DEADLINES FOR ASSIGNMENTS

DATE	ASSIGNMENTS	
MONDAY, NOVEMBER 11	1-2	
MONDAY, DECEMBER 2	3-4	
WEDNESDAY, DECEMBER 11	5-6	UNDERGRADS
	5-8	GRADS

HOW DO I GET A GOOD GRADE?

You are encouraged to work in group of 2 to 3 students. Please clearly identify the names of your partners in the project report. You need only submit one individual report per group of students.

Your report should be uploaded as a PDF file in the following format: lastname1_lastname2.pdf, where lastname1 and lastname2 are the last names of the students working on the project.

Your submission will consist of a report; for each question you will provide the following:

1. a few sentences that review the goal of the assignment;
2. a technical description of the approaches that you took to solve the question;
3. experimental results: where applicable: plots, figures, images, etc.
4. you need to include captions, axes legends with units, colorbar, etc.
5. a short *discussion* paragraph wherein you critically evaluate your experimental results (plots, images, etc.), and explain how they relate to the theory learnt in class.
6. For the experiments, you can use the programming language of your choice: C, Python, MATLAB, etc. You will not get credit for your code if you do not add comments in the code.
7. You will only get credit for an experiment or a result if it includes a discussion that connects the experiment with a theoretical analysis.
8. There is no need to include the MATLAB code in the report.

Question 1: /40

Comments:

Question 2: /60 = 12 x 5

Comments:

Question 3: /40

Comments:

Question 4: /60 = 12 x 5

Comments:

Question 5: /40

Comments:

Question 6: /60 = 12 x 5

Comments:

Question 7: /40

Comments:

Question 8: /60 = 12 x 5

Comments:

PART 0: Introduction

The goal of this project is to explore some of the recent techniques developed in the audio industry to organize, and search large music collection by content. The explosion of digital music has created a need to invent new tools to search and organize music. Several websites provide large database of musical tracks:

- http://magnatune.com/
- http://www.allmusic.com/
- http://www.last.fm/

and also allow users to find musical tracks and artists that are similar. Companies such as http://www.gracenote.com/ and http://www.shazam.com/ have application to recognize a song based solely on the actual music. Other examples of automated music analysis include

1. score following: Rock Prodigy, SmartMusic, Rockband
2. automatic music transcription: Zenph
3. music recommendation, playlisting: Google Music, Last.fm, Pandora, Spotify
4. machine listening: Echonest

At the moment these tools are still rudimentary. Fast computational methods are needed to expand these tools beyond their current primitive scope and integrate them on portable music players (e.g. iPod).

The goal of this project is to implement modern signal processing methods for content-based music information retrieval. In this project you will compute features to compare audio tracks, and to automatically detect the genre of a track. The features will be based on the human perception of music (psychoacoustic). These features can be computed using filterbanks, and lead to the definition of chroma.

In the first two parts of the project you will work on a small set of 12 music tracks. In the last part , you will test your algorithms on a a database of 150 training samples to learn the association between genres and songs. The performance of the algorithm will be evaluated using songs with know labels that are not part of the training set.

1.1 Terminology

For the purpose of this document we will be describing music using a terminology that is biased toward popular music. As a results we use the following words in a somewhat different context:

- an artist refers to the performer of a piece of music in the case of popular music, and a composer in the case of classical music. For instance, both Miles Davis and Johann Sebastian Bach are two artists.
- a song refers to a recording of (a part of) a piece of music. A song is identified with a CD track. The name does not imply any vocal performance.

1.2 The music

In the file tracks.zip you will find 12 tracks of various length. You will use these files to test the algorithms you develop in PARTS I & II. Please read the license posted on the webpage before downloading.

The 12 tracks are two examples of six different musical genres:

1. Classical
2. Electronic
3. Jazz
4. Metal and punk
5. Rock and pop
6. World

The name of the file indicates its genre. The musical tracks are chosen because they are diverse but also have interesting characteristics, which should be revealed by your analysis.

- track201-classical is a part of a violin concerto by J.S. Bach (13-BWV 1001 : IV. Presto).
- track204-classical is a classical piano piece composed by Robert Schumann (Davidsbundlertanze, Op.6. XVI).
- track370-electronic is an example of synthetic music generated a software that makes it possible to create notes that do not have the regular structure of Western music. The pitches may be unfamiliar but the rhythm is quite predictable.
- track396-electronic is an example of electronic dance floor piece. The monotony and the simplicity of the synthesizer and the bass drum are broken by vocals.
- track437-jazz is another track by the same trio as track439-jazz
- track439-jazz is an example of (funk) jazz with a Hammond B3 organ, a bass, and a percussion. The song is characterized by a well defined rhythm and a simple melody.
- track463-metal is an example of rock and metal with vocals, bass, electric guitars and percussion.
- track492-metal is an example of heavy metal with heavily distorted guitars, drums with double bass drum.
- track547-rock is an example of new wave rock of the 80's. The music varies from edgy to hard. It has guitars, keyboards, and percussion.
- track550-rock is an example of pop with keyboard, guitar, bass, and vocals. There is a rich melody and the sound of steel-string guitar.
- track707-world: this is a Japanese flute, monophonic, with background sounds between the notes. The notes are held for a long time.
- track729-world: this is a piece with a mix of Eastern and Western influence using electric and acoustic sarod and on classical steel-string guitars.

PART I: From the wav file to the psychoacoustic features

Sounds are produced by the vibration of air; sound waves produce variations of air pressure. The sound waves can be measured using a microphone that converts the mechanical energy into electrical energy. Precisely, the microphone converts air pressure into voltage levels. The electrical signal is then sampled in time at a sampling frequency f_s, and quantized. The current standard for high quality audio and DVD is a sampling frequency of f_s = 96kHz, and a 24-bit depth for the quantization. The CD quality is 16 bit sampled at f_s = 44.1 kHz.

2.1 Frames, and Samples

In order to automatically analyze the music, the audio file is divided into overlapping intervals of a few milliseconds over which the analysis is conducted. In this project the audio files are sampled at 22,050 Hz, and we only consider a monophonic reproduction. We use an interval of $N = 2048$ samples, of length 93 millisecond is called a **frame**. Throughout the project we will always split a track into overlapping frames. The number of frames is denoted by N_F.

Music is made up of notes of different pitch. Our goal is the reconstruction of a musical score. It is only natural that most of the automated analysis of music should be performed in the spectral (frequency) domain. Our analysis requires $N = 2048$ samples to compute notes over a frame. The spectral analysis proceeds as follows.

Each frame is smoothly extracted by multiplying the original audio signal by a tapper window w. The Fourier transform of the windowed signal is then computed. If x_n denotes a frame of size $N = 2048$ extracted at frame n, and w is a window of size N, then the Fourier transform, X_n (of size N) for the frame n is given by

```
Y  = fft (w.*xn);
K  = N/2 + 1;
Xn = Y(1:K);
```

2.2 Basic Psychoacoustic Quantities

In order to develop sophisticated algorithms to analyze music based on its content, we need to define several subjective features such as timbre, melody, harmony, rhythm, tempo, mood, lyrics, etc. Some of these concepts can be defined formally, while others are more subjective and can be formalized using a wide variety of different algorithms. We will focus on the features that can be defined mathematically.

2.3 Psychoacoustic

Psychoacoustics involves the study of the human auditory system, and the formal quantification of the relationships between the physics of sounds, and our perception of audio. We will describe some key aspects of the human auditory system:

1. the perception of frequencies and pitch for pure and complex tones;
2. the frequency selectivity of the auditory system: our ability to perceive two similar frequencies as distinct;

3. the modeling of the auditory system as a bank of auditory filters;
4. the perception of loudness;
5. the perception of rhythm.

2.4 Perception of frequencies

The auditory system, like the visual system, is able to detect frequencies over a wide range of scales. In order to measure frequencies over a very large range, it operates using a logarithmic scale. Let us a consider a pure tone, modeled by a sinusoidal signal oscillating at a frequency ω. If $\omega < 500$ Hz, then the perceived tone – or pitch – varies as a linear function of ω. When $\omega > 1,000$ Hz, then the perceived pitch increases logarithmically with ω.

Several frequency scales have been proposed to capture the logarithmic scaling of frequency perception.

2.5 The mel/Bark scale

The Bark (named after the German physicist Barkhausen) is defined as

$$z = 7\text{arcsinh}\left(\omega/650\right) = 7\log\left(x/650 + \sqrt{1 + (x/650)^2}\right),$$

where ω is measured in Hz. The mel-scale is defined by the fact that 1 bark = 100 mel. In this project we will use a slightly modified version of the mel scale defined by

$$m = 1127.01048 * \log\left(1 + \omega/700\right). \tag{1}$$

2.6 The cochlear filterbank

Finally, we need to account for the fact that the auditory system behaves as a set of filterbanks, with overlapping frequency responses. For each filter, the range of frequencies over which the filter response is significant is called the critical band. Our perception of pitch can be quantified using the total energy at the output of each filter bank. All spectral energy that falls into one critical band is summed up, leading to a single number for that frequency band.

We describe in the following a simple model of the cochlear filterbank. The filter bank is constructed using $N_B = 40$ logarithmically spaced triangle filters centered at the frequencies Ω_p, defined by

$$\text{mel}_n = 1127.01048 * \log\left(1 + \Omega_p/700\right), \tag{2}$$

where the sequence of mel frequencies is equally spaced in the mel scale,

$$\text{mel}_n = n\frac{\text{mel}_{\text{max}} - \text{mel}_{\text{min}}}{N_B}, \tag{3}$$

with

$$\text{mel}_{\text{max}} = 1127.01048 * \log\left(1 + 0.5 * \omega_s/700\right), \tag{4}$$
$$\text{mel}_{\text{min}} = 1127.01048 * \log\left(1 + 20/700\right), \tag{5}$$

and $N_B = 40$. Each filter H_p is centered around the frequency Ω_p, and defined by

$$H_p(\omega) = \begin{cases} \dfrac{2}{\Omega_{p+1} - \Omega_{p-1}} \dfrac{\omega - \Omega_{p-1}}{\Omega_p - \Omega_{p-1}} & \text{if} \quad \omega \in [\Omega_{p-1}, \Omega_p), \\ \dfrac{2}{\Omega_{p+1} - \Omega_{p-1}} \dfrac{\Omega_{p+1} - \omega}{\Omega_{p+1} - \Omega_p} & \text{if} \quad \omega \in [\Omega_p, \Omega_{p+1}). \end{cases} \tag{6}$$

Each triangular filter is normalized such that the integral of each filter is 1. In addition, the filters overlap so that the frequency at which the filter H_p is maximum is starting frequency for the next filter h_{n+1}, and the edge frequency of h_{n-1}.

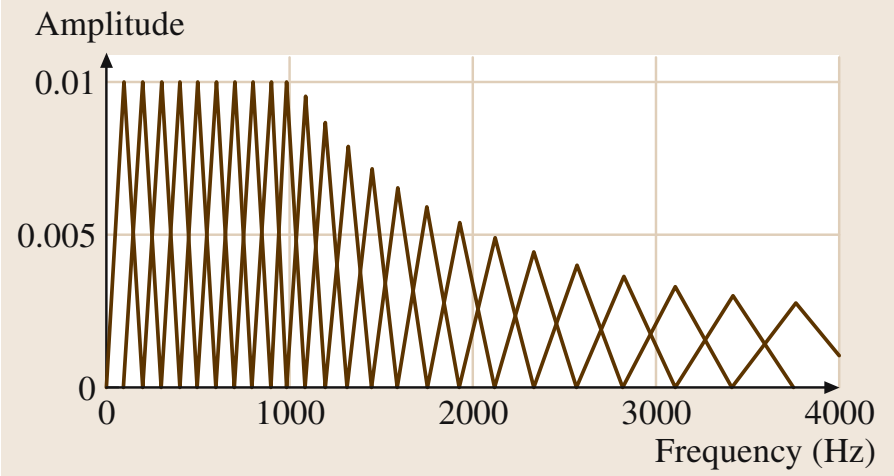

Figure 1: The filterbanks used to compute the mfcc.

Finally, the mel-spectrum (MFCC) coefficient of the n-th frame is defined for $p = 1, \ldots, N_B$ as

$$\texttt{mfcc}[p] = \sum_{k=1}^{K} |H_p(k) X_n(k)|^2 \tag{7}$$

where the filter H_p is a discrete implementation of the continuous filter defined by (6). The discrete filter is normalized such that

$$\sum_j H_p(j) = 1, p = 1, \ldots, N_B. \tag{8}$$

The MATLAB code in the following next three pages computes the `mfcc` coefficients. The computation of filter banks H_p is already completed. You need to implement (7) in the Fourier domain.

```
function [mfcc] = mfcc(wav, fs, fftSize, window)
%
% USAGE
%   [mfcc] = mfcc(wav, fs, fftSize,window)
%
% INPUT
```

```
%      vector of wav samples
%      fs : sampling frequency
%      fftSize: size of fft
%      window: a window of size fftSize
%
% OUTPUT
%   mfcc (matrix) size coefficients x nFrames


%   hardwired parameters

hopSize = fftSize/2;
nBanks = 40;

%    minimum and maximum frequencies for the analysis
fMin = 20;
fMax = fs/2;

%_____________________________________________________________________
%
%     PART 1 : construction of the filters in the frequency domain
%_____________________________________________________________________

%  generate the linear frequency scale of equally spaced frequencies from 0 to fs/2.

linearFreq = linspace(0,fs/2,hopSize+1);

fRange = fMin:fMax;

%  map the linear frequency scale of equally spaced frequencies from 0 to fs/2
% to an unequally spaced mel scale.

melRange = log(1+fRange/700)*1127.01048;

% The goal of the next coming lines is to resample the mel scale to create uniformly
% spaced mel frequency bins, and then map this equally spaced mel scale to the linear
& scale.

% divide the mel frequency range in equal bins

melEqui = linspace (1,max(melRange),nBanks+2);

fIndex = zeros(nBanks+2,1);

% for each mel frequency on the equally spaces grid, find the closest frequency on the
```

```
% unequally spaced mel scale

for i=1:nBanks+2,
    [dummy fIndex(i)] = min(abs(melRange - melEqui(i)));
end

% Now, we have the indices of the equally-spaced mel scale that match the unequally-spaced
% mel grid. These indices match the linear frequency, so we can assign a linear frequency
% for each equally-spaced mel frequency

fEquiMel = fRange(fIndex);

%  Finally, we design of the hat filters. We build two arrays that correspond to the center,
%  left and right ends of each triangle.

fLeft   = fEquiMel(1:nBanks);
fCentre = fEquiMel(2:nBanks+1);
fRight  = fEquiMel(3:nBanks+2);

% clip filters that leak beyond the Nyquist frequency

[dummy, tmp.idx] = max(find(fCentre <= fs/2));
nBanks = min(tmp.idx,nBanks);

% this array contains the frequency response of the nBanks hat filters.

freqResponse = zeros(nBanks,fftSize/2+1);

hatHeight = 2./(fRight-fLeft);


% for each filter, we build the left and right edge of the hat.

for i=1:nBanks,
    freqResponse(i,:) = ...
        (linearFreq > fLeft(i) & linearFreq <= fCentre(i)).* ...
        hatHeight(i).*(linearFreq-fLeft(i))/(fCentre(i)-fLeft(i)) + ...
        (linearFreq > fCentre(i) & linearFreq < fRight(i)).* ...
        hatHeight(i).*(fRight(i)-linearFreq)/(fRight(i)-fCentre(i));
end

%
% plot a pretty figure of the frequency response of the filters.

figure;set(gca,'fontsize',14);semilogx(linearFreq,freqResponse');
axis([0 fRight(nBanks) 0 max(freqResponse(:))]);title('FilterbankS');
```

```
%_____________________________________________________
%
% PART 2 : processing of the audio vector In the Fourier domain.
%_____________________________________________________
%
%  YOU NEED TO ADD YOUR CODE HERE
```

Assignment

1. Implement the computation of the mfcc coefficients, as defined in (7). You simply need to add your code in the MATLAB code in the previous pages.
2. Evaluate your MATLAB function `mfcc` on the 12 audio tracks, and display the output as an image using `imagesc`. You will use T =**24 seconds** from the middle of each track and compute a matrix of `mfcc` coefficients of size $N_B = 40$ rows and $24 \times 22,050/2048 = 258$ columns.

PART II: Tonality and Chroma

3.1 The equal-tempered scale

We now investigate the time-frequency structure related to the concept of tonality and chroma. The tonality is concerned with the distribution of notes at a given time. In contrast, the melody characterizes the variations of the spectrum as a function of time.

We first observe that the auditory sensation that is closely related to frequency, also known as pitch, corresponds to a logarithmic scale of the physical frequency. We have seen examples of such scales in PART I: the bark and the mel scales.

In this PART, we define another logarithmic scale, known as the *tempered scale*. First, we define a reference frequency f_0 associated with a reference pitch. While it is customary to use the frequency 440 Hz associated with the pitch A4, we will also use $f_0 = 27.5Hz$ (this known as A0, the lowest note on a piano) . The tempered scale introduces 12 frequency intervals – known as semi-tones – between this reference frequency f_0 and the next frequency also perceived as an A, $2f_0$. As a result, a frequency f_{sm} that is sm semitones away from f_0 is given by

$$f_{sm} = f_0 2^{sm/12}. \tag{9}$$

An interval of 12 semitones, $[f_{sm}, f_{sm+12})$, corresponds to an octave (see Fig. 2). The same notes (e.g. A, or C#) are always separated by an octave. For instance, the two previous As, A0 and A4 are separated by 4 octaves, since $440 = 2^4 \times 27.5$. The notion of note can be formally modeled by the concept of chroma.

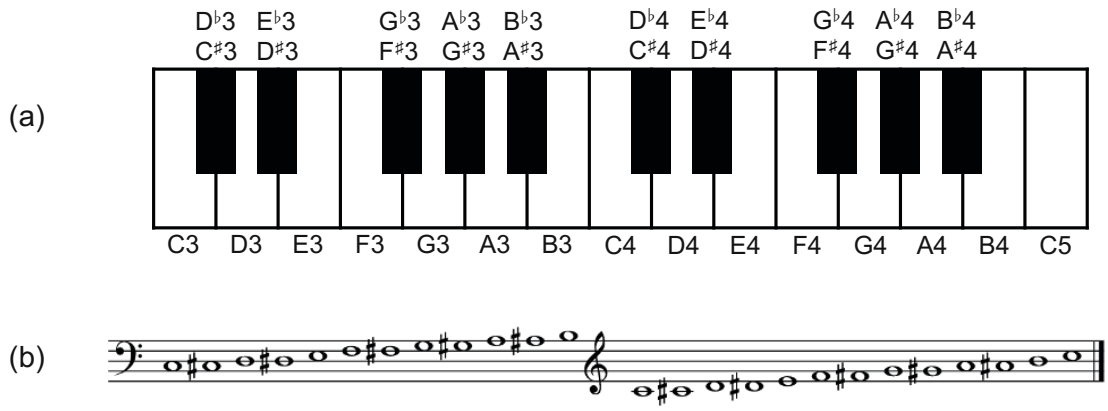

Figure 2: The well tempered-scale. (a) Section of piano keyboard with keys ranging from C3 to C5. (b) Corresponding notes using Western music notation.

3.2 Chroma

The chroma is associated with the relative position of a note inside an octave. It is a relative measurement that is independent of the absolute pitch. We describe an algorithm to compute a chroma feature vector for a frame n. To lighten the notation, we drop the dependency on the frame index n in this discussion.

We first present the idea informally, and then we make our statement precise. We are interested to map a given frequency f to a note. Given a reference frequency f_0, then we can use (9) to compute the distance sm between f and f_0 measured in semitones,

$$sm = \text{round}(12 \log_2(f/f_0)). \tag{10}$$

We note that f is usually not exactly equal to $f_0 2^{sm/12}$, and this is why we round $12 \log_2(f/f_0)$ to the closest integer.

We can then map this index sm to a note, or chroma, c within an octave, using

$$c = sm \pmod{12}, \tag{11}$$

or

$$sm = 12q + c, \quad \text{with} \quad 0 \leq c \leq 11, \quad \text{and} \quad q \in \mathbb{N}. \tag{12}$$

In other words, f_{sm} and f_{sm+12} given by (9) correspond to the same note c, which is exactly what you see if you look at the keys of a piano.

The problem with this approach is that we do not have an algorithm to extract the frequencies f of the notes that are being played at a given time. Rather, we obtain a distribution of the spectral energy provided by the Fourier transform. We need to identify the main peaks in the Fourier transform, and compute the chroma associated with these frequencies, using the equations above.

In the following we describe the different steps of the algorithm in details.

Step 1: spectral analysis and peak detection

The goal is to identify the main notes being played , and remove the noise and the harmonics coming from the timbre of each instrument.

For each frame n, we compute the windowed FFT as explained in PART I. We obtain $K = N/2 + 1$ (where N is even) Fourier coefficients. We detect the local maxima of the magnitude of the Fourier transform $|X_n|$, and count the number of maxima, or peaks, n_{peaks}. We denote by $f_1 < f_2 < \cdots < f_{n_{\text{peaks}}}$ the frequencies at which the peaks occur (see Fig. 3).

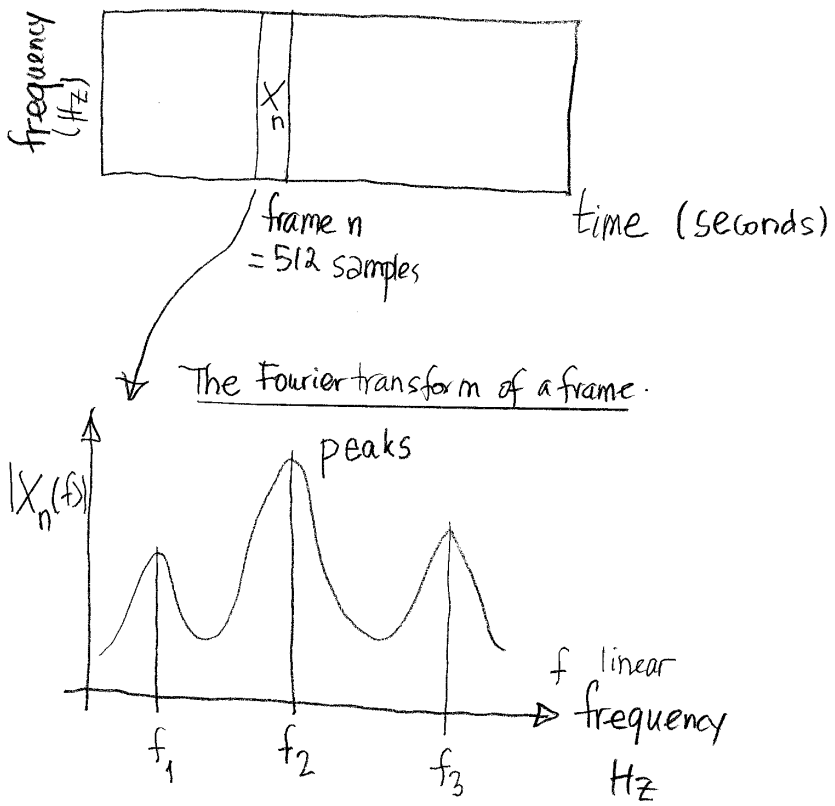

Figure 3: The frequencies $f_1, f_2, f_3, \ldots$ are associated with peaks in the spectral energy $|X_n|^2$ of frame n.

Step 2: Assignment of the peak frequencies to semitones

In the second step, we bin the peak frequencies into semitone bins. The semitone bin of index sm is centered at the frequency $f_0 2^{sm/12}$, and extends by $f_0 2^{sm/12-1}(1 + 2^{-1/12})$ to the left, and by $f_0 2^{sm/12-1}(1 + 2^{1/12})$ to the right (the bin is asymmetric because of the geometric progression of frequencies).

For each peak frequency, $f_k, k = 1, \ldots, n_{\text{peaks}}$, we find the semitone "bin" of index sm and the corresponding frequency $f_{sm} = f_0 2^{sm/12}$ closest to f_k, so that

$$f_0 2^{sm/12} \leq f_k < f_0 2^{(sm+1)/12}, \tag{13}$$

or equivalently,

$$sm \leq 12 \log_2(f_k/f_0) < sm + 1. \tag{14}$$

The equation for sm is thus

$$sm = \text{round}(12 \log_2(f_k/f_0)). \tag{15}$$

Finally, we map the index sm to the note c defined by

$$c = sm \pmod{12}. \tag{16}$$

For computational efficiency we use $f_0 = 27.5$ Hz instead of 440 Hz.

Step 3: the Pitch Class Profile: a weighted sum of the semitones

Instead of assigning each peak frequency f_k to a single semitone sm, we spread the effect of f_k to the two neighboring semitones sm and $sm + 1$ using a raised cosine function, defined in a weight function $w(k, c)$.

We define the *Pitch Class Profile* associated with the note, or chroma $c = 0, \ldots, 11$ as the weighted sum of all the peak frequencies that are mapped to the note c, irrespective of the octave they fall in. To wit, all the harmonics at the frequencies

$$2^{\text{octave}} f_0 2^{c/12}, \quad \text{where} \quad \text{octave} = 1, 2, \ldots \tag{17}$$

are perceived as being similar, and thus should be consolidated into a single number. To collect the energy associated with the note c and all its harmonics, we combine the contributions of all the harmonics in the following way.

We first lay down a sequence of weighting functions located at all the harmonics of the note c, $f_0 2^{c/12}, 2 f_0 2^{c/12}, 4 f_0 2^{c/12}, \ldots$ (see Fig. 4). The weighting functions are used to define the bins associated with the harmonics of the note c.

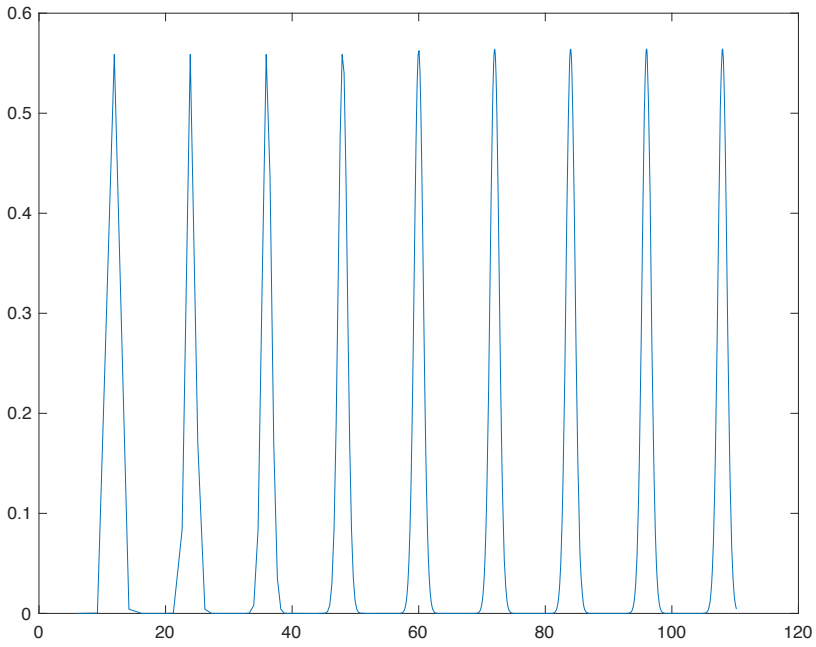

Figure 4: The note A and all its harmonics and associated weighting functions. The x-axis is in semitone units (A1 = 12 semitones, A2 = 24, A3 = 36, and A4 = 48, etc).

Instead of using a bin that counts a frequency f_k either inside or outside the bin, we use a "mollified assignment" that depends on the distance of each peak frequency f_k to the centre of a bin.

If the frequency f_k falls in the bin corresponding to the frequency $2^{\text{oct}} f_0 2^{c/12}$, associated with the octave oct, then the weight $w(k, c)$ measures the distance from f_k to the centre of the bin, $2^{\text{oct}} f_0 2^{c/12}$. Alternatively, if the frequency f_k does not fall in the bin $f_0 2^{c/12} 2^{\text{oct}}$, we set $w(k, c) = 0$.

Formally, we have

$$w(k, c) = \exp\left(-[\log_2(f_k/f_0) - 12 * \text{oct}]^2\right). \tag{18}$$

We note that we use a logarithmic scale to quantify the distance from f_k to the center frequency, $2^{\text{oct}} f_0 2^{c/12}$. Obviously, if $f_k = 2^{\text{oct}} f_0 2^{c/12}$, then $w(k, c) = 1$. Conversely, if $f_k = 2^{\text{oct}} f_0 2^{(c\pm1)/12}$, that is f_k should be in the previous or next bin, then $w(k, c) = 0$.

Figure 5 displays the 12 bumps associated with $w(c, k), C = 1, \ldots, 12$ over one octave for the semitones in the range [32,44]. The bump at $sm = 32$ corresponds to the note $F2$, since we have $32 = 2 \times 12 + 8$. The next harmonic, $F4$ corresponds to the last bump located at $sm = 44 = 4 \times 12 + 8$.

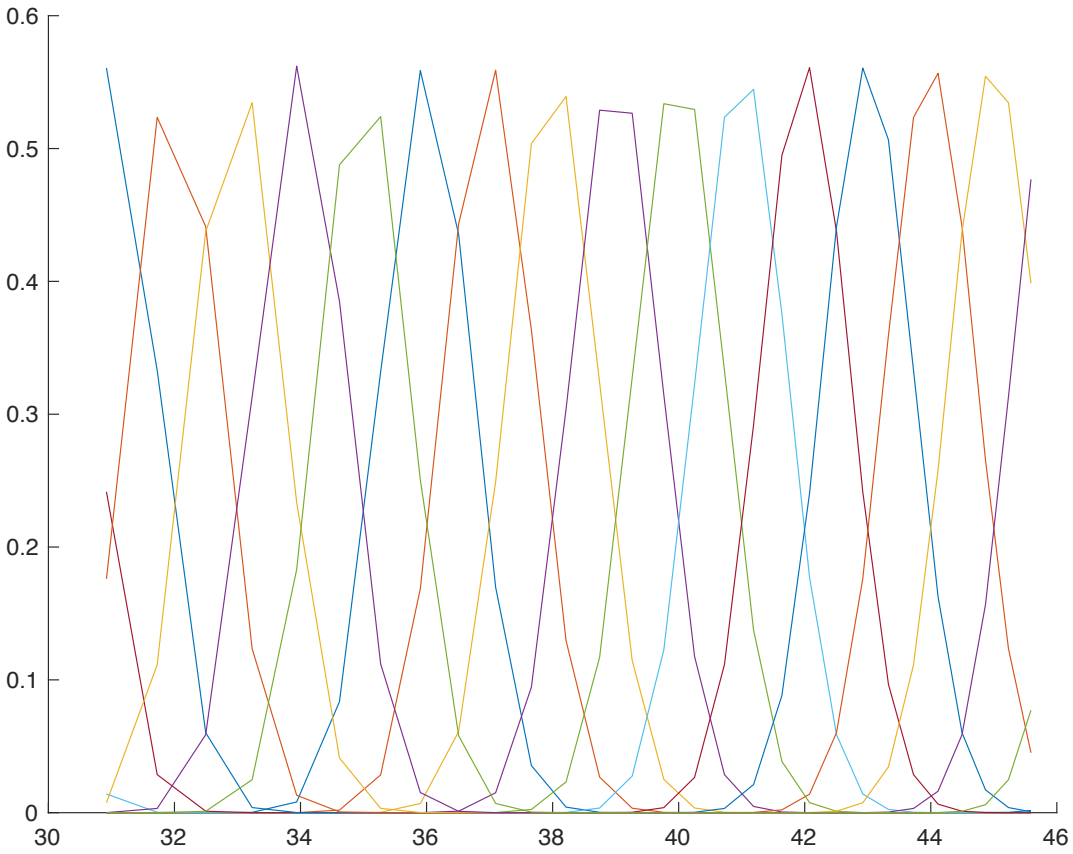

Figure 5: The 12 bumps $w(c, k), c = 1, \ldots, 12$ over the octave defined by semitone in the range [32,44]. Each bump function $w(c, k)$ corresponds to a note c. The bump located at $sm = 32$ corresponds to the note F (F2), since $32 = 2 \times 12 + 8$. The bump located at $sm = 36$ corresponds to an A (A3), since $36 = 3 \times 12 + 0$, etc.

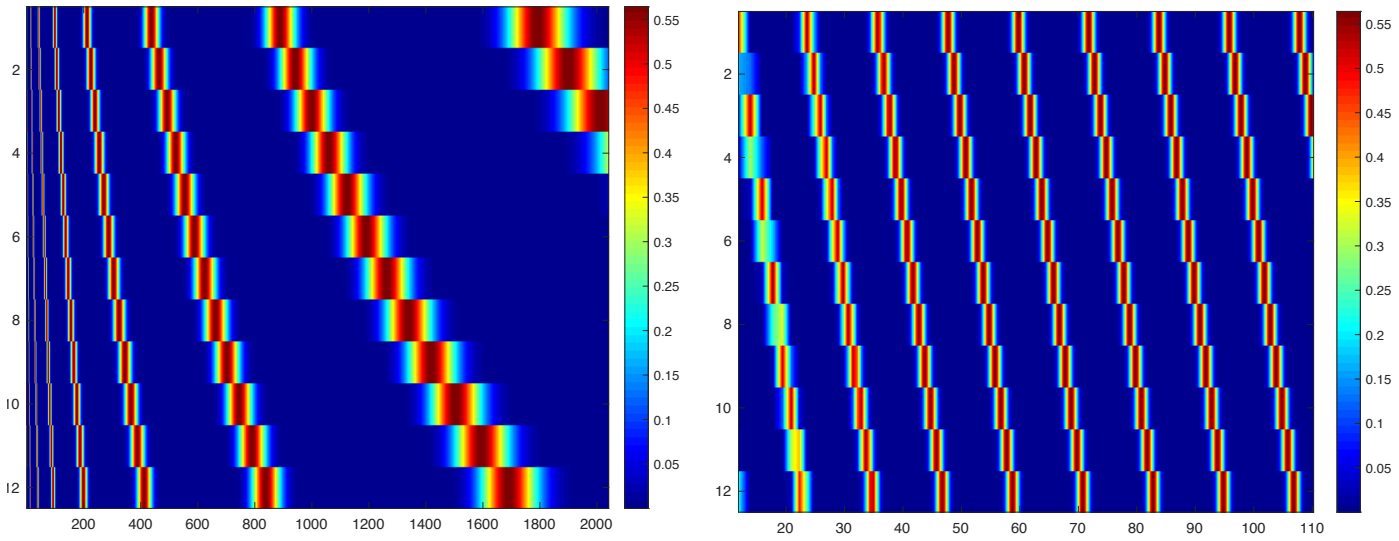

Figure 6: Left: the function $w(c,k)$ as a function of the index $k \in [0, 2048]$ of the Fourier transform. Right: the function $W(c,k)$ as a function of the semitone index sm that matches the frequency f_k as defined by equation (15).

Figure 6 displays the function $w(c,k)$ as a matrix: for each note (row) c, the entry $w(c,k)$ will be large in the columns corresponding to harmonics of the fundamental frequency $f_0 2^{c/12}$. Conversely, for each frequency (column), there exists a single note c – defined by (16), where $w(c,k)$ is large. Because the harmonics form a geometric progression, the lines where the notes lie are curved when the frequency is linear (left of Fig. 6), but are straight line when the frequency is expressed in semitones (right of Fig. 6). Figure 4 displays row 1 of the matrix $w(c,k)$ as a function of the semitone sm; the bumps associated with the function $w(c,k)$ are equidistant using the semitone scale.
Putting everything together, the weight $w(k,c)$ is given by

$$w(k,c) = \begin{cases} \exp(-r^2) & \text{if } -1 < r = 12\log_2(f_k/f_0) - sm < 1, \text{with } sm = \text{round}(12\log_2(f_k/f_0)), \\ & \text{and } c = sm \pmod{12}; \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

To compute the pitch class profile for note c, we compute the correlation (dot product) between the harmonically related weighting functions $w(k,c)$ (see Fig. 4), and the vector of peak frequencies (see Fig. 5), and we define the Pitch Class Profile (PCP)

$$PCP(c) = \sum_{\text{over all harmonics of note } c} (\text{distance from } f_k \text{ to the nearest harmonic of c}) \times |X_n(f_k)|^2. \tag{20}$$

Now, since we only care about the peak frequencies that are near the harmonics, we may as well visit each peak frequency f_k, and see if it falls within a bin associated with a harmonic of c, using $w(k,c)$.

Formally, we compute

$$PCP(c) = \sum_{k=1}^{n_{\text{peaks}}} w(k,c)|X_n(f_k)|^2. \tag{21}$$

Figure 7 shows that the matrix PCP can be computed using a simple matrix product, once the matrix w has been computed.

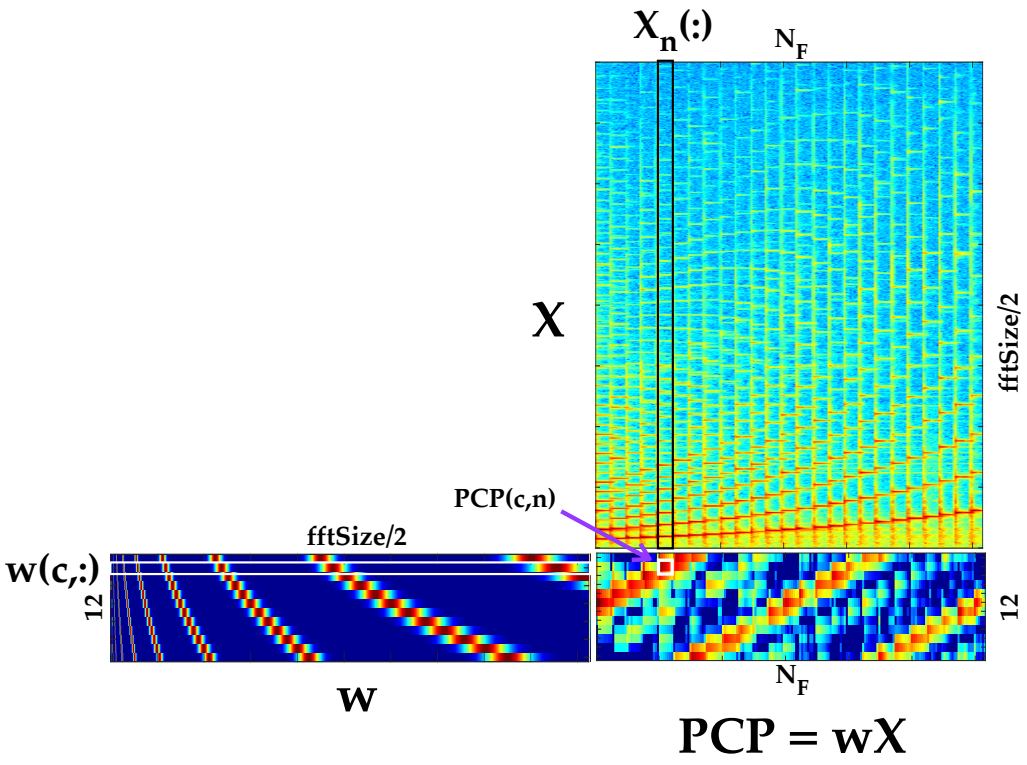

Figure 7: The matrix PCP is the product of the matrix w (see Fig. 6) and the short-time Fourier transform: PCP = w X. Entry PCP (c,n) corresponds to note c and time n, and is computed using equation (21), which is implemented as a dot product between w(c,:) and X_n(:). The matrix PCP has size 12 x N_F, where N_F is the number of frames.

Assignment

3. Implement the computation of the Pitch Class Profile, defined by (21). You will compute a vector of 12 entries for each frame n.

4. Evaluate and plot the PCP for the 12 audio tracks. You should first debug your code using the track chroma.wav, and compare the result to Fig. 8. You can then compare your results on the twelve tracks to Figs. 9,10.

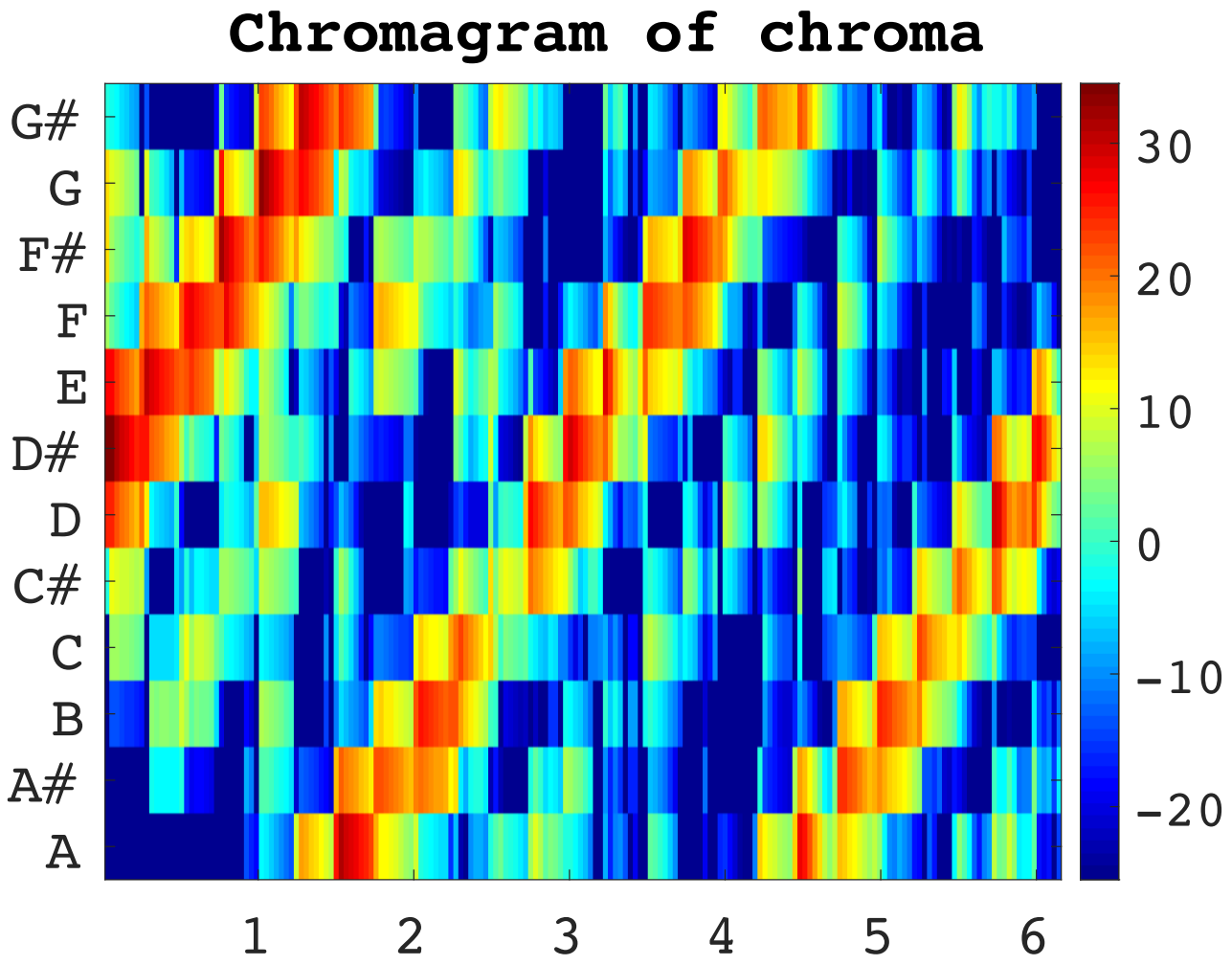

Figure 8: Chroma as a function of time (seconds).

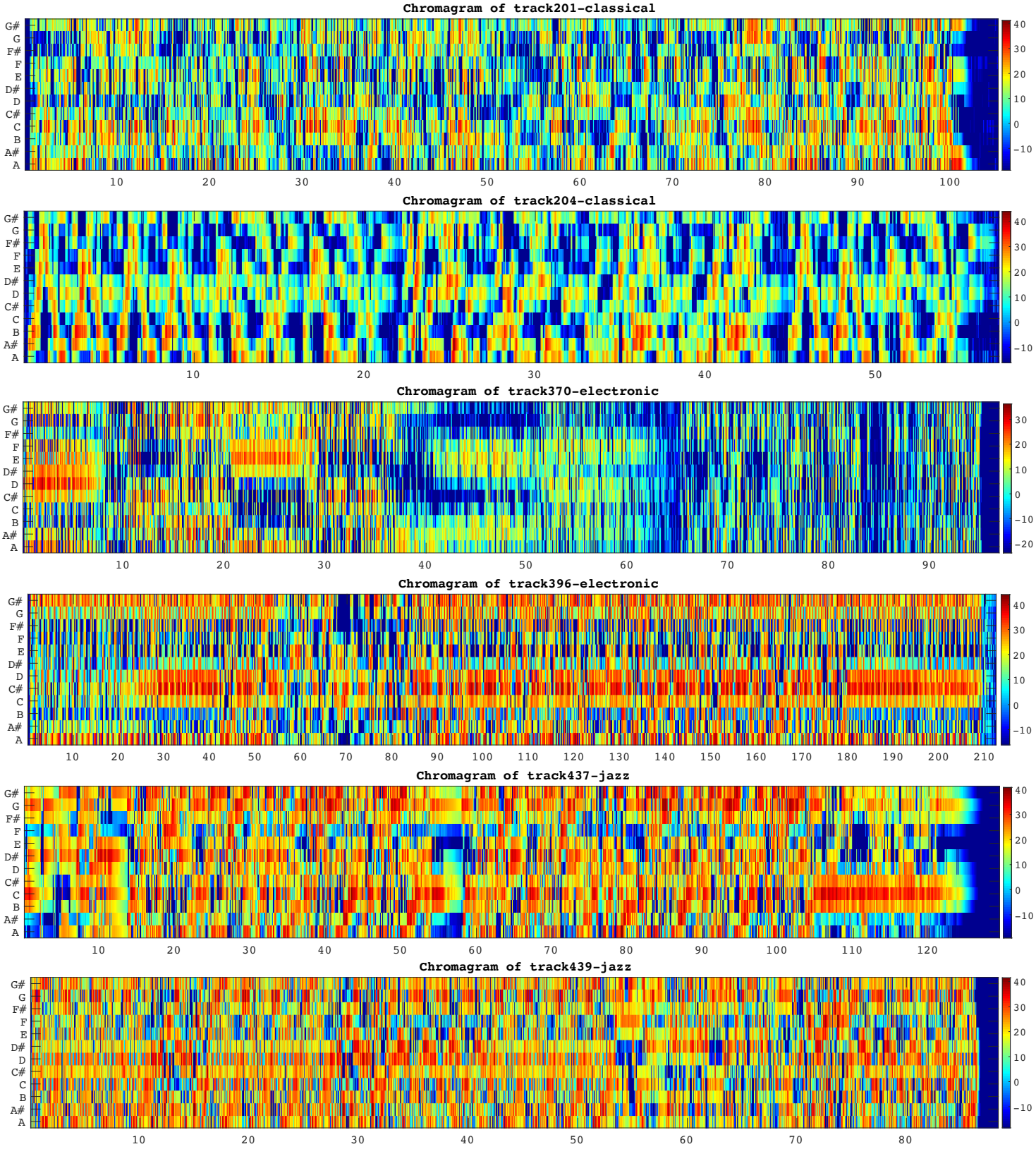

Figure 9: Chroma as a function of time (seconds).

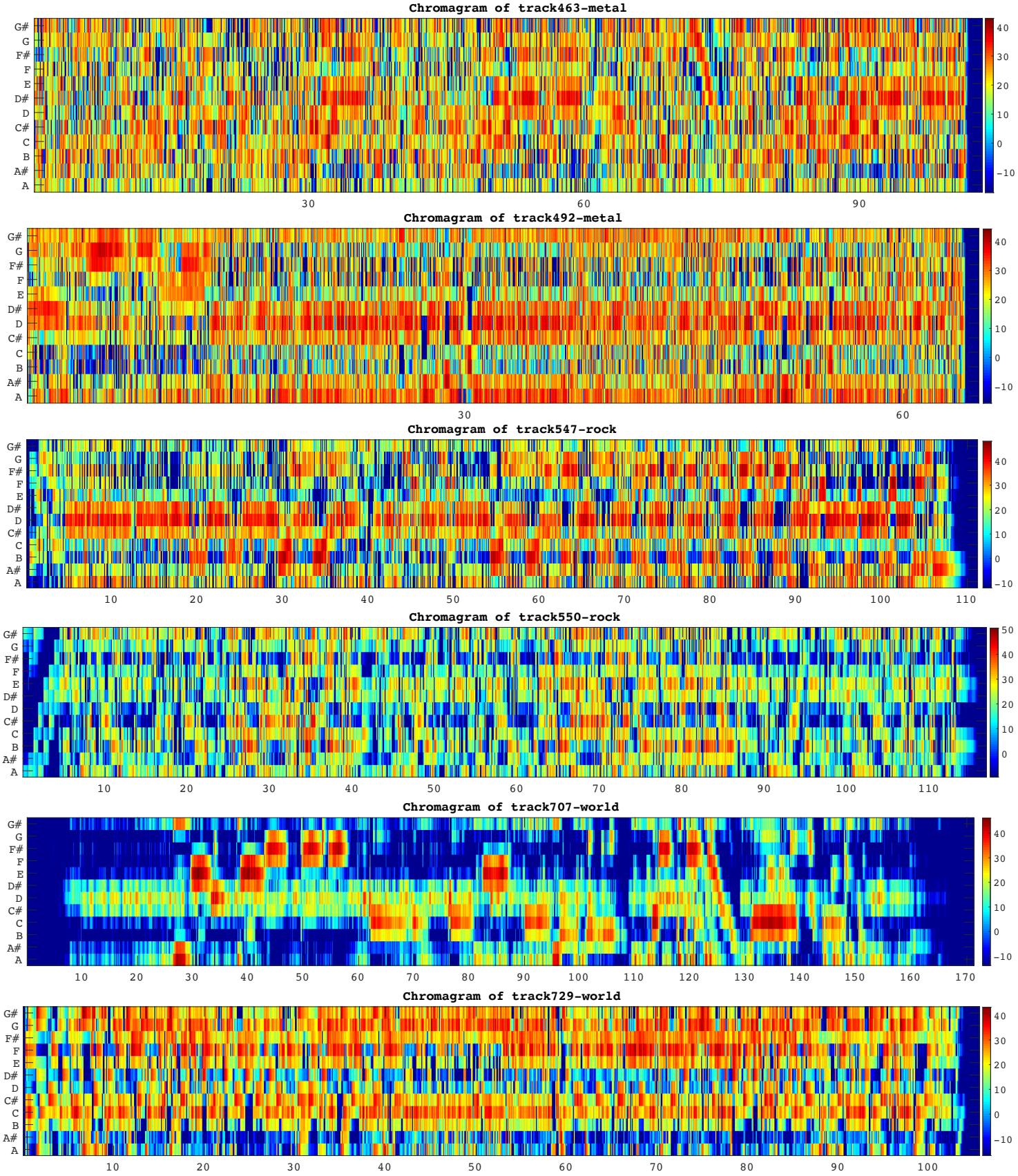

Figure 10: Chroma as a function of time (seconds).

PART III: Genre classification

While the definition of a musical genre is clearly arbitrary, and influenced by a culture, we will define genre classification as the problem of classifying music according to a set of (personal) labels. In particular, we will consider the following labels:

1. classical (western classical music)
2. electronic
3. jazz/blues
4. metal/punk
5. pop/rock
6. world.

The last label "world" is really a category that is only defined as being the complement of the union of the other categories. We note that you are not asked to refine the classification beyond the six categories.

The goal of PART III is to develop an algorithm that can upload a song and return a label with a certain probability. For instance, if the song "Recitative" from the album "Dark Intervals" by Keith Jarret is selected, one would expect that the algorithm classifies it as "classical" or "jazz" with a high probability. The artist Keith Jarrett is a clear example of the difficulty of such a genre classification, and the need for a probabilistic framework. The lab focuses on the genre classification question.

4.1 The data

4.2 Raw audio data

The data available are Pulse Coded Modulation (PCM) encode audio files. These format provides a digital sampling of the acoustic wave created by the sound pressures changes as a function of time. The audio has the following characteristics:

- sampling rate: 22,050 Hz
- sample size: 16 bit
- Number of channels: 1 (mono)
- Encoding: WAV

4.3 Available files

The composition of the training dataset is as follows:

- classical: 25 songs
- electronic: 25 songs

- jazz/blues: 25 songs
- metal/punk: 25 songs
- rock/pop: 25 songs
- world: 25 songs

Each song is a track of a CD and last for several minutes. While songs may have different lengths, we will work with an excerpt of fixed length extracted from the centre of the track.

4.4 But where are the songs?

A 1.4 GB archive with the 150 songs is available in the archive contest.zip.

4.5 Copyright

These songs are made available to you for educational purposes only. It is illegal to use these songs for commercial purposes. These tracks are made available under the non-commercial use as defined by the Creative Commons License: http://creativecommons.org/licenses/by-nc-sa/1.0/legalcode http://magnatune.com/info/licensing You must abide by the Attribution-NonCommercial-ShareAlike use restrictions placed by the license.

4.6 Features

We extract an audio excerpt of 2 minutes from the center of each piece, and compute several audio features, and combine them into a vector $\boldsymbol{X}$. If the track is shorter than 2 minutes, we use the entire track. We will work with frames of size 2048 samples (93 ms for the sampling rate equal to 22,050 Hz), and an overlap between frames of 50% = 1024 samples. We have $N_F = 2,584$ overlapping frames of size 2048 in 2 minutes of music.

We consider the following two vectors of features,

1. MFCC coefficients,
2. chroma (Normalized Pitch Class Profile)

Both sets of features yield a matrix $\boldsymbol{X}(k, n)$ that depends on a frequency (or pitch) index $k = 1, \ldots, K$, and a frame index $n = 1, \ldots, N_F$. We can think informally about $\boldsymbol{X}(1 : K, n)$ as the set of notes being played at time n.

In order to obtain more reliable classification results, and speed up the computation, we merge some of the Mel banks. We retain only 12 bands, as in the chroma representation. The new bands are defined by the next lines of MATLAB code.

```
t = zeros(1,36); (2.21)
t(1) =1;t(7:8)=5;t(15:18)= 9;
t(2)   = 2; t( 9:10) = 6; t(19:23) = 10;
t(3:4) = 3; t(11:12) = 7; t(24:29) = 11;
```

```
t(5:6) = 4; t(13:14) = 8; t(30:36) = 12;

mel2 = zeros(12,size(mfcc,2));

for i=1:12,
    mel2(i,:) = sum(mfcc(t==i,:),1);
end
```

In the remaining, we will consider that the mfcc coefficients are the 12 merged mfcc coefficients; in other words

```
mfcc = mel2;
```

4.7 Distance between features

Given two tracks s_1 and s_2, we are interested in comparing the distribution of notes between the two tracks. We proceed as follows.

For each track, the distribution of vectors $\boldsymbol{X}(:, n), n = 1, \ldots, N_F$ is modeled as a multivariate Gaussian distribution in $\mathbb{R}^K$, where $K = 12$. In essence, this approach collapses all the frames together, and summarizes the 2-minute excerpt by a mean note, and a covariance matrix. In MATLAB, we compute

```
>> mu = mean(mfcc,2);
>> co = cov(mfcc');
```

To compare two tracks, we compare the distance between the two Gaussian distributions $G^1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $G^2 = \mathcal{N}(\mu_2, \Sigma_2)$ that are estimated for each track. We use a standard approach that is used in statistics, and we compute a symmetric version of the Kullback-Leibler divergence

$$KL(G^1, G^2) = \frac{1}{2}\text{tr}(\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2) + \frac{1}{2}\text{tr}(\mu_1 - \mu_2)^T \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)(\mu_1 - \mu_2) - K \tag{22}$$

Finally, the KL distance is rescaled using an exponential kernel, and we define the distance between the tracks s_1 and s_2 as

$$d(s_1, s_2) = 1 - \exp\left(-\gamma/(KL(G^1, G^2) + \texttt{epsilon})\right), \tag{23}$$

where the parameter γ is chosen in the range $10 - 100$ to optimize the classification. The parameters `epsilon` is the smallest number in MATLAB, and prevents the potential division by zero.

The MATLAB code in Fig. 11 computes the distance d between the two tracks `s1` and `s2`.

```
ico1 = inv(co1);
ico2 = inv(co2);
KL   =  trace(co1*iCo2) + trace(co2*iCo1) + (mu1-mu2)'*(iCo1+iCo2)*(mu1-mu2);
gam  = 1e2;
d    = 1 - exp(-gam/(KL + eps));
```

Figure 11: Distance d between the two tracks `s1` and `s2` .

In the unlikely event that MATLAB complains that the covariance matrix co is singular, you could use the pseudo-inverse function `pinv`, as follows:

```
ico1 = pinv(co1);
ico2 = pinv(co2);
```

4.8 Distance matrix

Assignment

5. Implement the computation of the distance d given by (23), and Fig. 11. Your function should be able to use the 12 merged MFCC coefficients or the Normalized Pitch Class Profile.

   Compute the matrix of pairwise distance

   $$D(s_1, s_2), s, s_2 = 1, \ldots, 150, \tag{24}$$

   and display the distance matrix as an image, and discuss its structure.

6. Compute the 6×6 average distance matrix between the genres, defined by

   $$\overline{D}(i, j) = \frac{1}{25^2} \sum_{s_1 \in \text{genre } i,\ s_2 \in \text{genre } j} d(s_1, s_2), \quad i, j = 1, \ldots, 6, \tag{25}$$

   Experiment with different values of γ, and find a value of γ that maximizes the separation between the different genres, as defined by the 6×6 average distance matrix $\overline{D}$.

4.9 Classification method

We are now equipped with a distance to measure the similarity between two tracks. We describe here a simple procedure to classify a song with unknown genre using the training data.

K-Nearest Neighbors

Given a song s, we wish to determine its genre. We proceed as follows.

We compute the distance between the song s and every other song in the training data, and we determine the five nearest songs (according to d defined by (23)). Among the five nearest songs, we find the genre that is the most represented, and assign this genre to the track s.

Assignment

7. Implement a classifier based on the following ingredients, as explained above,
   - computation of the 12 mfcc coefficients, or 12 Normalized Pitch Class Profile
   - modified Kullback-Leibler distance d defined by (23)
   - genre = majority vote among the 5 nearest neighbors
8. Using cross validation, as explained in section 4.12, evaluate your classification algorithm. You will compute the mean and standard deviation for all the entries in the confusion matrices.

4.10 Statistical Evaluation of the Performance of My Algorithm

4.11 Confusion matrix

The accuracy of the classification will be quantified using confusion matrices. In this lab, a confusion matrix will measure the correct classification rate of each genre. For a given classification experiment, you will construct a 6×6 matrix where the rows are the true genres, and the columns are the genres classified by your algorithm. The entry $R(i, j)$ of the confusion matrix R is the number of songs of genre i that were classified as genre j. In the example in Table 1 the total number of songs per class was:

$$\begin{bmatrix} 21 & 3 & 1 & 0 & 0 & 0 \end{bmatrix},$$

so there were 21 good identifications (diagonal entries of the matrix) out of 25 queries. This corresponds to 84 % correct answers overall. The correct classification rate per class was:

$$\begin{bmatrix} 0.84 & 0.88 & 0.52 & 0.92 & 0.36 & 0.40 \end{bmatrix}, \tag{26}$$

which yields a 51.48% correct classification if we normalize with respect to the probability of each class.

	classical	electronic	jazz	punk	rock	world
classical	21	3	1	0	0	0
electronics	3	22	0	0	0	0
jazz	0	1	13	2	8	1
punk	0	0	0	23	1	1
rock	0	2	3	6	9	5
world	3	7	1	0	4	10

Table 1: Example of a confusion matrix. The rows are the true genres, and the columns are the predicted genres.

4.12 Cross-validation

In order to evaluate the performance of your method, you will perform a 5-fold cross-validation 10 times. Each cross-validation experiment consists in the following sequence of operations.

- let $\mathcal{S}$ be the set of all 150 songs.
- let $\mathcal{G}_i, i = 1, \cdots, 6$ be the six sets of songs organized by genres:

$$\mathcal{S} = \bigcup_{i=1}^{6} \mathcal{G}_i.$$

- **repeat n =1:10** `// average over the randomization`
  1. randomly divide each genre $\mathcal{G}_i$ into 5 subsets of size 5

$$\mathcal{G}_i = \bigcup_{k=1}^{5} \mathcal{G}_i^k$$

     where $|\mathcal{G}_i^k| = |\mathcal{G}_i|/5 = 5$.
  2. **for** $k = 1$ to 5 `// round robin over the testing songs`
     - form the set of test songs

$$\mathcal{U} = \bigcup_{i=1}^{6} \{\mathcal{G}_i^k, \}$$

       and the set of training songs

$$\mathcal{L} = \bigcup_{i=1}^{6} \left\{\mathcal{G}_i^l,\ l = 1, \cdots, 5, l \neq k\right\}$$

     - test the performance of your algorithm on the 30 songs in $\mathcal{U}$ using the 120 training songs $\mathcal{L}$
     - record the confusion matrix $R^{(k,n)}$

     end for

  end repeat

- **for** each entry (i, j) of the confusion matrix:
  - compute the average $\overline{R}(i, j)$
  - compute the standard deviation $\sigma_R(i, j)$

  end for

The final average confusion matrix, and the associated standard deviation should be included in the report.