Team 23: CS470 Final Project (Option #1)

Minsung Park, Youngjin Jin, Junghyun Lee

December 1, 2019

1 Introduction

Ever since deep learning came out, it has revolutionized many different fields: from natural language processing, bioinformatics, to computer vision. Recently, deep learning has found its way in the field of music information retrieval (MIR) research. and more people are getting interested in this relatively new field. While there were 2 deep learning articles in 2010 in ISMIR (International Society for Music Information Retrieval) conferences and 6 articles in 2015, it increases to 16 articles in 2016. This is showing us the current trend trend in using deep learning in MIR research.

MIR research is an interdisciplinary science of retrieving information from music that include:

- Recommender systems
- Track separation and instrument recognition
- Automatic music transcription
- **Automatic categorization**
- Music generation

Here, we are considering the problem of **automatic categorization**; specifically, **music genre classification** i.e. classifying the give set of music audio files into different genres such as classical, jazz, hiphop...etc.

2 Audio data representations

2.1 Audio signal

Depending on the deep learning architecture that one will utilize, different "version" of input sound data is required. Also, an efficient representation of audio data is necessary so that the we can easily use it to train the given, without too heavy memory usage/computation.

The "original" representation is 1-dimensional (discrete) audio signal. It may be appropriate if one uses techniques such as RNNs or LSTMs, but if one wants to use techniques such as CNNs, the data representation has to be processed to a different one. The point is, the initial input audio data representations do matter, and this section is devoted to providing the necessary theoretical background.

2.2 Spectrum/Spectrogram

Spectrum, or energy spectral density, describes how the energy of a signal (or a time series) is distributed with frequency.

Suppose we have a time signal $x(t)$. Then the energy of the signal is defined as

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt$$

By Parseval's theorem, we have that

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{x}(f)|^2 df$$

where $\hat{x}(t)$ represents the Fourier transform of the signal.

Therefore, we can interpret $|\hat{x}(f)|^2$ as a density function describing the energy per unit frequency contained in the signal at the frequency f. From this, the natural definition of the energy spectral density of a signal is

$$S_{xx}(f) = |\hat{x}(f)|^2$$

Note that S_{xx} is a function of frequency, not a function of time. However, the spectral density of small windows of a longer signal may be calculated, plotted versus time associated with the window. Such graph is called a spectrogram. In other words, spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. A common format is a 2-D representation: one axis for time, one axis for frequency, and the intensity (or color) of each point.

The equation for deriving the spectrogram of a time signal $x(t)$ is

$$\text{spectrogram}\{x(t)\}(\tau, \omega) = |X(\tau, \omega)|^2$$

where $X(\tau, \omega)$ is the discrete-time SFTF(Short-Time Fourier Transform):

$$\textbf{STFT}\{x[n]\}(\tau, \omega) \equiv X(\tau, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-\tau]e^{-jwn}$$

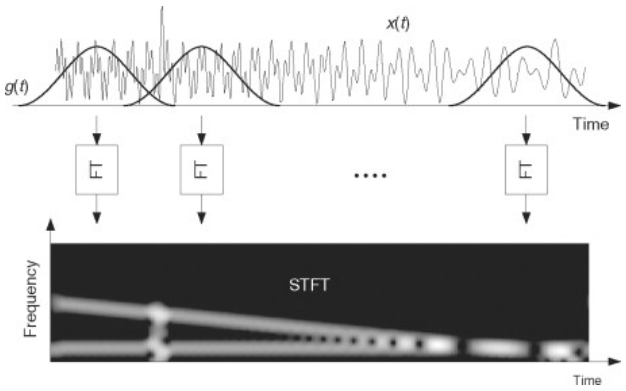

Figure 1: Computing STFT (g is the (sliding) window function)

Here, $\omega[n]$ is the window function, typically Hann window or Gaussian window. (Window function is a function that is zero-valued outside some chosen interval, and usually symmetric.)

2.3 Mel scale/Mel spectrogram

Mel scale, introduced by Stevens, Volkmann, and Newman in 1937, is a perceptual scale of pitches judged by listeners to be equal in distance from one another. In other words, the Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also "sound" to humans as they are equal in distance from one another. (Think of Mel scale as a psychoacoustic scale)

There isn't a "single" Mel scale formula, but the popular formula from O'Shaughnessy's book is as the following:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Mel spectrogram is the combination of above concepts: spectrogram with Mel scale as its y-axis.

2.4 Mel-frequency cepstrum (MFC)

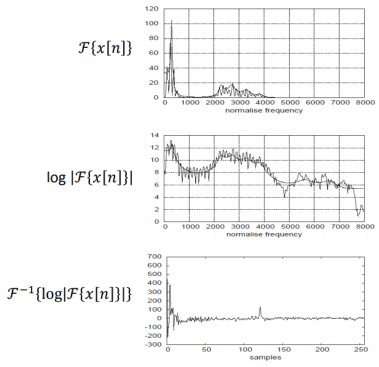

Figure 2: Computing cepstrum

The power cepstrum of a time signal $x[n]$, as defined by Bogert *et al.* in 1963, is

$$c[n] = \left|\mathcal{F}^{-1}\left\{\log\left(|\mathcal{F}\{x[n]\}|^2\right)\right\}\right|^2$$

Intuitively, this gives us the information about the rate of change in the different spectrum bands. This is often utilized as a feature vector for representing musical signals.

But in musical signals, the spectrum is usually first transformed using the Mel scale. The result is called the Mel-frequency cepstrum (MFC), and we call its coefficients as the Mel-frequency cepstral coefficients (MFCCs). In the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs of $x[n]$ are commonly derived as follows (with some variantions in the process, depending on the task):

1. Take the Fourier transform of a windows excerpt of $x[n]$
2. Map the powers of the obtained spectrum onto the Mel scale (using triangular overlapping windows)
3. Take the logs of the powers at each of the mel frequencies
4. Take the discrete cosine transform of the list of mel log powers
5. The MFCCs are the amplitudes of the resulting spectrum

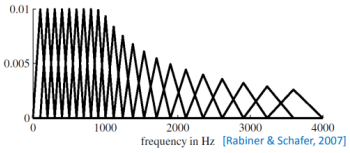

Figure 3: $V_r[k]$

Summarizing above using mathematical expressions:

$$MFCC[m] = \frac{1}{R}\sum_{r=1}^{R}\log(MF[r])\cos\left[\frac{2\pi}{R}\left(r+\frac{1}{2}\right)m\right]$$

where

$$MF[r] = \left(\sum_{k=L_r}^{U_r}|V_r[k]|^2\right)^{-1}\sum_{k=L_r}^{U_r}|V_r[k]X(n,k)|$$

($V_r[k]$ is the triangular weighting function for the r-th filter, ranging from DFT index L_r to U_r. Usually, $MFCC[m]$ is evaluated for a number of coefficients N_{MFCC} that is less than the number of mel-filters R)

2.5 Spectral statistics

2.5.1 Spectral centroid

Spectral centroid indicates where the "center of mass" of the mass of the spectrum is located *i.e.* it represents *at which frequency the energy of a spectrum is centered upon.*

$$f_c = \frac{\sum_k S(k)f(k)}{\sum_k S(k)}$$

where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k.

2.5.2 (Octave-based) Spectral Contrast

MFCC is a representation of spectral characteristics of music that averages the spectra in each sub-band and reflects the average spectral characteristics. (It is sometimes called the average spectral envelope) But MFCC cannot represent the relative spectral characteristics in each sub-band, which might be more important in our music genre classification problem.

Octave-based Spectral Contrast feature considers the strength of spectral peaks and spectral valleys in each sub-band separately to represent the relative spectral characteristics. For most music, the strong spectral spectral peaks roughly correspond with harmonic components; while non-harmonic components, or noises, often appear at spectral valleys. This is why the *Spectral Contract feature roughly reflects the relative distribution of harmonic and nonharmonic components in the spectrum*, while MFCC loses that information.

To make this point more clear, consider two spectra with different spectral distribution and similar average spectral characteristics. Then, MFCC won't be able to distinguish between them while Spectral Contrast may perform better.

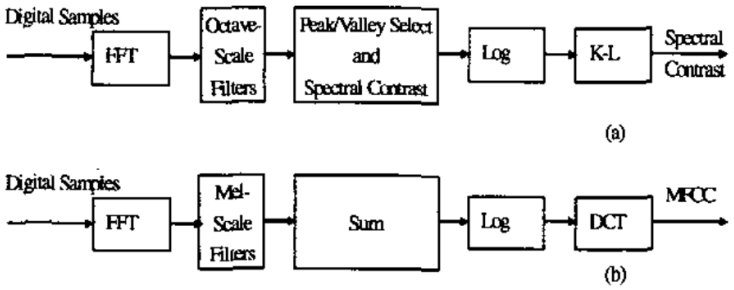

Figure 4: Comparison of (a) Octave-based Spectral Contrast and (b) MFCC

Main differences between the Spectral Contrast and MFCC are...

- The filter bank is different; Octave-based Spectral Contrast uses octave-scale filters while MFCC uses Mel-scale filters. (Although Mel-scale is suitable for general auditory model, octave-scale filter is more suitable for music processing.)
- Spectral Contrast extracts the spectral peaks, valleys, and their differences in each sub-band while MFCC sums the FFT amplitudes
- Spectral Contrast feature uses the Karhunen-Loéve Transform (KLT) while MFCC uses a Discrete Cosine Transform (DCT).

To solve the problem of some relativity among the different dimensions of raw feature, the KLT is performed. Then by the Karhunen-Loéve Theorem, the feature vector is mapped into an orthogonal space, and the covariance matrix is diagonalized. These properties of KLT make the classifying procedure easier and lead to good classification performance, even with simple classifier with Spectral Contrast (Jiang *et al.*).

2.5.3 Chroma

...?

3 Method

Please refer to our github repository[1] for the code.

[1] `https://github.com/iJinjin/AI_Team23_Proj`

4 Contributions

- Minsung Park: TBD
- Youngjin Jin: TBD
- Junghyun Lee: TBD