SCATTERING REPRESENTATION OF MODULATED SOUNDS

Joakim Andén,[*]
CMAP, Ecole Polytechnique
Palaiseau, France
anden@cmap.polytechnique.fr

Stéphane Mallat,[*]
CMAP, Ecole Polytechnique
Palaiseau, France

ABSTRACT

Mel-frequency spectral coefficients (MFSCs), calculated by averaging the spectrogram along a mel-frequency scale, are used in many audio classification tasks. Their efficiency can be partly explained by their stability to deformation in a Euclidean norm. However, averaging the spectrogram loses high-frequency information. This loss is reduced by keeping the window size small, around 20 ms, which in turn prevents MFSCs from capturing large-scale structures. Scattering coefficients recover part of this lost information using a cascade of wavelet decompositions and modulus operators, enabling larger window sizes. This representation is sufficiently rich to capture note attacks, amplitude and frequency modulation, as well as chord structure.

1. INTRODUCTION

Mel-frequency spectral coefficients (MFSCs) are defined as the spectrogram averaged along a mel-frequency scale. They have proven useful in many audio classifiers, which can be partly attributed to their stability to deformation. To explain this concept, we consider an audio signal $x(t)$, deformed using a time-warping $\tau(t)$ to yield $\tilde{x}(t) = x(t - \tau(t))$. A feature representation Φ, mapping x to Φx, is stable to deformation if the Euclidean distance $\|\Phi x - \Phi \tilde{x}\|$ is small for τ small. Since small deformations cause small changes in perception, such distances provide useful measurements in a variety of tasks.

The Fourier transform and the spectrogram are unstable to deformation since high-frequency components are more sensitive to deformation than the in low frequencies. Averaging using a mel scale that is logarithmic above a certain frequency removes this instability.

However, mel-scale averaging loses information in the high frequencies. To reduce this loss the window size is kept small, around 20 ms. MFSCs thus cannot capture large-scale structures. Methods such as modulation spectra [1], correlograms [2], and stabilized auditory images (SAIs) [3] attempt to remedy this. Unfortunately, modulation spectra are calculated using a Fourier transform, so they inherit its instability. Correlograms and SAIs are equally unstable since their high-frequency channels are better resolved in time than their low-frequency channels and thus more sensitive to deformation.

Scattering coefficients [4] recover the information lost in MFSCs while remaining stable through a cascade of wavelet decompositions and modulus operators. Using larger window sizes, larger-scale structures are thus captured. In a music classification task, scattering coefficients performed significantly better than standard MFSC methods [5]. The representation also resembles various neurophysiological models of auditory processing in the brain [6, 7]. In addition, scattering coefficients have proven useful in image classification tasks [8].

Despite discarding information to obtain deformation stability, scattering coefficients succeed in capturing important auditory phenomena such as note attacks, amplitude and frequency modulations, as well as chord structure.

Section 2 describes the deformation stability and information loss properties of MFSCs. In Section 3, this information is partially recovered while remaining stable to deformation using the scattering transform. Section 4 shows that scattering coefficients capture important timbral information in modulated sounds.

A MATLAB software package is available at `http://www.cmap.polytechnique.fr/scattering/`.

2. MEL-FREQUENCY SPECTRAL COEFFICIENTS

MFSCs stabilize the spectrogram to deformation by averaging according to a scale that is logarithmic in the high frequencies. Rewriting this using a wavelet modulus decomposition, the information lost in the average can be seen as the high-frequency temporal variation of the decomposition.

The Fourier transform of x is $\hat{x}(\omega) = \int x(u) e^{-i\omega u} du$. Using a window ϕ such that $\int \phi(t) dt = 1$, we define the windowed Fourier transform as

$$\hat{x}(t, \omega) = \int x(u) \phi(u - t) e^{-i\omega u} du. \tag{1}$$

While informative, the spectrogram $|\hat{x}(t, \omega)|$ is rarely used for classification due to instability in the high frequencies.

To illustrate this, we consider a signal of harmonic structure $x(t) = \sum_n a_n \cos(n\xi t)$ and a deformed version $\tilde{x}(t) = x(t - \tau(t))$, where $\tau(t) = \epsilon t$ is a scaling for small ϵ. As shown in Figure 2, this shifts the frequency component at $n\xi$ by $\epsilon n \xi$ to $(1 - \epsilon) n \xi$. The low frequencies therefore move little compared to the high frequencies, so the Euclidean distance between $|\hat{x}(t, \omega)|$ and $|\hat{\tilde{x}}(t, \omega)|$ is large even for a small ϵ. This instability occurs in any representation based on the Fourier transform.

To remedy this, MFSCs average along the frequency axis, giving

$$Mx(t, \lambda) = \int |\hat{x}(t, \omega)| \hat{\psi}_\lambda(\omega) d\omega, \tag{2}$$

where $\hat{\psi}_\lambda$ is a mel-frequency filter centered at λ. The bandwidth of these filters is constant for $\lambda \leq \omega_0$ and proportional to λ for $\lambda > \omega_0$, where $\omega_0 = 1$ kHz. To cover the frequency domain, MFSCs are calculated for $\lambda \in \Lambda$, where Λ is linearly spaced below ω_0 and logarithmically spaced above. We use Gabor filters to define

$$\hat{\psi}_\lambda(\omega) = \exp\left(-\frac{(\omega - \lambda)^2}{2\sigma^2 Q^{-2} \max(\lambda, \omega_0)^2}\right), \tag{3}$$

[*] This work is funded by the ANR grant 0126 01.

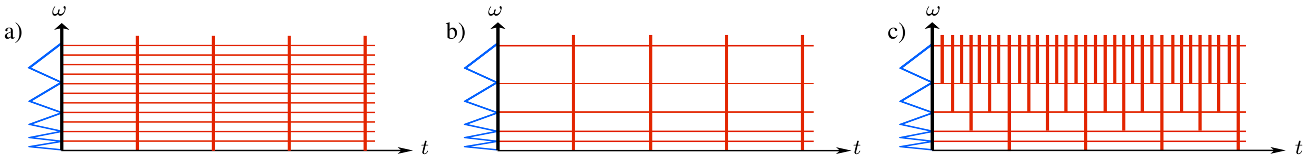

Figure 1: *Time-frequency resolution of a spectrogram $|\hat{x}(t,\omega)|$ in (a), a mel-spectrogram $Mx(t,\lambda)$ in (b), and a scalogram $|x \star \psi_\lambda|$ in (c).*

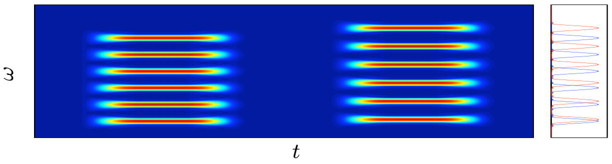

Figure 2: *Spectrogram of a musical note $x(t)$ and a scaled version $\tilde{x}(t) = x(t-\epsilon t)$. The plot to the right shows the Fourier transform of $x(t)$ in blue and $\tilde{x}(t)$ in red.*

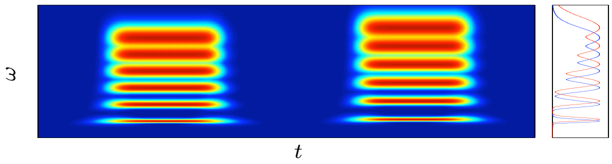

Figure 3: *Mel-spectrogram and mel-frequency plots of the notes $x(t)$ and $\tilde{x}(t)$ in Figure 2 for $Q = 4$.*

where σ is chosen so that λ/Q is the half-power bandwidth for $\lambda > \omega_0$. For the examples in this paper, $Q = 16$. MFSCs thus transform a representation of uniform time-frequency resolution (Figure 1a) into one of varying frequential resolution (Figure 1b).

The mel-spectrograms of $x(t)$ and $\tilde{x}(t)$ are shown in Figure 3. Since the bandwidth of $\hat{\psi}_\lambda$ in (3) is proportional to λ for $\lambda > \omega_0$ and $n\xi$ is shifted by $\epsilon n\xi$, the averaging yields the same overlap for each pair of frequency components. The Euclidean distance between the mel-spectra of the signals is thus proportional to ϵ.

We can rewrite the information loss incurred by frequency-scale averaging as a loss of temporal resolution. The filters in (3) are dilations $\hat{\psi}_\lambda(\omega) = \hat{\psi}(\lambda^{-1}\omega)$ of a mother wavelet $\psi(t)$ for $\lambda > \omega_0$, so we consider $\{\psi_\lambda\}_{\lambda\in\Lambda}$ as a wavelet filter bank. Convolving x with these filters and computing the modulus yields a scalogram $|x \star \psi_\lambda|$. The scalogram is of mel-scale frequential resolution (see Figure 1c), so averaging in time using ϕ yields

$$|x \star \psi_\lambda| \star \phi(t), \tag{4}$$

which is of uniform temporal resolution (see Figure 1b). Since they are both measures of energy in the same time-frequency grid, MFSCs are equivalent to (4).

3. SCATTERING WAVELETS

By calculating a cascade of wavelet decompositions and modulus operators, part of the information lost in MFSCs can be recovered while maintaining deformation stability.

The information lost in (4) consists of the high frequencies of $|x \star \psi_{\lambda_1}|$. These are described by the wavelet coefficients $|x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}$. Since the high frequencies are extracted using wavelets, taking the modulus and averaging using ϕ stabilizes them to deformation, giving

$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi. \tag{5}$$

The high frequencies of $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$ are now lost after convolving with ϕ, so another wavelet transform is computed to recover them, and the process is repeated indefinitely.

For a path of frequencies $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$, we call

$$S[p]x = || \cdots ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \cdots | \star \psi_{\lambda_m}| \star \phi \tag{6}$$

the windowed scattering coefficient of x at p. We define the order of p as its length m. First-order scattering coefficients thus correspond to MFSCs while higher orders provide complementary information.

The maximum temporal bandwidth of the wavelets is proportional to Q/ω_0. To capture larger-scale structure, we thus decrease ω_0 in (3) and enlarge the window ϕ. For a window size of N samples, one can show that the number of filters $|\Lambda|$ is of the order of $Q \log_2 N$. According to (6), the number of mth-order $|\Lambda|^m$ paths is therefore of the order of $(Q \log_2 N)^m$.

The calculation of (6) can be viewed as a cascade of a wavelet modulus propagator U, defined as $Ux = \{x \star \phi, |x \star \psi_\lambda|\}_{\lambda\in\Lambda}$. This cascade is illustrated in Figure 4. Its first layer is Ux, and each successive layer is generated from the previous by applying U to each wavelet modulus coefficient. The output then consists of the coefficients averaged using ϕ from all layers.

We suppose that the filters in (3) and ϕ satisfy

$$1 - \epsilon \le |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{\lambda\in\Lambda} |\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \le 1, \tag{7}$$

for all ω where $0 \le \epsilon < 1$. For $\|Ux\|^2 = \|x \star \phi\|^2 + \sum_{\lambda\in\Lambda} \|x \star \psi_\lambda\|^2$, Plancherel's theorem and the contractivity of the modulus together with (7) show that U is contractive, that is $\|Ux - Uy\| \le \|x - y\|$. If $\epsilon = 0$, the same argument gives $\|Ux\| = \|x\|$. The contractivity of U ensures that the above cascade does not diverge.

This cascade of wavelet decompositions and non-linearities is found in several neurophysiological models which simulate auditory processing in the brain [6, 7]. Similar cascades have also been used to in generative models of auditory textures [9] and in convolutional networks [10]. The mathematical framework presented here can thus provide theoretical insights for these models.

Letting $\mathcal{P}$ be the set of all paths, the windowed scattering representation is $Sx = \{S[p]x\}_{p\in\mathcal{P}}$. Defining a Euclidean norm through $\|Sx\|^2 = \sum_{p\in\mathcal{P}} \|S[p]x\|^2$ on Sx, it can be shown that S is stable to deformation [4].

Since S is a repeated application of U, S is also contractive. For appropriate wavelets, it can also be shown that if U preserves the norm, so does S and $\|Sx\| = \|x\|$ [4].

The sound x can be reconstructed from its scattering representation Sx by inverting U using a phase retrieval method. When reconstructing from first-order coefficients, the quality is acceptable

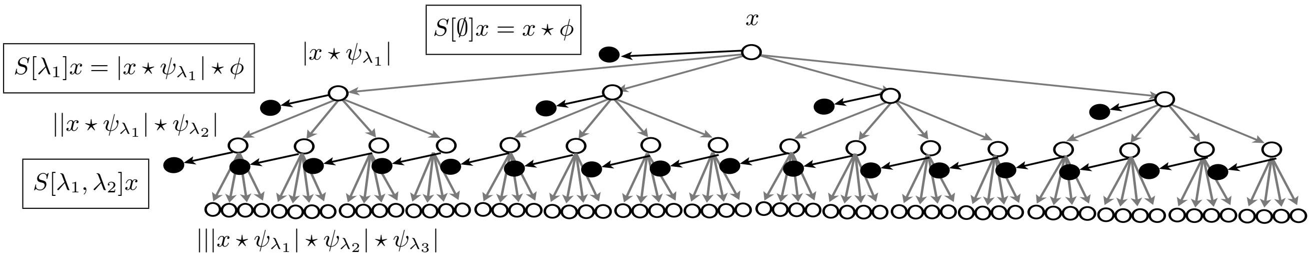
x
$S[\emptyset]x = x \star \phi$
$|x \star \psi_{\lambda_1}|$
$S[\lambda_1]x = |x \star \psi_{\lambda_1}| \star \phi$
$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$
$S[\lambda_1, \lambda_2]x$
$|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|$

Figure 4: *The scattering cascade decomposes an input signal x into its wavelet modulus coefficients $|x \star \psi_{\lambda_1}|$, which are in turn averaged to yield $|x \star \psi_{\lambda_1}| \star \phi$ and redecomposed, giving $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$. The latter are again averaged and redecomposed, furthering the cascade. The output of the cascade is contained in the averages (the boxed expressions).*

for window sizes N such that $N < |\Lambda| = O(Q \log_2 N)$, so $N \lesssim 200$. Second-order coefficients, characterizing larger structures, allow for good reconstruction when $N < |\Lambda|^2 = O([Q \log_2 N]^2)$, hence $N \lesssim 29000$. Examples can be found at `http://www.cmap.polytechnique.fr/scattering/audio/`.

4. MODULATED SOUNDS

While first-order scattering coefficients extract the pitch and filter structures of sounds, second-order coefficients can be shown to characterize different types of modulation.

4.1. Amplitude modulation

Let us consider an impulse train of fundamental frequency ξ_1

$$e(t) = \frac{2\pi}{\xi_1} \sum_n \delta\left(t - \frac{2\pi n}{\xi_1}\right) = \sum_k e^{k\xi_1 t}, \tag{8}$$

filtered by a filter h and modulated in amplitude by a to give

$$x(t) = (e \star h)(t) \cdot a(t). \tag{9}$$

Since h is often highly localized in time, $\hat{h}$ is smooth and so for λ_1 small enough, $\hat{h}(\omega) \approx \hat{h}(\lambda_1)$ on the support of $\hat{\psi}_{\lambda_1}$. Similarly, a is constant on the support of ψ_{λ_1} for λ_1 large enough. When calculating $|x \star \psi_{\lambda_1}|$, we thus get

$$|x \star \psi_{\lambda_1}|(t) \approx |\hat{h}(\lambda_1)||e \star \psi_{\lambda_1}|(t) \cdot a(t). \tag{10}$$

If $\lambda_1 \ll \xi_1 Q$, only one frequency component is contained in the support of $\hat{\psi}_{\lambda_1}$ so $e \star \psi_{\lambda_1}(t) \approx \hat{\psi}_{\lambda_1}(k_1\xi_1)e^{k_1\xi_1 t}$ for some k_1 such that $|k_1\xi_1 - \lambda_1|$ is minimized. As a result, we can approximate $S[\lambda_1]x(t) = |x \star \psi_{\lambda_1}| \star \phi(t)$ by plugging $|e \star \psi_{\lambda_1}|(t) \approx |\hat{\psi}_{\lambda_1}(k_1\xi_1)|$ into (10) and averaging using ϕ, which yields

$$S[\lambda_1]x(t) \approx |\hat{h}(\lambda_1)||\hat{\psi}_{\lambda_1}(k_1\xi_1)|(a \star \phi)(t). \tag{11}$$

As with MFSCs, first-order scattering coefficients describe the filter and pitch structure of h and e, respectively. We can also approximate $S[\lambda_1, \lambda_2]x(t) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t)$ using the same method, giving

$$S[\lambda_1, \lambda_2]x(t) \approx |\hat{h}(\lambda_1)||\hat{\psi}_{\lambda_1}(k_1\xi_1)|S[\lambda_2]a(t), \tag{12}$$

since $S[\lambda_2]a(t) = |a \star \psi_{\lambda_2}| \star \phi(t)$. Hence the second order is proportional to the scalogram of the envelope averaged in time.

Let $e(t)$ be a Gaussian white noise process of unit variance in (9), giving an unvoiced sound. The approximation (10) still holds and we can further decompose $|e \star \psi_{\lambda_1}|(t)$ into $\mathbb{E}(|e \star \psi_{\lambda_1}|) + \varepsilon(t)$. Using the fact that $e \star \psi_{\lambda_1}(t)$ is normally distributed, one can show that $\mathbb{E}(|e \star \psi_{\lambda_1}|) = C\sqrt{\lambda_1}$ where $C = \sqrt{\pi/4}\|\psi\|$ for $\lambda_1 > \omega_0$. When convolving $|x \star \psi_{\lambda_1}|$ with ϕ, it can be shown that $C\sqrt{\lambda_1}a \star \phi(t)$ dominates over $(\varepsilon a) \star \phi(t)$ for large window sizes, provided the support of a is large. We thus have

$$S[\lambda_1]x(t) \approx C|\hat{h}(\lambda_1)|\sqrt{\lambda_1}(a \star \phi)(t). \tag{13}$$

To obtain the second order, we convolve $|x \star \psi_{\lambda_1}|$ with ψ_{λ_2}. The term $C\sqrt{\lambda_1}a \star \psi_{\lambda_2}(t)$ dominates over $(\varepsilon a) \star \psi_{\lambda_2}(t)$ when the variation of a is large with respect to its average. Specifically, one can show that when $\lambda_1|a \star \psi_{\lambda_2}|^2(t) \gg a^2 \star |\psi_{\lambda_2}|^2(t)$, we have

$$S[\lambda_1, \lambda_2]x(t) \approx C|\hat{h}(\lambda_1)|\sqrt{\lambda_1}S[\lambda_2]a(t). \tag{14}$$

We thus have results analogous to the impulse train case.

To illustrate the effect of the envelope in (12) and (14), we consider four sounds. The first two are tones with a smooth and sharp attack, respectively. The third is a tone with a smooth attack and exhibiting a tremolo, during which $a(t) = 1 + \epsilon\cos(\xi_2 t)$ for some $\epsilon < 1$. Finally, the fourth is a white noise modulated by a sharp attack. Figure 5 shows the first-order scattering coefficients for a small window of 20 ms as well as first- and second-order scattering coefficients of window size 256 ms for these four and other sounds.

Despite noticeable differences in timbre, the first-order coefficients $S[\lambda_1]x(t)$ of the first three sounds are similar since the difference in a is averaged out in (11). However, for λ_1 equal to the third partial, 2080 Hz, the second order $S[\lambda_1, \lambda_2]x(t)$ provides a clear distinction between the notes: the sharp attack excites higher frequencies λ_2 than does the smooth attack while the tremolo gives rise to a maximum at $\lambda_2 = \xi_2$. The attack of the white noise in the fourth sound is also captured by the second order.

4.2. Frequency modulation

Frequency modulation, or vibrato, is created by applying a periodic deformation $\tau(t) = \epsilon\cos(\xi_2 t)$ to a pitched source e, such as (8), yielding $\tilde{e}(t) = e(t - \epsilon\cos(\xi_2 t))$. Let $x(t) = \tilde{e} \star h(t)$.

Calculating $|x \star \psi_{\lambda_1}|$, we can move h outside as in (10). Carson's rule on the bandwidth of frequency-modulated sinusoids gives that the bandwidth at λ_1 is of the order of $\epsilon\lambda_1\xi_2$. To ensure that the partials do not overlap, we therefore suppose $\lambda_1 \ll \xi_1/\epsilon\xi_2$. As in the previous case, we also suppose $\lambda_1 \ll \xi_1 Q$ to isolate one partial. For the partial at frequency $k_1\xi_1$ closest to λ_1, the instantaneous frequency is $k_1\xi_1[1 + \epsilon\xi_2\sin(\xi_2 t)]$. If λ_1 is large enough, $\sin(\xi_2 t)$ varies little on the support of ψ_{λ_1}, so

$$|x \star \psi_{\lambda_1}|(t) \approx |\hat{h}(\lambda_1)||\hat{\psi}_{\lambda_1}(k_1\xi_1[1 + \epsilon\xi_2\sin(\xi_2 t)])|. \tag{15}$$

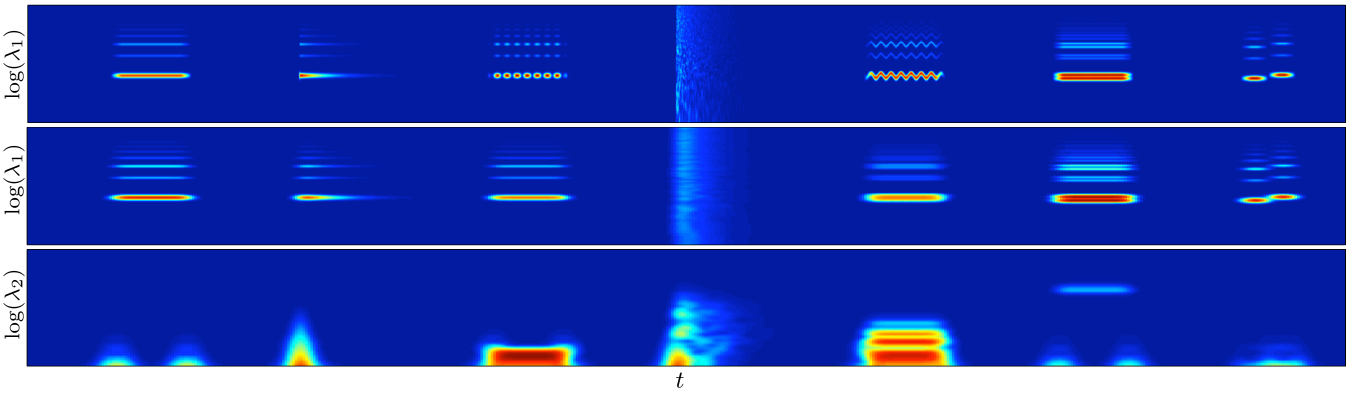

Figure 5: *Top:* $S[\lambda_1]x(t)$ *for small window size* 20 ms *of a regular note, a note with a sharp attack, with tremolo, a white noise with a sharp attack, a note with vibrato, a chord and its arpeggio. Middle:* $S[\lambda_1]x(t)$ *for window size* 256 ms. *Bottom:* $S[\lambda_1, \lambda_2]x(t)$ *for* $\lambda_1 = 2080$ Hz*, corresponding to the third partial of the first three sounds.*

The wavelet modulus coefficient is thus periodic in time, with frequency ξ_2. As a result, its spectrum exhibits a harmonic structure of fundamental frequency ξ_2, which is reproduced in $S[\lambda_1, \lambda_2]\tilde{e}(t)$. For λ_1 fixed, $S[\lambda_1, \lambda_2]\tilde{e}(t)$ is thus characterized by peaks at frequencies $\lambda_2 = k_2\xi_2$ for positive integers k_2.

In Figure 5, the fifth sound exhibits a vibrato, but cannot be distinguished from the first three using the first order since the oscillations in $|\tilde{e} \star \psi_{\lambda_1}|$ have been averaged out by ϕ. However, the second order reveals a harmonic structure in the fourth note that is not present in the others.

4.3. Interference modulation

In the previous examples, the modulation was specified explicitly. Here, we consider the modulations implicit in a chord due to interference between frequency components.

Let us consider $x(t) = 2a_1 \cos(\xi_1 t) + 2a'_1(\xi'_1) \cos(\xi'_1 t)$. Since

$$x \star \psi_{\lambda_1}(t) = e^{i\xi_1 t}\left(a_1\hat{\psi}_{\lambda_1}(\xi_1) + a'_1\hat{\psi}_{\lambda_1}(\xi'_1)e^{i[\xi'_1-\xi_1]t}\right), \quad (16)$$

the wavelet modulus

$$|x \star \psi_{\lambda_1}|(t) = \left|a_1\hat{\psi}_{\lambda_1}(\xi_1) + a'_1\hat{\psi}_{\lambda_1}(\xi'_1)e^{i[\xi'_1-\xi_1]t}\right| \quad (17)$$

is periodic with frequency $|\xi'_1 - \xi_1|$. Thus $S[\lambda_1, \lambda_2]x(t)$ exhibits a harmonic structure in λ_2 of fundamental frequency $|\xi'_1 - \xi_1|$, as described in the previous subsection.

Next we consider the sum of two impulse trains of fundamental frequencies ξ_1 and ξ'_1, a chord. Whenever two partials $n\xi_1$ and $n'\xi'_1$ share the support of $\hat{\psi}_{\lambda_1}$ we obtain an interference modulation at multiples of $|n\xi_1 - n'\xi'_1|$. In addition, for a single note of fundamental frequency ξ_1, we have a modulations at multiples of ξ_1 whenever two partials are in the support of $\hat{\psi}_{\lambda_1}$.

The sixth and seventh sounds in Figure 5 are a two-note chord and its associated arpeggio. In the second order, a modulation at $\lambda_2 = 131$ Hz (the difference in fundamental frequency of the notes) is found for the chord which is not present in the arpeggio.

5. CONCLUSION

The value of MFSCs in audio classification can be partly explained by their stability to deformation which is not found in the spectrogram. To reduce the loss of information at high frequencies, they are calculated on small windows and cannot capture large-scale structures. Scattering coefficients extend MFSCs to recover these high frequencies while maintaining stability to deformation. This allows larger windows and captures timbral structures, such as attacks, amplitude and frequency modulation, as well as chords.

6. REFERENCES

[1] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[2] M. Slaney and R. Lyon, *Visual representations of speech signals*, chapter On the importance of time–a temporal representation of sound, pp. 95–116, M. Cooke, S. Beet and M. Crawford (Eds.) John Wiley and Sons, 1993.

[3] R. D. Patterson, "Auditory images: How complex sounds are represented in the auditory system," *Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 4, pp. 183–190, 2000.

[4] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, to appear, http://arxiv.org/abs/1101.2286.

[5] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proc. of ISMIR*, Miami, Florida, Unites States, Oct. 24-28 2011, pp. 657–662.

[6] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am*, vol. 118, no. 2, pp. 887–906, 2005.

[7] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *J. Acoust. Soc. Am*, vol. 102, no. 5, pp. 2892–2905, 1997.

[8] J. Bruna and S. Mallat, "Invariant scattering convolution network," *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear, http://arxiv.org/abs/1203.1513.

[9] J.H. McDermott and E.P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[10] H. Lee, P. Pham, Y. Largman, , and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. of NIPS*, 2009.