A Frequency-Uniform and Pitch-Invariant Time-Frequency Representation

Sören Schulze1,* and Emily J. King1
1 University of Bremen, FB3, AG Computational Data Analysis, Bibliothekstr. 5, 28359 Bremen, Germany.

We introduce the terms frequency-uniformity and pitch-invariance in order to characterize time-frequency representations. A frequency-uniform representation has the property that it displays Dirac transients as a straight line in the spectrogram, while a pitch-invariant representation translates pitch change into shifts, which is adequate for melodic instruments. We propose a novel representation that fulfills both criteria.

© 2019 The Authors Proceedings in Applied Mathematics & Mechanics published by Wiley-VCH Verlag GmbH & Co. KGaA Weinheim

1 Introduction

A time-frequency representation $U$ transforms a given signal $X: \mathbb{R} \to \mathbb{C}$ into a spectrogram $UX: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that $UX(t, f)$ signifies (in some sense) how much a given frequency $f$ is present in $X$ at a particular time point in time $t$. Especially in audio, spectrograms are very good at encapsulating the relevant information for both human inspection and digital processing. They are widely used for certain audio effects, source separation, speech processing, and music information retrieval (MIR). Different kinds of time-frequency representations are used dependent on the properties required by the application. We formalize two of these properties and define a novel representation that fulfills both of them.

2 Analysis of existing representations

One building block for time-frequency representations is the short-time Fourier transform (STFT), given by (cf. [1]):

$$V_w X(t, f) = F(X \cdot w(-t))(f), \quad X \in S' \mathbb{R}, \quad w \in \mathcal{S} \mathbb{R},$$

where $S \mathbb{R}$ is the Schwartz space, and $S' \mathbb{R}$ is its dual, the space of tempered distributions. It can be shown via standard arguments that $V_w X \in C^\infty \mathbb{R} \times \mathbb{R}$.

**Definition 2.1** A time-frequency representation $U: S' \mathbb{R} \to C(\mathbb{R} \times \Omega)$ with $\Omega \subseteq \mathbb{R}$ is frequency-uniform if $U\delta(t, f)$ is constant in $f$, where $\delta$ is the Dirac distribution.

**Proposition 2.2** For any $w \in S \mathbb{R}$, the time-frequency representation given by $U_w^{\text{STFT}} X = |V_w X|^2$ is frequency-uniform.

**Proof.** Due to $\delta \cdot w(-t) = \delta \cdot w(-t)$, we have $V_w \delta(t, f) = F(\delta \cdot w(-t))(f) = w(-t)$, independently of $f$.

This condition ensures that a signal with a constant frequency spectrum will have a frequency-independent footprint in the spectrogram. This is important when representing short, percussive sounds. Voice and melodic musical instruments such as wood and string instruments make sounds that consist mainly of sinusoids. These instruments are often capable of producing similar sounds at different pitch. When modulating the fundamental frequency, the frequencies of the sinusoids are multiplied with a constant factor, resulting in a dilation along the frequency axis of the STFT spectrogram.

Since dilations are hard to handle computationally, more “musically meaningful” representations have been developed:

- The mel spectrogram is derived from the STFT spectrogram by applying smoothing and a logarithmic transform along the frequency axis:

$$U_w^{\text{mel}} X(t, \log f) = \int_{-\infty}^{\infty} |V_w X(t, \xi)|^2 \cdot \Lambda_f(f - \xi) \, d\xi,$$

  for appropriate $\Lambda_f \in S \mathbb{R}$, $f > 0$.

- The constant-Q transform [2] is a wavelet transform:

$$U_w^{\text{cq}} X(t, \log f) = |F(X \cdot w_\beta f(-t))(f)|^2,$$

  $w_\beta(t) = \frac{1}{\beta} w(t/\beta), \quad w \in S \mathbb{R}, \quad f > 0$.

**Definition 2.3** A time-frequency representation $U: S' \mathbb{R} \to C(\mathbb{R} \times \Omega)$ is pitch-invariant if, for $x_\nu(t) = e^{i2\pi \nu t}$ with arbitrary $\nu > 0$, the spectrogram $U x_\nu(t, \log f)$ only depends on the ratio $\nu/f$.  

* Corresponding author: e-mail sschulze@uni-bremen.de, phone +49 421 218 59896
Proposition 2.4 The mel spectrogram is pitch-invariant for $f > f_0 > 0$ with:

$$w(t) = \sqrt{2} e^{-t^2/\sigma^2}, \quad \Lambda_f(\xi) = \frac{1}{\sqrt{1 - \frac{f^2}{f_0^2}}} \exp\left(-\frac{(2\pi \sigma \xi)^2}{2 (\frac{f^2}{f_0^2} - 1)}\right), \quad \sigma > 0.$$

The constant-Q transform is pitch-invariant for all $w \in \mathcal{S}(\mathbb{R})$.

**Proof.** The latter claim is shown by substitution:

$$U_{\text{cQ}}^w x_\nu(t, \log f) = \left| \int_{-\infty}^\infty f w(f(\eta - t)) e^{-i2\pi (f-\nu) \eta} \, d\eta \right|^2 = |\mathcal{F}w(1-\nu/f)|^2.$$

For the first statement, we note that $|\mathcal{V}_w x_\nu(t, f)|^2 = 2\pi \sigma^2 e^{-((2\pi \sigma f_0 (1-\nu/f))^2/2}$. When convolving two Gaussians, their means and variances add and their $L_1$-norms multiply, yielding $U_{\text{cQ}}^w(t, \log f) = \sqrt{2} \pi \sigma^2 e^{-((2\pi \sigma f_0 (1-\nu/f))^2/2}.$

Regarding the response to a Dirac distribution, we get $U_{\text{cQ}}^w \delta(t, \log f) = |w_1(\nu/f)|^2$, which is only constant in $f$ for the trivial $w = 0$. Thus, the constant-Q transform is not frequency-uniform.

In the mel spectrogram, the response is $U_{\text{mel}}^w \delta(t, \log f) = |w(-t)|^2 \cdot \int_{-\infty}^\infty \Lambda_f$, which can be frequency-uniform if the integral over $\Lambda_f$ is independent of $f$. However, for Gaussian $w$, it is not possible to have frequency-uniformity and pitch-invariance at the same time.

3 Proposed representation

**Proposition 3.1** The time-frequency representation given by

$$U_{\text{w}}^*X(t, \log f) = \int_{-\infty}^\infty |\mathcal{F}(X \cdot w_{\nu/f}(\cdot - s))(f)| \cdot w_{\beta}(s) \cdot \sqrt{1 - \frac{f^2}{f_0^2}} \pi(t - s) \, ds, \quad w_\beta(s) = \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{s^2}{2\pi\beta}}$$

is both frequency-uniform and pitch-invariant for $f > f_0 > 0$.

**Proof.** We first check for frequency-uniformity. By adding the variances of the Gaussians, we get:

$$U_{\text{w}}^* \delta(t, \log f) = (w_{\nu/f} * w_{\beta})'(t) = w_1(t).$$

Now we apply the representation to a complex exponential. Via the result from the constant-Q transform, we obtain:

$$U_{\text{w}}^*x_\nu(t, \log f) = |\mathcal{F}w_{\nu}(1-\nu/f)| \cdot \int_{-\infty}^\infty w_{\beta} \sqrt{1 - \frac{f^2}{f_0^2}} \pi(t - s) \, ds = |\mathcal{F}w_{\nu}(1-\nu/f)|.$$

Thus, our representation is also pitch-invariant.

The representation generalizes the first layer of the scattering transform [3]. Independently, a related transform was developed by Dörfler et al. [4], which is equivalent to the mel spectrogram and therefore not frequency-uniform and pitch-invariant at the same time.

4 Conclusion

Our representation combines favorable properties from existing transforms and is thus particularly suitable for the analysis of audio signals with both percussive and melodic components (cf. Fig. 1). However, due to the Heisenberg uncertainty principle, it only reaches down to a certain frequency $f_0$, and the convolution makes its inversion an ill-posed problem.

**Acknowledgements** The first author acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 281474342/GRK2224/1.

**References**

[1] K. Gröchenig, Foundations of Time-Frequency Analysis (Springer, 2001).
[2] J. C. Brown, J. Acoust. Soc. Am. 89(1), 425–434 (1991).
[3] J. Andén and S. Mallat, IEEE Trans. Signal Process. 62(16), 4114–4128 (2014).
[4] M. Dörfler, T. Grill, R. Bammer, and A. Flexer, Neural Comput. Appl. pp. 1–14 (2017).