Invariance and Stability of Gabor Scattering for Music Signals

Roswitha Bammer and Monika Dörfler

University of Vienna, Faculty of Mathematics

The Gabor Scattering

The feature extractor we consider is called **Gabor Scattering** [1, 2] and is based on

- Gabor frames
- Mallat's scattering transform [3]

⇒ This feature extractor has certain properties.

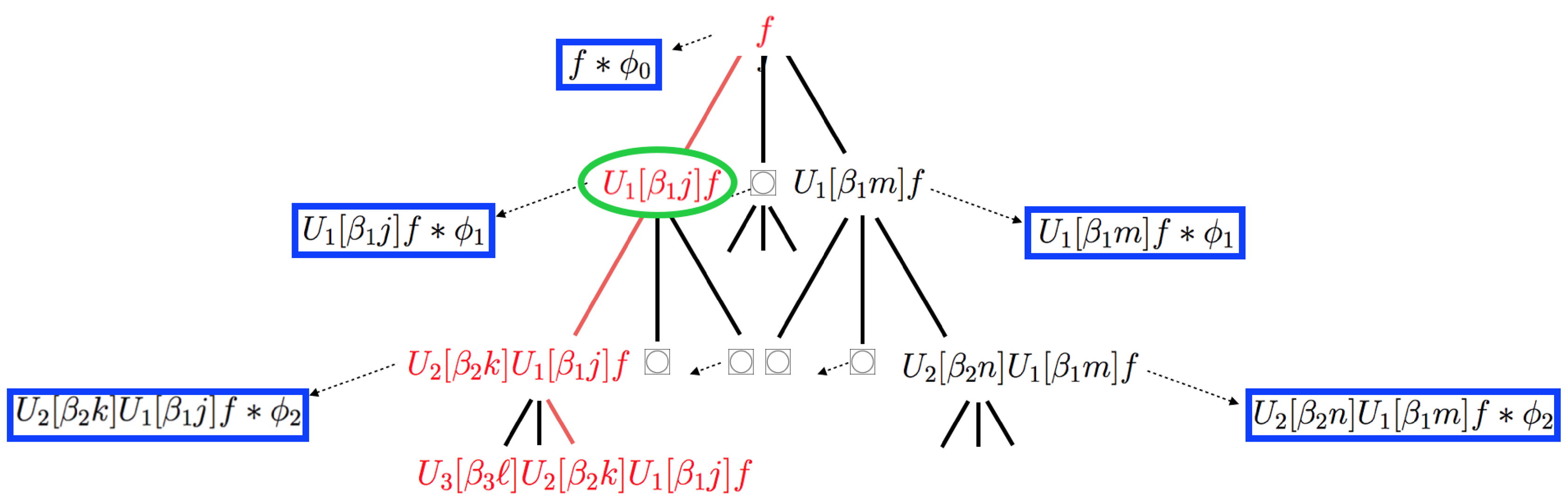
Definitions

In order to define the feature extractor, we need the following definitions [4]:

Triplet Sequence $\Omega = \big((\Psi_\ell, \sigma_\ell, S_\ell)\big)_{\ell\in\mathbb{N}}$:

- $\Psi_\ell := \{g_{\lambda_\ell}\}_{\lambda_\ell\in\Lambda_\ell}$ with $g_{\lambda_\ell} = M_{\beta_\ell j}T_{\alpha_\ell k}g_\ell$, $\lambda_\ell = (\alpha_\ell k, \beta_\ell j)$, is a Gabor frame indexed by a lattice $\Lambda_\ell = \alpha_\ell\mathbb{Z}\times\beta_\ell\mathbb{Z}$.
- Pointwise non-linearity function $\sigma_\ell : \mathbb{C}\to\mathbb{C}$, here: modulus function with Lipschitz constant $L_\ell = 1$.
- Pooling factor $S_\ell > 0$, which leads to dimensionality reduction, here: choosing specific lattices Λ_ℓ in each layer, i.e. $S_\ell = \alpha_\ell$.

Gabor Scattering Network:



Gabor Scattering ℓ−th Layer Element ($\ell = 1 \to$ the element in the green circle): is defined as the output of the operator $U_\ell : \beta_\ell\mathbb{Z}\times\mathcal{H}_{\ell-1}\to\mathcal{H}_\ell$:

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

where $f_{\ell-1}$ is the output-vector of the previous layer. Here $\mathcal{H}_0 = L^2(\mathbb{R})$ and $\mathcal{H}_\ell = \ell^2(\mathbb{Z})$ $\forall\ell > 0$.

Path extension (red path):
$q := (q_1, ..., q_\ell) \in \beta_1\mathbb{Z}\times ...\times\beta_\ell\mathbb{Z} =: \mathcal{B}^\ell$, $\ell\in\mathbb{N}$ and obtain

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

Output-generating atom (elements in the blue boxes):
$\phi_{\ell-1} := g_{\lambda^*_\ell}$, $\lambda^*_\ell\in\Lambda_\ell$.

Definition (Feature Extractor)

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

$$\Phi_\Omega(f) := \bigcup_{\ell=0}^{\infty}\{(U[q]f)*\phi_\ell\}_{q\in\mathcal{B}_1^\ell}.$$

Here $\mathcal{Q} := \bigcup_{\ell=0}^{\infty}\mathcal{B}^\ell$ and the space $(\ell^2(\mathbb{Z}))^{\mathcal{Q}}$ of sets $s := \{s_q\}_{q\in\mathcal{Q}}$, $s_q\in\ell^2(\mathbb{Z})$ for all $q\in\mathcal{Q}$.

Signal Model

In order to verify the properties of the feature extractor in audio, we need a signal model. The simplest model for audio one can think of, is the **class of tones:**
$\mathcal{T} = \{\sum_{n=1}^N A_n(t)e^{2\pi i n\xi_0 t} | A_n\in\mathcal{C}_c^\infty(\mathbb{R})\}$.

ξ_0... fundamental frequency
$A_n(t)$... envelope for each harmonic
N... number of harmonics is finite, since our ear is limited to 20kHz

Properties

- **Invariance:**
  Different layers create invariances to certain signal features [5], we have a look at the output of layer one and two.

Proposition (1st layer output)

Let $f(t)\in\mathcal{T}$ with $\|A_n\|_\infty\le 1$, $\|A_n'\|_\infty<\infty$ $\forall n\in\{1,...,N\}$, $g_1 : |\hat{g}_1(\omega)|\le C_{\hat{g}_1}(1+|\omega|^s)^{-1}$ for some $s>1$ and $\|tg_1(t)\|_1 = C_{g_1}<\infty$. For fixed j, n_0 is chosen s. t. $n_0 = \underset{n}{\operatorname{argmin}}|\beta_1 j - \xi_0 n|$. Moreover let $\phi_1\in\Psi_2$, then the output of the first layer is

$$U_1[\beta_1 j]f*\phi_1(k) = |\hat{g}_1(\beta_1 j - n_0\xi)|(A_{n_0}*\phi_1)(k) + \epsilon_1(k),$$

where

$$\epsilon_1(k)\le C'_{g_1}\cdot\sum_{n=1}^N\|A_n'\cdot T_k\chi[-\alpha_1;\alpha_1]\|_\infty + C'_{\hat{g}_1}\sum_{n>0}\Big(1+|\xi_0|^s|n-\frac{1}{2}|^s\Big)^{-1}.$$

⇒ for slowly varying amplitude $A_n\to$ relevant contribution only near the frequencies of the tone's harmonics.
⇒ ϕ_1 low pass filter $\to$ in dependence on pooling factor α_1 temporal fine-structure is averaged out.
⇒ 1st layer is invariant w.r.t. envelope changes.

Corollary (2nd layer output)

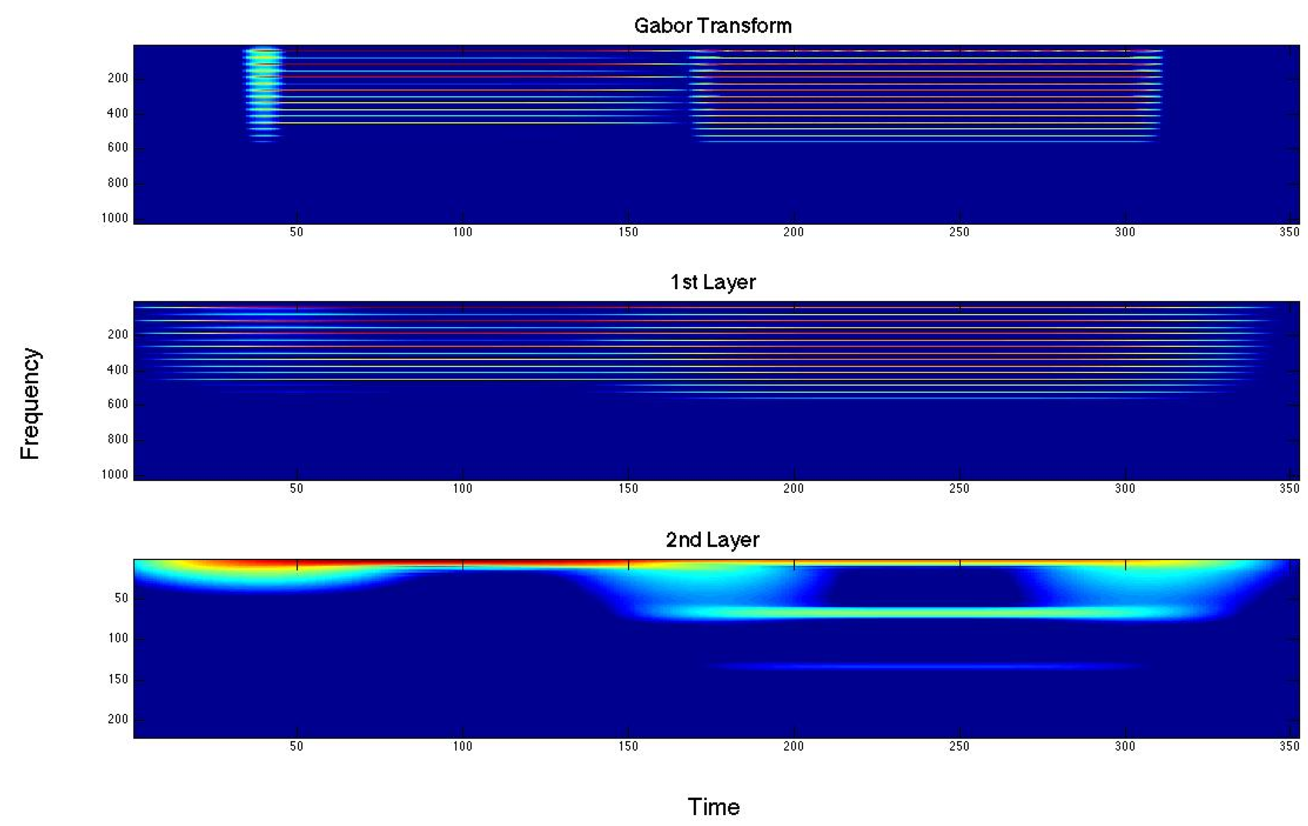
Let $f(t)\in\mathcal{T}$, $\sum_{k\neq 0}|\hat{A}_{n_0}(.-\frac{k}{\alpha_1})|\le\varepsilon_{\alpha_1}$, $|\hat{g}_2(h)|\le C_{\hat{g}_2}(1+|h|^s)^{-1}$ and $\phi_2\in\Psi_3$. Then the second layer output is

$$U_2[\beta_2 h]U_1[\beta_1 j]f*\phi_2(m) = |\hat{g}_1(\beta_1 j-\xi_0 n_0)||\big\langle M_{-\beta_2 h}A_{n_0}, T_{\alpha_2 m}g_2\big\rangle|*\phi_2+\epsilon_2(m)$$

with

$$\epsilon_2(m)\le\varepsilon_{\alpha_1}C'_{\hat{g}_2}|\hat{g}_1(\beta_1 j-\xi_0 n_0)|\sum_r\big(1+|\beta_2 h - r|^s\big)^{-1}+\|E_1\|_\infty\|\phi_2\|_1.$$

⇒ applying $\phi_2\to$ removes fine temporal structure.
⇒ 2nd layer is invariant w.r.t. pitch, reveals information contained in the envelopes A_n.



- **Deformation stability:**
  Stability is obtained by using the decoupling technique [6] relying on the contractivity of feature extractor $\|\Phi_\Omega(f)-\Phi_\Omega(h)\|_2\le\|f-h\|_2$ and error bound of signal class w.r.t. a small deformation τ :

* Envelope changes $\mathfrak{F}_\tau(f)(t) = \sum_{n=1}^N A_n(t+\tau(t))e^{2\pi i n\xi_0 t}$ lead to

Lemma

Let $f(t)\in\mathcal{T}$ and $|A_n'(t)|\le C_n(1+|t|^s)^{-1}$, for some constant $C_n>0$, $n=1,...,N$ and $s>1$. Moreover let $\|\tau\|_\infty\ll 1$, then $\|f-\mathfrak{F}_\tau(f)\|_2\le D\|\tau\|_\infty\sum_{n=1}^N C_n$ for $D>0$ not depending on f and τ.

* Frequency modulation $\mathfrak{F}_\tau(f)(t) = \sum_{n=1}^N A_n(t)e^{2\pi i(n\xi_0 t+\tau_n(t))}$ leads to

Lemma

Let $f(t)\in\mathcal{T}$ and $\|A_n\|_2\le C_n$ for all $n\in\{1,...,N\}$. Moreover let $\|\tau_n\|_\infty<\frac{\arccos(1-\frac{\varepsilon^2}{2})}{2\pi}$, then $\|f-\mathfrak{F}_\tau(f)\|_2\le\varepsilon\sum_{n=1}^N C_n$.

Acknowledgement

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

[2] R. Bammer and M. Dörfler. Gabor Frames and Deep Scattering Networks in Audio Processing. *arXiv preprint: 1706.08818v1*, 2017. http://homepage.univie.ac.at/monika.doerfler/GaborScattering.htm.

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

[4] 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.

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

[6] Philipp Grohs, Thomas Wiatowski, and Helmut 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.