallat. Group Invariant Scattering. Comm. Pure Appl. Math., 65(10):1331–1398, 2012.

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

[8] T. Wiatowski and H. Bölcskei. Deep Convolutional Neural Networks Based on Semi-Discrete Frames. In Proc. of IEEE International Symposium on Information Theory (ISIT), pages 1212–1216, June 2015.

[9] T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. CoRR, abs/1512.06293, 2015.

[10] T. Wiatowski, M. Tschannen, A. Stanic, P. Grohs, and H. Bölcskei. Discrete deep feature extraction: A theory and new architectures. In Proc. of International Conference on Machine Learning (ICML), June 2016.