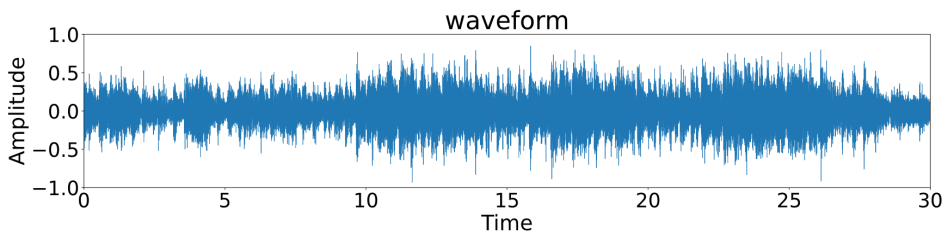
Topological Fingerprints for Audio ID

W. REISE, X. FERNÁNDEZ, etc.

(SPOTIFY, (OX)
EPFL)

Spectral Representations of Audio

An audio recording is the measurement of the pressure fluctuations induced by a sound wave in the vicinity of a microphone, which is represented as a continous real valued function $s: [0, T] \to \mathbb{R}$ over a time interval (i.e., the waveform).



Advantage: Loudness of the sound is related to the amplitude of the signal.

Drawback: Other audio features such as pitch are linked to frequency changes, which cannot be clearly discerned in the waveform representation.

M. MÜLLER, Fundamentals of music processing: Audio, algorithms, applications.

The short-time Fourier transformation (STFT) provides a decomposition for the time-varying frequency and phase content of an audio signal. Give $t \in [0, T]$ and a frequency $f \geq 0$, the continous STFT of the waveform s represents the amplitude of f over a window around t. The STFT is computed as

$$S(f, t) = \int_{\mathbb{R}} s(\tau)\, w(\tau - t) \exp(-i f \tau)\, d\tau,$$

where $w(t)$ is a window function, which is typically a bell-shaped function centered at zero with finite support.

In digital audio processing (DAP), the signal is a finite collection of samples $(s_n)_{n=1}^N$ of s at equally-spaced time points $(t_n)_{n=1}^N$ in $[0, T]$. The size of the sample N is Tf_s, where f_s is the sampling rate. A spectral representation of $(s_n)_{n=1}^N$ can be obtained via the discrete STFT. Give a discretization of the frequency change $\{f_m\}_{m=1}^M$, the magnitude of the frequency f_m around t_n is

$$\hat{S}(n, m) = \sum_{k=-\infty}^{\infty} s_k w_{k-n} \exp(-ikf_m),$$

where $(w_k)_{k=0}^{N_w-1}$ is a discrete version of a window function, N_w is the size of the window.

JULIUS O. SMITH, Spectral Audio Signal Processing.

The Spectrogram of $(S_n)_{n=1}^{N}$ is a matrix $S \in \mathbb{R}^{N \times M}$ whose entries contain the absolute magnitude of the spectral decomposition

$$S_{n,m} = |\hat{S}(n,m)|$$

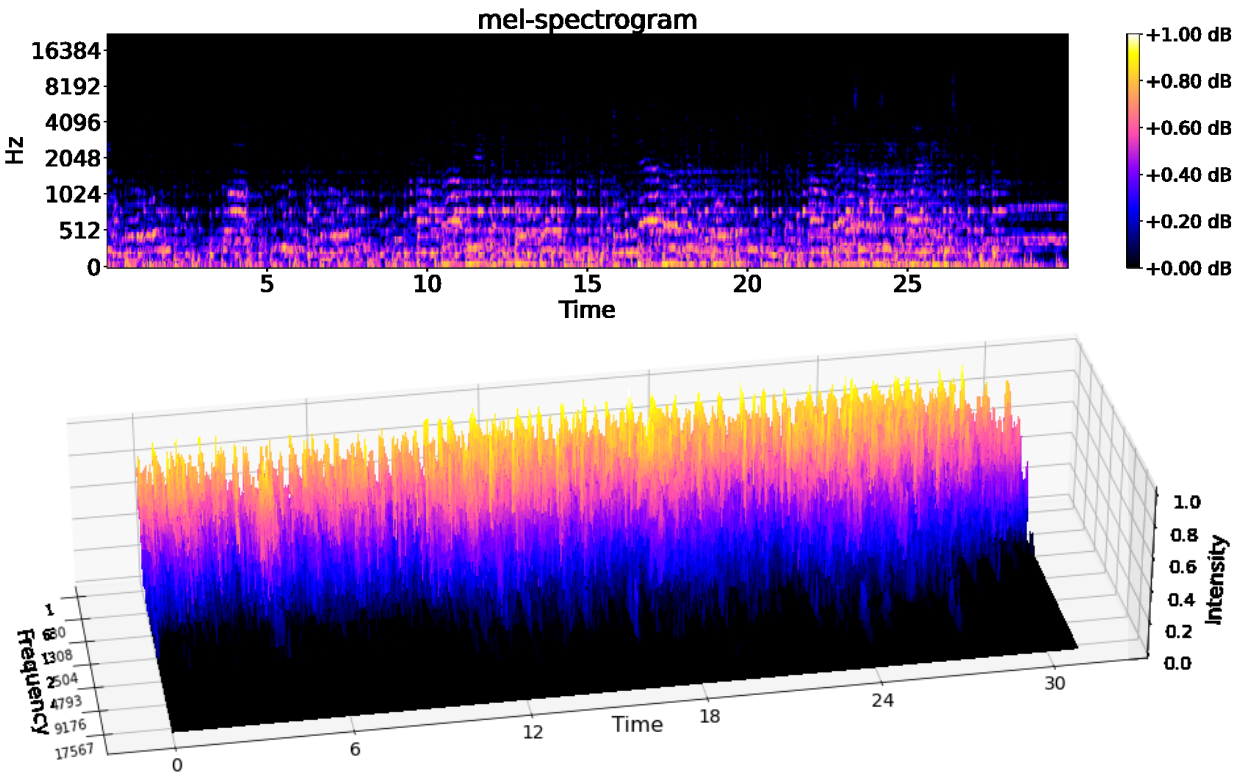
The value of $S_{n,m}$ can be interpreted as the loudness of a pitch with frequency f_m around time t_n in the audio signal.

Note that: the human ear has a logarithmic frequency resolution rather than linear, the frequency and amplitude scales of spectrograms for audio processing applications use the mel-scale* and the Decibel-scale respectively.

This spectral representation of an audio signal is called the Mel-spectrogram

* A Scale for the Measurement of the Psychological Magnitude Pitch

A typical visual representation of a spectrogram is as a heatmap or as a 3D surface.



Topological Fingerprints for Audio ID

The relationship between auditory features and their visual patterns in mel-spectrograms can be analyzed using image processing techniques in the problem of audio identification

Problem: The image representation of an audio signal is not invariant under common audio obfuscations.

Solution: Associate to each image a local low-dimensional representation known as a fingerprint.[1]

Example: In Shazam[2], a fingerprint is the relative position of local maxima in the spectrogram.

Ad/Disad: allow ID under "rigid" obfuscations such as noise or reverb, but sensitive to "topological deformations" such as stretching or pitch-shifting

(1) M. COVELL. Audio Fingerprinting: Combing CV & Data Stream Processing
(2) A. L.-C. WANG, An industrial-Strength Audio Search Algorithm.

Topology of Spectrograms

Cubical complexes: A cube $Q = I_1 \times \dots \times I_d$ is a product of elementary intervals $I_1, \dots, I_d$ of the form $[a, a]$, $[a, a+1]$, $a \in \mathbb{Z}$. We say that

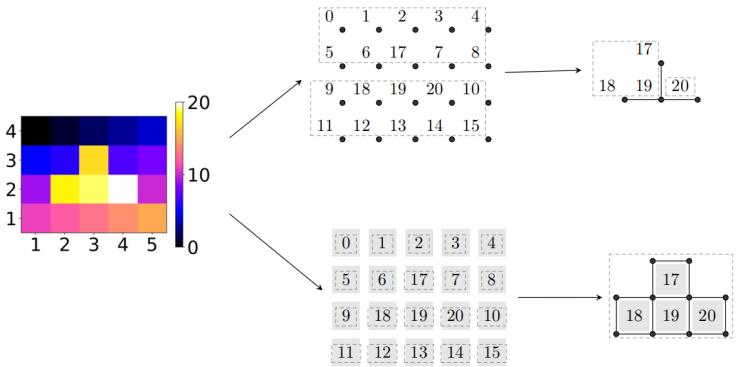
- d is the embedding number of Q
- $\dim(Q) = \left| \{ l \mid I_l \text{ is not degenerate} \} \right|$
- Q is called a vertex if $\dim(Q) = 0$.

Cubes have geometric faces. A cube Q' is said to be a face of a cube Q if $Q' \subset Q$. Moreover, if $\dim(Q') = \dim(Q) + 1$, Q' is a proper face of Q.

Def Let K be a collection of cubes of the same embedding dimension. Then K is a cubical complex if

- for any cube $Q \in K$, its faces are also in K
- for any cubes $Q_1, Q_2 \in K$, the intersection $Q_1 \cap Q \in K$ is either empty or a face of Q_1 and Q_2.

There are two concurring ways of representing an image as a cubical complex: the Top-construction[1] and the Vertex-construction[2]



V and T-constructions. On the left, a grayscale image. In the vertex construction, at the top, each pixel from the image is a 0-cube, while in the top-cell construction, at the bottom, it is a 2-cube. The top-right figure illustrates the super-level set K^{17}. The 1 and 2 cubes were added to the 0 cubes on the figure in the centre and assigned filtration values according to the upper star co-filtration construction. An analogous process is shown on the bottom figure, where 0 and 1 cubes were added.

(1) M. ZEPPELZAUER CV for Music ID

(2) V. ROBINS Percolating length scales from topological persistence analysis of micro-CT images of porous material

def Let $f : V(K) \to \mathbb{R}$ be a function defined on the vertices $V(K)$ of a cubical complex K. The **lower-star filtration** associated to f is $K_f = (K_s)_{s \in \mathbb{R}}$, where

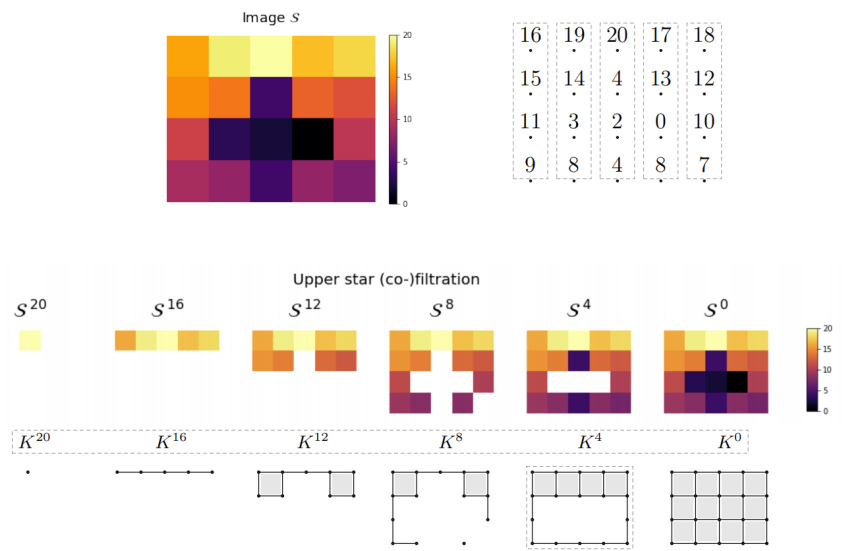
$$K_s = \{Q \subset K \mid f(v) \leq s, \forall v \in V(Q)\}$$

The **upper star co-filtration** $(K^s)_{s \in \mathbb{R}}$ is defined analogously but reversing the order

$$K^s = \{Q \subset K \mid f(v) \geq s, \forall v \in V(Q)\}$$

Given a spectrogram S and an intensity i, we associate a (2-dimensional) cubical complex K^i as follows:

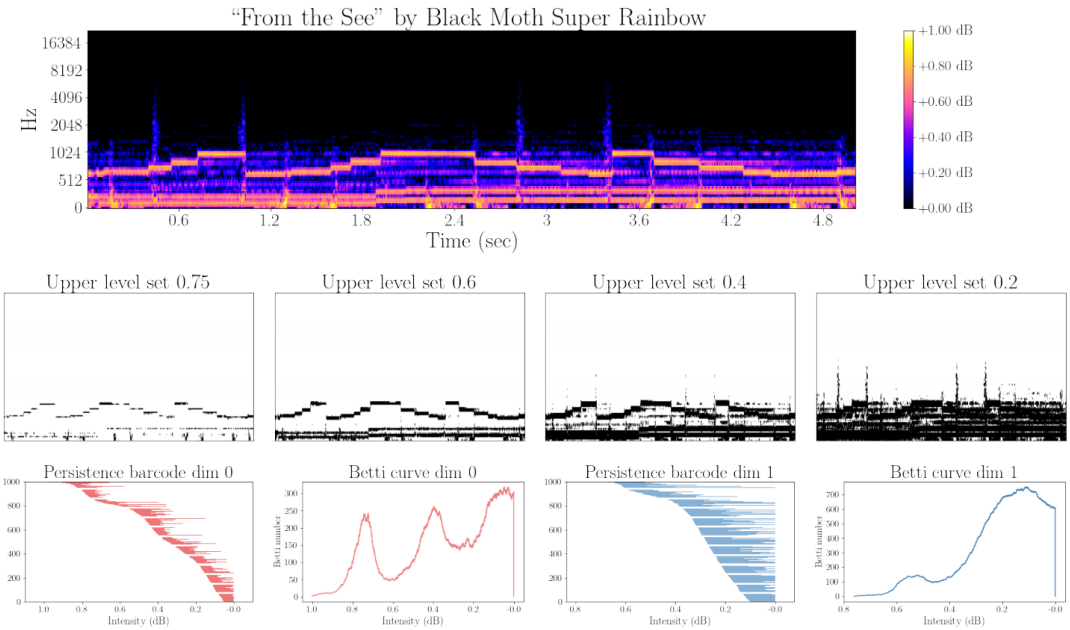
- the vertices (or 0-cubes) of K^i are the pixels in the spectrogram whose intensity value is greater than or equal to i,
- the edges (or 1-cubes) join every pair of vertices in K^i associated to adjacent pixels in S,
- the 2-cubes fill every set of four vertices $\{(m,n), (m+1,n), (m,n+1), (m+1,n+1)\}$ in K^i.

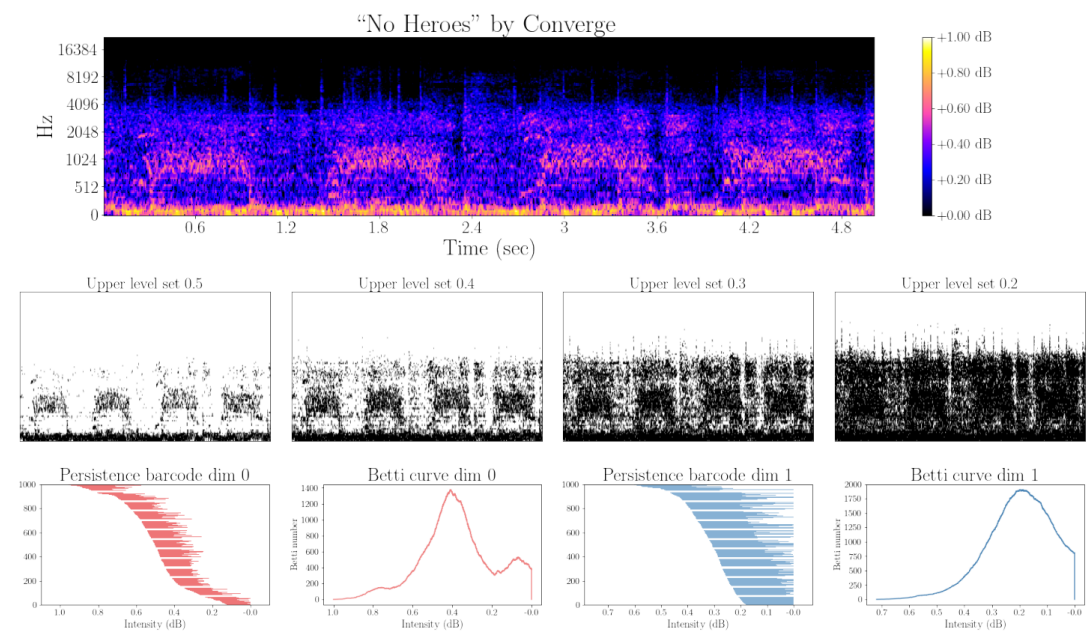


Filtered cubical complex. *Top:* An image $\mathcal{S}$ representing a small patch of a mel-spectrogram, and the intensity values for every pixel. *Bottom:* The upper-star co-filtration of the image $\mathcal{S}$ and the cubical vertex construction on $\mathcal{S}$ for some values of the intensity parameter.

Given a spectrogram S, we compute the PH of the upper star co-filtration of the cubical construction on S. Different audio features, such as melodies at specific intensities, can be interpreted in the topological fingerprints, such as peaks in the Betti curves.

Example: Two 5-seconds extracts of different tracks and how their marked audio styles are displayed on their associated barcodes and Betti curves.


"From the See" by Black Moth Super Rainbow
16384
8192
4096
2048
1024
512
0
Hz
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time (sec)
+1.00 dB
+0.80 dB
+0.60 dB
+0.40 dB
+0.20 dB
+0.00 dB
Upper level set 0.75
Upper level set 0.6
Upper level set 0.4
Upper level set 0.2
Persistence barcode dim 0
Betti curve dim 0
Persistence barcode dim 1
Betti curve dim 1
Intensity (dB)
Betti number


"No Heroes" by Converge
16384
8192
4096
2048
1024
512
0
Hz
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time (sec)
+1.00 dB
+0.80 dB
+0.60 dB
+0.40 dB
+0.20 dB
+0.00 dB
Upper level set 0.5
Upper level set 0.4
Upper level set 0.3
Upper level set 0.2
Persistence barcode dim 0
Betti curve dim 0
Persistence barcode dim 1
Betti curve dim 1
Intensity (dB)
Betti number

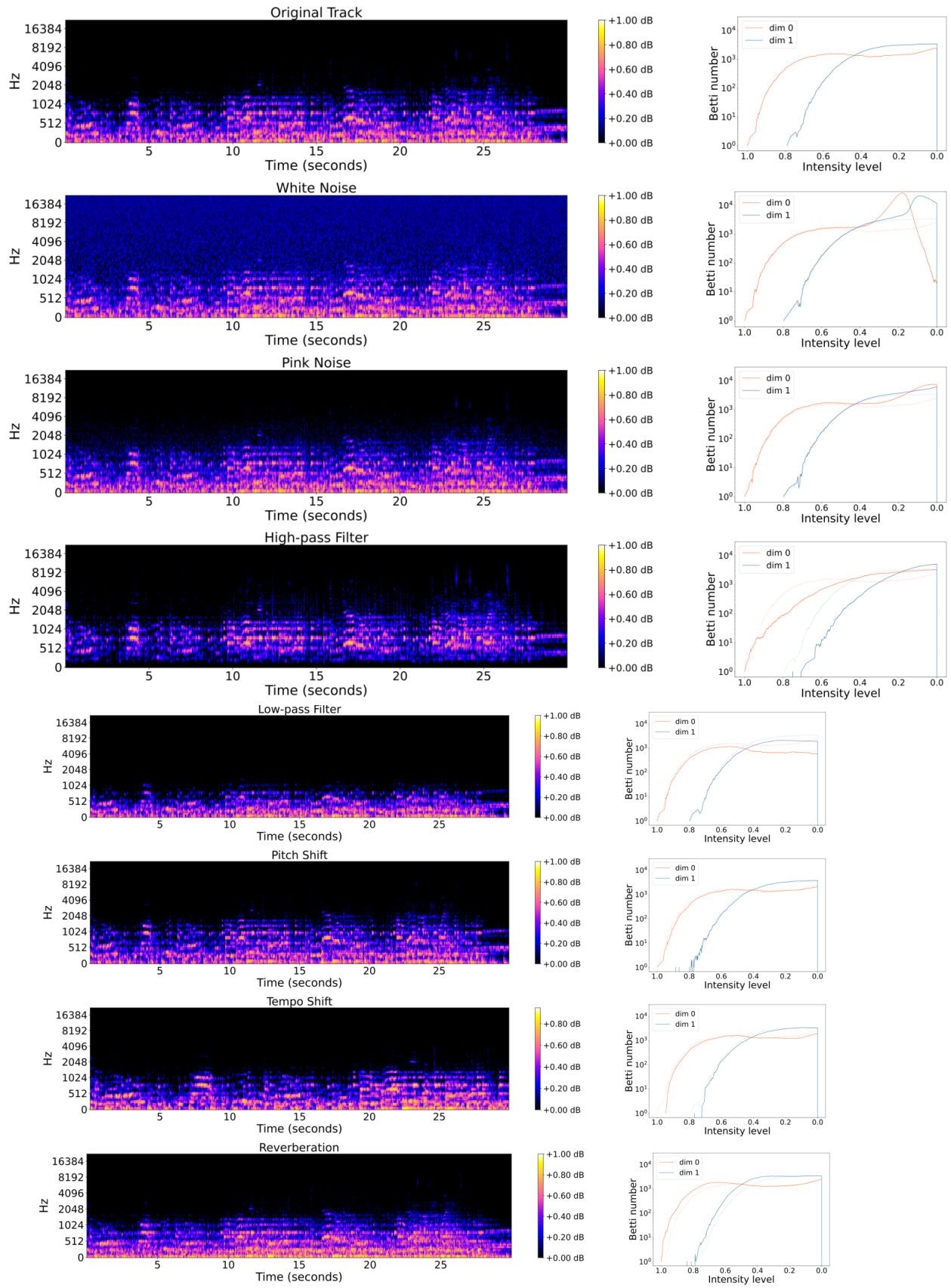
There is a range of audio obfuscation techniques that allow to modify an audio signal without affecting its recognition by human ear.

(1) *Noise:* Addition of a random noise to the main signal. The most common types of noise are: *white noise*, with constant power spectral density (implying that all frequency ranges are affected equally), and *pink noise*, whose power spectral density logarithmically decreases as frequency increases.
(2) *Reverb:* Addition of reverberation to the main signal. Reverberation is the persistence of the sound after the original sound has stopped and it is generated using the algorithm *freeverb* [37, 38].
(3) *High-pass filter:* A filter that allows only the higher-frequency content of the signal to pass through. This is accomplished by specifying a cutoff frequency, below which signals are attenuated, and above which signals are allowed to pass through with little or no attenuation. It is reflected on the darkening or reduction of intensity in the lower frequency region of the spectrogram.
(4) *Low-pass filter:* Analog to high-pass filter, but removing high-frequency components from the signal and letting only the low-frequency content pass through. It darkens the upper frequency region of the spectrogram.
(5) *Tempo shift:* Time stretch the audio signal without changing its pitch. This effect uses WSOLA algorithm [43]. It is reflected as a continuous linear deformation in time of the spectrogram.
(6) *Pitch shift:* Alter the frequencies of signal components while preserving their harmonic relationships. This transformation is evidenced as a continuous vertical deformation in the spectrogram.

(1) ~ (4): rigid obfuscations

(5) and (6): topological obfuscations

Note: Unlike rigid obfuscations, topological obfuscations involve the distortion of time or frequency variables, resulting in a continous deformation of the spectrogram either in the horizontal or vetical direction.


Original Track
White Noise
Pink Noise
High-pass Filter
Low-pass Filter
Pitch Shift
Tempo Shift
Reverberation
Hz
16384
8192
4096
2048
1024
512
0
5
10
15
20
25
Time (seconds)
+1.00 dB
+0.80 dB
+0.60 dB
+0.40 dB
+0.20 dB
+0.00 dB
dim 0
dim 1
Betti number
Intensity level
1.0
0.8
0.6
0.4
0.2
0.0