MULTISCALE SCATTERING FOR AUDIO CLASSIFICATION

Joakim Andén
CMAP, Ecole Polytechnique, 91128 Palaiseau
anden@cmap.polytechnique.fr

Stéphane Mallat
CMAP, Ecole Polytechnique, 91128 Palaiseau

ABSTRACT

Mel-frequency cepstral coefficients (MFCCs) are efficient audio descriptors providing spectral energy measurements over short time windows of length 23 ms. These measurements, however, lose non-stationary spectral information such as transients or time-varying structures. It is shown that this information can be recovered as spectral co-occurrence coefficients. Scattering operators compute these coefficients with a cascade of wavelet filter banks and modulus rectifiers. The signal can be reconstructed from scattering coefficients by inverting these wavelet modulus operators. An application to genre classification shows that second-order co-occurrence coefficients improve results obtained by MFCC and Delta-MFCC descriptors. [1]

1. INTRODUCTION

Many speech and music classifiers use mel-frequency cepstral coefficients (MFCCs), which are cosine transforms of mel-frequency spectral coefficients (MFSCs). Over a fixed time interval, MFSCs measure the signal frequency energy over mel-frequency intervals of constant-Q bandwidth. As a result, they lose information on signal structures that are non-stationary on this time interval. To minimize this loss, short time windows of 23 ms are used in most applications since at this resolution most signals are locally stationary. The characterization of audio properties on larger time scales is then done by aggregating MFSC coefficients in time, with multiple ad-hoc methods such as Delta-MFCC [5] or MFCC segments [1]. This paper shows that the non-stationary behavior lost by MFSC coefficients is captured by a scattering transform which computes multiscale co-occurrence coefficients. A scattering representation includes MFSC-like measurements together with higher-order co-occurence coefficients that can characterize audio information over much longer time intervals, up to several seconds. This yields efficient representations for audio classification.

Section 2 relates MFSCs and wavelet filter bank coefficients. It is shown that information lost by spectral energy measurements can be recovered by a scattering operator introduced in [8]. It computes co-occurrence coefficients by cascading wavelet filter banks and rectifiers calculated with modulus operators. A scattering transform has strong similarities with auditory physiological models based on cascades of constant-Q filter banks and rectifiers [4, 10]. It is shown that second-order co-occurrence coefficients carry an important part of the signal information. Section 3 gives an application to musical genre classification, which shows that scattering co-occurence coefficients reduce classification errors obtained with MFCCs and Delta-MFCCs. A MATLAB software is available at http://www.cmap.polytechnique.fr/scattering/.

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

2. SCATTERING REPRESENTATION

2.1 From Mel-Frequency Spectra to Wavelets

To understand the information lost by mel-frequency spectral coefficients, we relate them to a wavelet transform. The Fourier transform of $x(t)$ is written $\hat{x}(\omega) = \int x(u)e^{-i\omega u}du$. A short-time Fourier transform of x is computed as the Fourier transform of $x_{t,T}(u) = x(u)w_T(u-t)$, where w_T is a time window of size T:

$$\hat{x}_{t,T}(\omega) = \int x_{t,T}(u)e^{-i\omega u}du.$$

MFSCs are obtained by averaging the spectrogram $|\hat{x}_{t,T}(\omega)|^2$ over mel-frequency intervals. These intervals have a constant frequency bandwidth below 1000 Hz and a constant octave bandwidth above 1000 Hz. The MFSCs can thus be written

$$M_T x(t,j) = \frac{1}{2\pi}\int |\hat{x}_{t,T}(\omega)|^2 |\hat{\psi}_j(\omega)|^2 d\omega \qquad (1)$$

where each $\hat{\psi}_j(\omega)$ covers a mel-frequency interval indexed by j. Applying Parseval's theorem yields

$$M_T x(t,j) = \int |x_{t,T} \star \psi_j(u)|^2 du. \qquad (2)$$

It results that $M_T x(t, j)$ is the energy of x in a neighborhood of t of size T and in the mel-frequency interval indexed by j. It is unable to capture non-stationary structures of duration shorter than T, which is why T is chosen to be small, typically 23 ms.

At high frequencies, the filters ψ_j are constructed by dilating a single filter ψ whose octave bandwidth is $1/Q$:

$$\psi_j(t) = a^{-j}\psi(a^{-j}t) \text{ with } a = 2^{1/Q} \text{ and } j \leq J. \quad (3)$$

These filters can thus be interpreted as dilated wavelets. The filter ψ is normalized so that its support is about 1 s. It is a complex filter whose transfer function approximately covers the frequency interval $[2Q\pi - \pi, 2Q\pi + \pi]$. For $j < J$, the time support of ψ_j is thus smaller than a^J and it covers the frequency interval $[2Q\pi a^{-j} - \pi a^{-j}, 2Q\pi a^{-j} + \pi a^{-j}]$. Frequencies below $2\pi Q a^{-J}$ are covered by P filters ψ_j (for $J \leq j < J + P$), having the same frequency bandwidth as ψ_J, which is $2\pi a^{-J}$, and a time support equal to a^J. Although these low-frequency filters are not dilations of ψ, for the sake of simplicity we shall still call them wavelets. The resulting wavelet transform is a filter bank defined by:

$$W_J x(t) = \begin{pmatrix} x \star \phi_J(t) \\ x \star \psi_j(t) \end{pmatrix}_{j<J+P}.$$

The first filter ϕ_J is a low-pass filter covering the interval $[-\pi a^{-J}, \pi a^{-J}]$, which is not covered by other wavelet filters and whose temporal support is about a^J.

Wavelet filters are designed so that for all frequencies ω

$$1-\epsilon \leq |\hat{\phi}_J(\omega)|^2 + \frac{1}{2}\sum_{j<J+P} |\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2 \leq 1 \quad (4)$$

for a small ϵ. The squared norm of a signal is written $\|x\|^2 = \int |x(t)|^2 dt$ and the norm of its wavelet transform is defined by:

$$\|W_J x\|^2 = \|x \star \phi_J\|^2 + \sum_{j<J+P} \|x \star \psi_j\|^2.$$

Thus by applying Parseval's theorem one can verify that the filter admissibility condition (4) implies that

$$(1-\epsilon)\|x\|^2 \leq \|W_J x\|^2 \leq \|x\|^2.$$

The wavelet filter bank is thus contractive and if $\epsilon = 0$, it is also unitary. This energy equivalence also implies that x can be recovered from its wavelet transform.

In numerical applications we use Gabor filters $\psi(t) = \theta(t)e^{i2\pi Qt}$ where θ is Gaussian, with $Q = 16$ and $P = 23$, which satisfy (4) for $\epsilon = 0.02$. The resulting filter bank is shown in Figure 1.

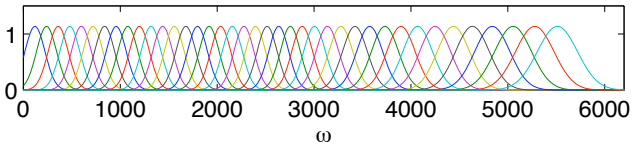

Figure 1. Wavelet filter bank of Gabor filters at sampling frequency 11025 Hz.

2.2 Scattering Wavelets

An MFSC coefficient $M_T x(t, j)$ in (2) gives the squared energy of wavelet coefficients at the scale a^j, over a time interval of size T around t. Let us choose the maximum wavelet scale to be $a^J = T$. The square does not play an important role on the derived MFCC audio descriptors which are calculated with a logarithm. Replacing the squared amplitude by the amplitude yields similar measurements which can be computed directly by averaging the wavelet coefficient amplitudes of x:

$$|x \star \psi_j| \star \phi_J(t). \quad (5)$$

This measures the signal amplitude in the frequency interval covered by ψ_j, averaged over a neighborhood of t of duration $T = a^J$. The larger T, the more information is lost by this averaging.

To recover the information lost by averaging, observe that $|x \star \psi_{j_1}| \star \phi_J$ can be written as the low-frequency component of the wavelet transform of $|x \star \psi_{j_1}|$:

$$W_J|x \star \psi_{j_1}|(t) = \begin{pmatrix} |x \star \psi_{j_1}| \star \phi_J(t) \\ |x \star \psi_{j_1}| \star \psi_{j_2}(t) \end{pmatrix}_{j_2<J+P}.$$

Since the wavelet transform is invertible, the information lost by the convolution with ϕ_J is recovered by the wavelet coefficients $|x \star \psi_{j_1}| \star \psi_{j_2}(t)$. Averaged measurements are obtained with a low-pass filtering of the modulus of these complex wavelet coefficients:

$$||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t). \quad (6)$$

These provide co-occurrence information at the scales a^{j_1} and a^{j_2}. Such coefficients are called scattering coefficients because they compute the interferences of the signal x with two successive wavelets ψ_{j_1} and ψ_{j_2}. They measure the amplitude of time variations of $|x \star \psi_{j_1}(t)|$ in the frequency intervals covered by the wavelets ψ_{j_2}. Figure 2 shows first-order scattering coefficients of a musical recording sampled at 11025 Hz, calculated with $T = 800$ ms. Co-occurrence coefficients $||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t)$ are shown in Figure 2, for a fixed scale a^{j_1}.

Averaging $||x \star \psi_{j_1}| \star \psi_{j_2}|$ by ϕ_J in (6) again entails a loss of high frequencies, which can be recovered by a new wavelet transform. The same procedure is thus iterated,

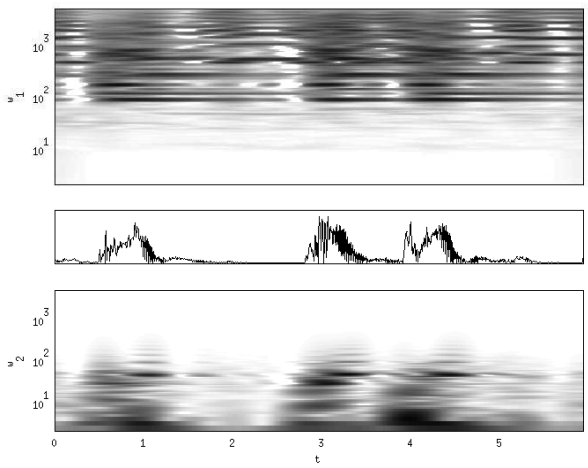

Figure 2. Top: $\log[||x \star \psi_{j_1}| \star \phi_J(t)]$ as a function of time t and of $\omega_1 = 2\pi Q a^{-j_1}$ for $T = a^J = 800$ ms. Middle: graph of $|x \star \psi_{j_1}|$ for $\omega_1 = 855$ Hz. Bottom: $\log[||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t)]$ as a function of t and of $\omega_2 = 2\pi Q a^{-j_2}$ for $|x \star \psi_{j_1}|$ shown above.

defining a cascade of filter banks and modulus operators illustrated in Figure 3.

Let U_J be the wavelet modulus operator which computes the modulus of complex wavelet coefficients while keeping the phase of $x \star \phi_J$:

$$U_J x(t) = \begin{pmatrix} x \star \phi_J(t) \\ |x \star \psi_j(t)| \end{pmatrix}_{j<J+P}. \quad (7)$$

A scattering transform first computes $U_J x$ and outputs the low-frequency signal $x \star \phi_J$. At the next layer, each $|x \star \psi_{j_1}|$ is retransformed by U_J, which outputs $|x \star \psi_{j_1}| \star \phi_J$ and computes $||x \star \psi_{j_1}| \star \psi_{j_2}|$. These coefficients are themselves again transformed by U_J, which outputs $||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J$ and computes third-order wavelet signals, which are further subdecomposed by U_J, and so on.

Applying this transformation m times and discarding the coefficients not filtered by ϕ_J yields a scattering vector of size $m+1$ at time t:

$$S_J x(t) = \begin{pmatrix} x \star \phi_J(t) \\ |x \star \psi_{j_1}| \star \phi_J(t) \\ ||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t) \\ \vdots \\ ||\cdots|x \star \psi_{j_1}|\cdots| \star \psi_{j_m}| \star \phi_J(t) \end{pmatrix}_{j_1, j_2, \ldots < J+P}$$

This scattering transform is a cascade of modulated filter banks and non-linear rectifications, as in the auditory physiological models studied in [4, 10]. It has an architecture similar to convolutional networks used in computer vision [6] and to convolutional deep belief networks used in audio classification [7]. However, a scattering gathers outputs from all layers as opposed the last one. Indeed, the energy of coefficients of order q decays to zero when q increases.

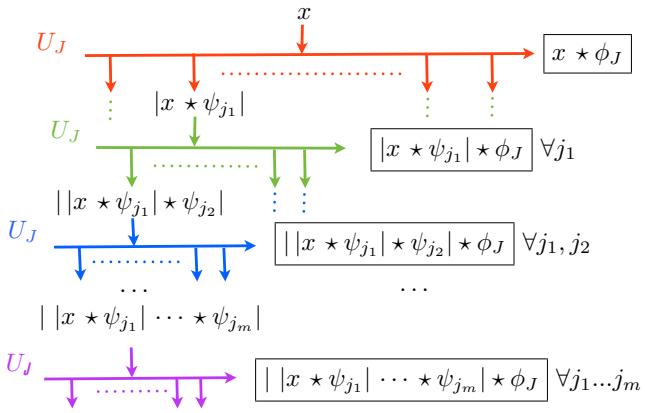

Figure 3. A scattering operator is a cascade of wavelet modulus operators U_J. It outputs convolutions with ϕ_J shown in boxes.

The squared norm of this scattering signal is the sum of the squared norms of its components:

$$\|S_J x\|^2 = \sum_q \sum_{j_1, \ldots, j_q < J+P} \| \, ||x \star \psi_{j_1}| \cdots \star \psi_{j_q}| \star \phi_J \|^2.$$

Since W_J and the modulus are both contractive operators, the wavelet modulus operator U_J is also contractive. Because S_J is calculated with a cascade of U_J, it remains contractive, and thus for any signals x and y

$$\|S_J x - S_J y\| \leq \|x - y\| .$$

The wavelet transform is unitary if the wavelet filters satisfy the admissibility condition (4) with $\epsilon = 0$. For wavelets satisfying this and additional criteria, it is proved in [8] that the energy of all scattering coefficients of order q decays to zero as q increases. It results that the whole signal energy is carried by a scattering vector consisting of co-occurrence coefficients of all orders from $q = 0$ to $q = \infty$:

$$\|S_J x\| = \|x\| .$$

Table 1 gives the average value of $\|S_J x\| / \|x\|$ over all audio signals x in the GTZAN dataset, sampled at 11025 Hz, as a function of m and T. For $m = 0$, $S_J x(t) = f \star \phi_J(t)$. Observe that for $T \leq 6$ s, first- and second-order coefficients carry more than 98% of the energy.

2.3 Second-Order Scattering Decomposition and Reconstruction

In the following, the scattering transform is computed for $m = 2$ because first- and second-order scattering coefficients carry most of the signal energy in the interesting range

T	$m=0$	$m=1$	$m=2$	$m=3$
23 ms	23.7%	98.9%	99.6%	99.6%
93 ms	1.9%	97.7%	99.4%	99.4%
370 ms	1.2%	92.7%	99.3%	99.4%
1.5 s	1.0%	82.0%	98.9%	99.3%
5.9 s	0.99%	73.0%	98.1%	99.1%
22 s	0.97%	67.5%	96.5%	99.0%

Table 1. Averaged ratio $\|S_J x\|/\|x\|$ on the GTZAN dataset, as a function of the maximum scattering order m and of $T = a^J$.

of window sizes T. The signals $|f \star \psi_{j_1}| \star \phi_J(t)$ and $||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t)$ are uniformly sampled at intervals $T = a^J$ because the frequency bandwidth of $\hat{\phi}_J$ is $2\pi a^{-J}$. A sampled second-order scattering vector is thus defined by:

$$S_J x(na^J) = \begin{pmatrix} |x \star \psi_{j_1}| \star \phi_J(na^J) \\ ||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(na^J) \end{pmatrix}_{j_1, j_2 < J+P}. \quad (8)$$

We now show that if $j_2 < j_1 + \log_a Q/2$ then $||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t) \approx 0$, so second-order coefficients need only be calculated for $j_2 \geq j_1 + \log_a Q/2$. Indeed, since $\psi(t) = \theta(t) e^{i2\pi Qt}$, it results that

$$|x \star \psi_{j_1}(t)| = |x_{j_1} \star \theta_{j_1}(t)| \text{ with } x_{j_1}(t) = x(t) e^{-i2\pi Q a^{-j_1} t}.$$

The Fourier transform of $|x \star \psi_{j_1}(t)|$ is thus approximately located in the low-frequency interval covered by $\hat{\theta}_{j_1}$ where $\theta_j(t) = a^{-j}\theta(a^{-j}t)$. One can verify that if $j_2 < j_1 + \log_a Q/2$ then the supports of $\hat{\psi}_{j_2}$ and $\hat{\theta}_{j_1}$ barely overlap, which implies that $||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t) \approx 0$. Non-zero scattering coefficients (8) are computed with the following algorithm.

Algorithm 1 Second-order scattering calculations

for $j_1 < J + P - 1$ **do**
  Compute $||f \star \psi_{j_1}(a^{j_1} n)| \; \forall n$
  Output $||f \star \psi_{j_1}| \star \phi_J(a^J n) \; \forall n$
  for $j_2 = j_1 + \log_a(Q/2)$ to $J + P - 1$ **do**
    Compute and output $||f \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(a^J n) \; \forall n$
  end for
end for

An audio frame of duration $T = a^J$ containing N samples yields about $Q \log_2(N/Q)$ and $Q^2/2 \log_2^2(N/Q^2)$ first-order and second-order scattering coefficients, respectively. If $N = 8192$, there are 150 first-order coefficients and 5500 second-order coefficients, approximately. Using FFTs, these coefficients are computed with $O(N \log(N/Q))$ operations.

Since the scattering transform is computed by iterating the wavelet modulus operator U_J, its inversion is reduced to inverting U_J. The wavelet transform W_J is invertible with a stable inverse but U_J loses the complex phase of wavelet coefficients. Inverting U_J then amounts to retrieving the complex phase from the modulus information. A surprising new result [12] proves that for appropriate wavelets, the operator U_J is invertible and that its inverse is continuous, which is a weak stability result. This inversion is made possible because of the redundancy of wavelet signals $x \star \psi_j(t)$, which can be exploited with a reproducing kernel projector. Numerical reconstructions are computed with an alternating projection algorithm, which alternates between a projector on the modulus constraint and the wavelet transform reproducing kernel projector [12]. However, this algorithm does not compute the exact inverse of U_J because it is a non-convex optimisation which can be trapped in local minima.

Even though U_J is invertible, x cannot be recovered exactly from $S_J x$ calculated at a finite order m because all scattering coefficients of order larger than m are set to 0. For $T \leq 100$ ms most of the audio signal energy is concentrated in first-order coefficients according to Table 1 and the reconstruction from these first-order coefficients (which correspond to MFSCs) is indeed of good audio quality. As T increases, reconstructions from first-order coefficients progressively lose more information on transient structures and lose all melodic structures for $T \geq 3$ s. Second-order coefficients recover this transient information and fully restores melodic structures when $T = 3$ s. Reconstruction examples are available at `http://www.cmap.polytechnique.fr/scattering/audio/`.

2.4 Cosine Log-Scattering

MFCC coefficients are computed as a cosine transform of the logarithm of MFSC coefficients. Indeed, many musical and voiced sounds can be approximated by an excitation $e(t)$ filtered by resonator corresponding to a filter $h(t)$: $x(t) = e \star h(t)$ [2]. MFCCs separate h from e with a logarithm and a discrete cosine transform (DCT). The same property applies to scattering coefficients, which are therefore retransformed with a logarithm and a DCT.

The impulse response $h(t)$ is typically very short so $\hat{h}(\omega)$ is a regular function of ω. Supposing that $\hat{h}(\omega)$ is nearly constant over the frequency support of $\hat{\psi}_{j_1}$, one can verify that

$$x \star \psi_{j_1}(t) \approx \hat{h}(2\pi Q a^{-j_1}) \cdot e \star \psi_{j_1}(t). \quad (9)$$

It results that

$$\log |x \star \psi_{j_1}| \star \phi_J(t) \approx \log |\hat{h}(2\pi Q a^{-j_1})| + \log \left[|e \star \psi_{j_1}(t)| \star \phi_J(t) \right]. \quad (10)$$

Since $|\hat{h}(\omega)|$ is a regular function of ω, $\log |\hat{h}(2\pi Q a^{-j_1})|$ is also a regular function of j_1 whereas this is typically false

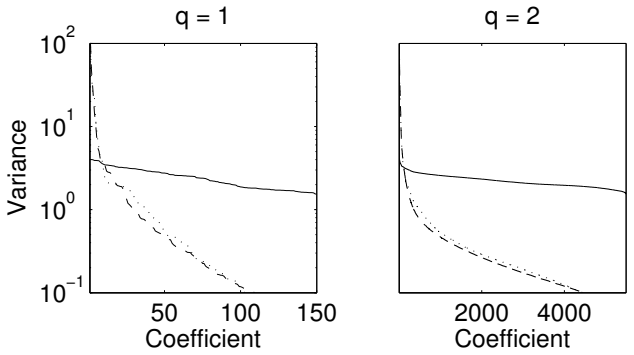

Figure 4. Variances, in decreasing order, of log-scattering coefficients in different bases for $q = 1$ and $q = 2$ computed on GTZAN for $T = 1.5$ s. Solid curve: Variance of log-scattering coefficients. Dashed curve: Variance of a PCA basis computed on log-scattering coefficients. Dotted curve: Variance of cosine log-scattering coefficients.

for $|e \star \psi_{j_1}(t)|$. Both components can thus be partially separated with a DCT along j_1, which carries the information depending on h over to low-frequency DCT coefficients.

Similarly, (9) implies

$$||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t) \approx |\hat{h}(2\pi Q a^{-j_1})| \cdot |e \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t),$$

and hence

$$\begin{aligned} \log\left[||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t)\right] &\approx \log|\hat{h}(2\pi Q a^{-j_1})| \\ &+ \log\left[|e \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t)\right]. \end{aligned}$$

These coefficients are transformed with a DCT along j_2 and then along j_1, yielding a representation parametrized by k_2 and k_1 respectively. The first term, depending only on j_1, only contributes to the zero DCT coefficient ($k_2 = 0$) along j_2. The second DCT along j_1 separates the remaining low-frequency components along j_1 from high-frequency ones.

Figure 4 indicates that the DCTs efficiently decorrelate log-scattering coefficients and concentrate the energy over fewer coefficients. Variances were calculated for $q = 1$ and $q = 2$ on part of the GTZAN dataset in three bases: standard log-scattering (solid), a PCA basis computed on another part of the dataset (dashes), and the DCT basis (dotted). The PCA basis decorrelates the log-scattering coefficients and since the variances in the DCT basis closely follow those in PCA basis, the DCT basis decorrelates them as well.

For classification, the final representation using cosine log-scattering (CLS) coefficients is obtained by keeping only the low-frequency DCT coefficients as with MFCCs. For $q = 1$, the first a_1 coefficients are retained. When $q = 2$, a square defined by $k_1 < a_1$ and $k_2 < a_2$ is selected. This adds a_2 bands of information on the non-stationary part corresponding to the coefficients in $q = 1$. In addition, for $k_1 < b_1$, where $b_1 \ll a_1$ (capturing the spectral outline), $b_2 \gg a_2$ bands are included instead of a_2 to better model the time-varying aspects of the spectral shape (e.g. the filter h mentioned). For the numerical results presented in this paper, we have $a_1 = 100$, $b_1 = 10$, $a_2 = 2$ and $b_2 = 10$ (chosen so that classification errors do not differ from the uncompacted representation for relevant scales). The size of the representation is then at most 100 coefficients for $m = 1$ and 380 coefficients for $m = 2$. For $m = 1$, this is larger than the standard MFCC vector of 20 coefficients when $T = 23$ ms since the compacitification is optimized for all scales and smaller scales need less coefficients.

3. CLASSIFICATION

Music and speech classification algorithms are often based on MFCCs computed over 23 ms time windows. To capture longer-range structures, these MFCCs are either aggregated in segments [1] that cover longer time intervals or are complemented with other features such as Delta-MFCCs [5]. Sophisticated GMM, HMM, AdaBoost, sparse coding classifiers have been developed on such feature vectors to optimize audio classification. The next section studies classifications results obtained with simple classifiers to concentrate on the properties of feature vectors as opposed to a specific classifier.

3.1 Musical Genre Classification

The performance of MFCC and log-scattering vectors are compared for musical genre classification, on the GTZAN genre database [11]. This database includes 10 genres, each containing 100 clips of 30 seconds each.

Delta-MFCC coefficients [5] are defined as the difference between MFCC coefficients of two consecutive audio frames and thus cover a time interval of twice the size. These complement the ordinary MFCCs, providing information on the temporal audio dynamics over longer time intervals. The classification performances of feature vectors are evaluated with an SVM classifier computed with a Gaussian kernel $k(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2)$ or an affine space classifier.

Each audio track is decomposed in frames of duration T which are represented using MFCCs, Delta-MFCCs, or cosine log-scattering. A multi-class SVM is implemented over the audio frames with a 1vs1 approach which trains an SVM to discriminate each pair of classes. To classify a whole track, each frame is classified using the SVM and the class with the largest number of frames in the track is selected. The Gaussian kernel parameter γ and the SVM slack variable C are optimized with a cross-validation on a subset of the training set.

Due to the large number of training examples available for small window sizes, training an SVM in these circum-

T/classifier	0.023 s/PCA	0.19 s/PCA	1.5 s/SVM
MFCC	46	36	28
Delta-MFCC	37	33	26
CLS, $m = 1$	46	36	28
CLS, $m = 2$	34	23	18

Table 2. Error rates (in percent) on GTZAN using five-fold cross-validation for different window sizes (T) and features.

stances is infeasible. Therefore, we also compare the performance of the features using an affine space classifier [3] which uses a PCA to create an affine space approximation for each class and then assigns a given track to the class whose affine space model best approximates the feature vector.

The results of five-fold cross-validation on the GTZAN dataset are shown in Table 2. As expected, the error rates for MFCC and first-order CLS are close since they measure similar quantities. Second-order CLS vectors achieve significantly higher accuracy since they recover lost non-stationary structure of the signal. Delta-MFCC perform better than regular MFCCs, but are outperformed by CLS vectors which provide richer representations. With increasing T, the error decreases as larger-scale musical information is encoded, yielding the lowest error of 18% for $T = 1.5$ s with an SVM. At larger time scales, however, classification suffers since even second-order CLS vectors are unable to accurately represent the signal, as seen during reconstruction. Incorporating third-order scattering coefficients ($m = 3$) marginally improves the classification results while greatly increasing the computational load.

State-of-the-art results on GTZAN are obtained with classifiers better adapted than SVMs. These classifiers can also be applied to CLS vectors to improve classification results. With MFCCs on 23 ms and other local features, an AdaBoost classifier yields an error of 17% in [1]. The cascade filter bank of cortical representations in [10], which is similar to a scattering representation, yields an error of 7.6% [9] with a sparse coding classifier.

4. CONCLUSION

Scattering representations are shown to provide complementary co-occurence information which refines MFCC descriptors. We demonstrated that second-order scattering coefficients can bring an important improvement over MFCCs for classification. The ability to characterize non-stationary signal structures opens the possibility to discriminate more sophisticated phenomena such as transients, time-varying filters and rhythms with co-occurrence scattering coefficients, which is not possible with MFCCs. It opens a wide range of applications for music and speech signal processing.

5. REFERENCES

[1] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate Features and AdaBoost for Music Classification," *Machine Learning,* Vol. 65, No. 2-3, pp. 473–484, 2006.

[2] J. Brown: "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, Vol. 105, No. 3, pp. 1933–1941, 1999.

[3] J. Bruna and S. Mallat: "Classification with Scattering Operators," *Proc. of CVPR*, 2011.

[4] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, Vol. 102, No. 5, pp. 2892–2905, 1997.

[5] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 1, pp. 52–59, 1986.

[6] Y. LeCun, K. Kavukvuoglu, and C. Farabet: "Convolutional Networks and Applications in Vision," *Proc. of ISCAS*, 2010.

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

[8] S. Mallat: "Group Invariant Scattering," http://arxiv.org/abs/1101.2286, to appear in *Communications in Pure and Applied Mathematics*.

[9] Y. Panagakis, C. Kotropoulos, and G. Arce, "Music Genre Classification Using Locality Preserving Non-Negative Tensor Factorization and Sparse Representations," *Proc. of the International Society for Music Information Retrieval*, 2009.

[10] T. Chi, P. Ru, and S. Shamma: "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, Vol. 118, No. 2, pp. 887–906, 2005.

[11] G. Tzanetakis and P. Cook: "Musical Genre Classification of Audio Signals," *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293–302, 2002.

[12] I. Waldspurger and S. Mallat: "Wavelet and Scattering Phase Retrieval", CMAP Tech. Report, http://www.cmap.polytechnique.fr/scattering/, 2011.