Chroma Feature Extraction

A. K. Shah, M. Kattel, A. Nepal, D. Shrestha
Department of Computer Science and Engineering, School of Engineering
Kathmandu University, Nepal

ayush.kumar.shah@gmail.com, manasikattel1@gmail.com, araju7nepal@gmail.com, deepeshshrestha@outlook.com

Abstract

The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore chroma features can be considered as an important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. A better quality of the extracted chroma feature enables much better results in these high-level tasks. Short-Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.

Keywords: Fourier transform, spectrogram, chroma representation, chroma vector

1. Introduction

Over the past few years the need of music information retrieval and classification systems has become more urgent and this brought to the birth of a research area called Music Information Retrieval (MIR). It is an important task in the analysis of music and music transcription in general, and it can contribute to applications such as key detection, structural segmentation, music similarity measures, and other semantic analysis tasks.

Pitch

Pitch is a perceptual property of sounds that allows their ordering on a frequency-related scale, or more commonly, pitch is the quality that makes it possible to judge sounds as "higher" and "lower" in the sense associated with musical melodies. Pitch can be determined only in sounds that have a frequency that is clear and stable enough to distinguish from noise. Pitch is a major auditory attribute of musical tones, along with duration, loudness, and timbre [1].

Chroma

Chroma feature, a quality of a pitch class which refers to the "color" of a musical pitch, which can be decomposed in into an octave-invariant

value called "chroma" and a "pitch height" that indicates the octave the pitch is in [2].

Chroma Vector

A chroma vector is a typically a 12-element feature vector indicating how much energy of each pitch class, {C, C#, D, D#, E, ..., B}, is present in the signal. The Chroma vector is a perceptually motivated feature vector. It uses the concept of chroma in the cyclic helix representation of musical pitch perception. The Chroma vector thus represents magnitudes in twelve pitch classes in a standard chromatic scale [3].

Chroma features

In music, the term chroma feature or chromagram closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale. One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation.

Chroma features aim at representing the harmonic content (eg:keys,chords) of a short-time window of audio. The feature vector is extracted from the magnitude spectrum by using a short time fourier transform(STFT), Constant-Q transform(CQT), Chroma Energy Normalized (CENS), etc[5].

Harmonic Pitch Class Profile (HPCP)

Harmonic pitch class profiles (HPCP) is a group of features that a computer program extracts from an audio signal, based on a pitch class profile—a descriptor proposed in the context of a chord recognition system. HPCP are an enhanced pitch distribution feature that are sequences of feature vectors that, to a certain extent, describe tonality, measuring the relative intensity of each of the 12 pitch classes of the equal-tempered scale within an analysis frame. Often, the twelve pitch spelling attributes are also referred to as chroma and the HPCP features are closely related to chroma features or chromagrams [4].

2. Background

Features of audio

- Frequency
- Amplitude

Frequency is the speed of the vibration, and this determines the pitch of the sound. It is only useful or meaningful for musical sounds, where there is a strongly regular waveform. It is measured as the number of wave

cycles that occur in one second. The unit of frequency measurement is Hertz.

Amplitude is the size of the vibration, and this determines how loud the sound is. We have already seen that larger vibrations make a louder sound. Amplitude is important when balancing and controlling the loudness of sounds, such as with the volume control on your CD player.

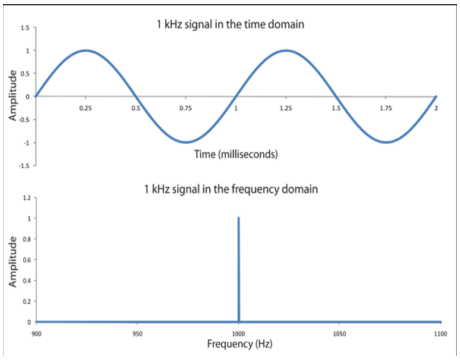


Fig 2.1: Features of audio

2.1 Chroma Features

The underlying observation is that humans perceive two musical pitches as similar in color if they differ by an octave. Based on this observation, a pitch can be separated into two components, which are referred to as tone height and chroma. Assuming the equal-tempered scale, one considers twelve chroma values represented by the set

{C, C♯, D, D♯, E ,F, F♯, G, G♯, A, A♯, B}

that consists of the twelve pitch spelling attributes as used in Western music notation. Note that in the equal-tempered scale different pitch spellings such C♯ and D♭ refer to the same chroma. Enumerating the chroma values, one can identify the set of chroma values with the set of integers {1,2,...,12}, where 1 refers to chroma C, 2 to C♯, and so on. A pitch class is defined as the set of all pitches that share the same chroma. For example, using the scientific pitch notation, the pitch class corresponding to the chroma C is the set

{..., C−2, C−1, C0, C1, C2, C3 ...}

consisting of all pitches separated by an integer number of octaves. Given a music representation (e.g. a musical score or an audio recording), the main idea of chroma features is to aggregate for a given local time window (e.g. specified in beats or in seconds) all information that relates to a given chroma into a single coefficient. Shifting the time window across the music representation results in a sequence of chroma features each expressing how the representation's pitch content within the time window is spread over the twelve chroma bands. The resulting time-chroma representation is also referred to as chromagram. The figure below shows chromagrams for a C-major scale, once obtained from a musical score and once from an audio recording [5].

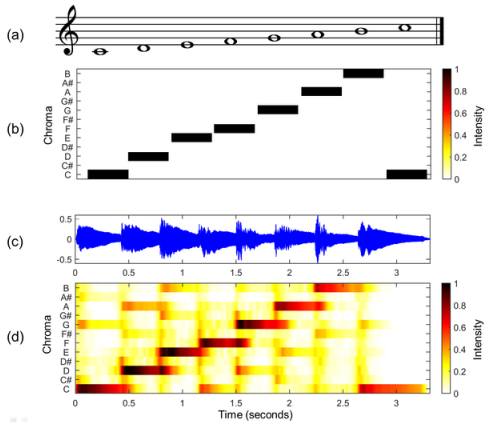


Fig 2.2 : (a) Musical score of a C-major scale. (b) Chromagram obtained from the score. (c) Audio recording of the C-major scale played on a piano. (d) Chromagram obtained from the audio recording.

2.1.1 Types of Chroma Features [7]

2.1.1.1 CP Feature

From the Pitch representation, one can obtain a chroma representation by simply adding up the corresponding values that belong to the same chroma. To archive invariance in dynamics, we normalize each chroma vector with respect to the Euclidean norm. The resulting features are referred to as Chroma-Pitch denoted by CP.

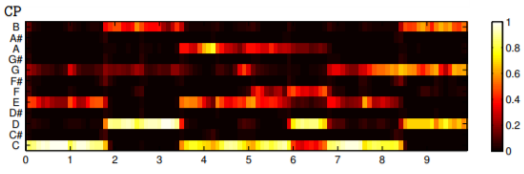


Fig 2.3 : CP Feature

2.1.1.2 CLP Features

To account for the logarithmic sensation of sound intensity, one often applies a logarithmic compression when computing audio features [11]. To this end, the local energy values e of the pitch representation are logarithmized before deriving the chroma representation.

Here, each entry e is replaced by the value

$\log(\eta \cdot e+1)$, where η is a suitable positive constant. The resulting features, which depend on the compression parameter η, are referred to as Chroma-Log-Pitch denoted by CLP[η].

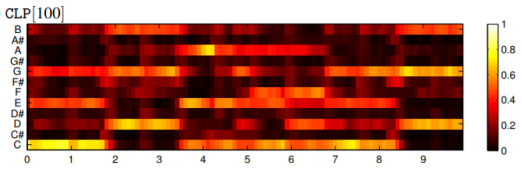


Fig 2.4 : CLP Features

2.1.1.3. CENS Features

Adding a further degree of abstraction by considering short-time statistics over energy distributions within the chroma bands, one obtains CENS (Chroma Energy Normalized Statistics) features, which constitute a family of scalable and robust audio features. These features have turned out to be very useful in audio matching and retrieval applications. In computing CENS features, a quantization is applied based on logarithmically chosen thresholds.

This introduces some kind of logarithmic compression similar to the CLP[η] features. Furthermore, these features allow for introducing a temporal smoothing. Here, feature vectors are averaged using a sliding window technique depending on a window size denoted by w (given in frames) and a downsampling factor denoted by d. In the following, we do not change the feature rate and consider only the case d = 1 (no downsampling). Therefore, the resulting feature only depends on the parameter w and is denoted by CENS[w].

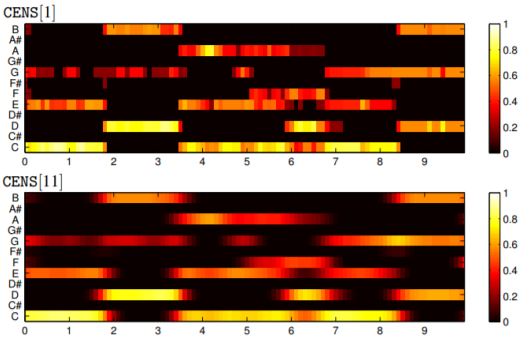


Fig 2.5 : CENS Features

2.1.1.4 CRP Features

To boost the degree of timbre invariance, a novel family of chroma-based audio features has been introduced. The general idea is to discard timbre-related information in a similar fashion as pitch-related information is discarded in the computation of mel-frequency cepstral coefficients (MFCCs). Starting with the Pitch features, one first applies a logarithmic compression and transforms the logarithmized pitch representation using a DCT. Then, one only keeps the upper coefficients of the resulting pitch-frequency cepstral coefficients (PFCCs), applies an inverse DCT, and finally projects the resulting pitch vectors onto 12-dimensional chroma vectors.

These vectors are referred to as CRP (Chroma DCT Reduced log Pitch) features. The upper coefficients to be kept are specified by a parameter p ∈ [1 : 120].

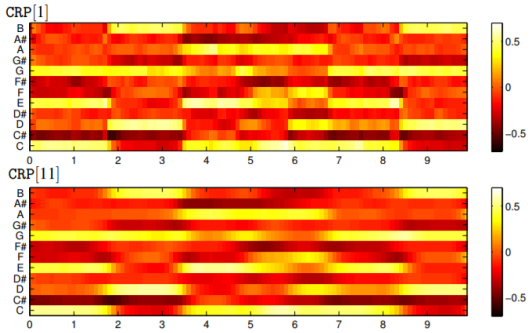


Fig 2.6 :CRP Features

3. Methodology

General chroma feature extraction procedure

The block diagram of the procedure is shown in Fig.

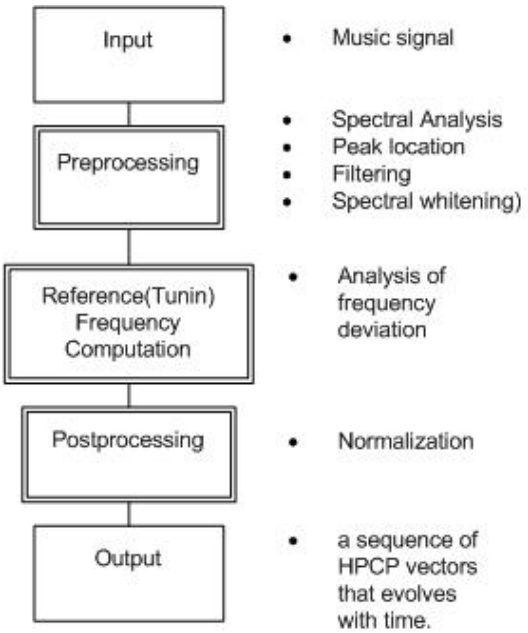


Fig 3.1: HPCP block diagram

The General HPCP (chroma) feature extraction procedure is summarized as follows:

1. Input musical signal.
2. Do spectral analysis to obtain the frequency components of the music signal.
3. Use Fourier transform to convert the signal into a spectrogram. (The Fourier transform is a type of time-frequency analysis.)
4. Do frequency filtering. A frequency range of between 100 and 5000 Hz is used.
5. Do peak detection. Only the local maximum values of the spectrum are considered.
6. Do reference frequency computation procedure. Estimate the deviation with respect to 440 Hz.
7. Do Pitch class mapping with respect to the estimated reference frequency. This is a procedure for determining the pitch class value from frequency values. A weighting scheme with cosine function is used. It considers the presence of harmonic frequencies (harmonic summation procedure), taking account a total of 8 harmonics for each frequency. To map the value on a one-third of a semitone, the size of the pitch class distribution vectors must be equal to 36.
8. Normalize the feature frame by frame dividing through the maximum value to eliminate dependency on global loudness. And then we can get a result HPCP sequence like Figure [4].

There are many ways for converting an audio recording into a chromagram. For example, the conversion of an audio recording into a chroma representation (or chromagram) may be performed either by using short-time Fourier transforms in combination with binning strategies or by employing suitable multirate filter banks. Furthermore, the properties of chroma features can be significantly changed by introducing suitable pre- and post-processing steps modifying spectral, temporal, and dynamical

aspects. This leads to a large number of chroma variants, which may show a quite different behavior in the context of a specific music analysis scenario [5].

3.1 Performing Fourier Transform

3.1.1 Short Time Fourier Transform

The Fourier transform maps a time-dependent signal to a frequency-dependent function which reveals the spectrum of frequency components that compose the original signal. Loosely speaking, a signal and its Fourier transform are two sides of the same coin. On the one side, the signal displays the time information and hides the information about frequencies. On the other side, the Fourier transform reveals information about frequencies and hides the time information [8].

To obtain back the hidden time information, Dennis Gabor introduced in the year 1946 the modified Fourier transform, now known as short-time Fourier transform or simply STFT. This transform is a compromise between a time- and a frequency-based representation by determining the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In this way, the STFT does not only tell which frequencies are "contained" in the signal but also at which points of times or, to be more precise, in which time intervals these frequencies appear.

The Short-Time Fourier Transform (STFT) is a powerful general-purpose tool for audio signal processing. It defines a particularly useful class of *time-frequency distributions* which specify complex amplitude versus time and frequency for any signal. Fourier transform is a well-known tool for analyzing the frequency distribution of a signal. Let us denote the uniformly sampled f(t) and g(t) functions by f[n] and g[n]. Then the discrete (D) STFT over a compactly supported g window function can be written as

$$\mathcal{F}_g f[n,k] = \sum_{m=0}^{M-1} f[n-m]g[m]\epsilon_k[m]\,,$$

Where

$$\epsilon_k[m] = e^{-2\pi m \frac{k}{N}}$$

M is the window length of g and N is the number of samples in f. This algorithm can be interpreted as a successive evaluation of Fourier transforms over short segments of the whole signal. Additionally, the frequencies can be visually represented by displaying the squared magnitude of the Fourier coefficients at each section. This diagram is called as the spectrogram of the signal f [15].

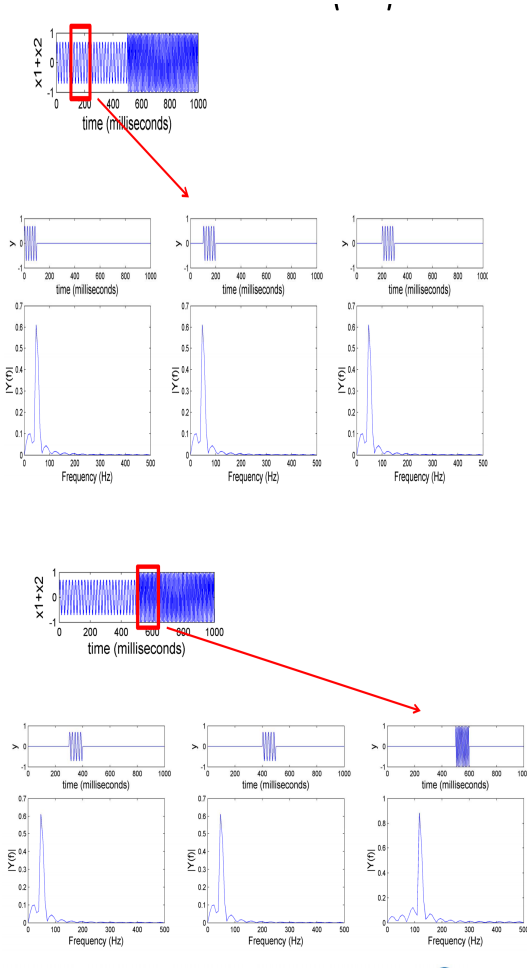


Fig 3.2 : Short Time Fourier Transform [9]

3.1.2 Constant-Q Transform

Like the Fourier transform a constant Q transform is a bank of filters, but in contrast to the former it has geometrically spaced center frequencies:

$$f_k = f_0 \cdot 2^{\frac{k}{b}} \quad (k = 0, \ldots),$$

where b dictates the number of filters per octave.

What makes the constant Q transform so useful is that by an appropriate choice for f0 (minimal center frequency) and b the center frequencies directly correspond to musical notes.

Another nice feature of the constant Q transform is its increasing time resolution towards higher frequencies. This resembles the situation in our auditory system. It is not only the digital computer that needs more time to perceive the frequency of a low tone but also our auditory sense. This is related to music usually being less agitated in the lower registers.The constant Q-transform can be viewed as a wavelet transform. There are at least three reasons why the CQT has not widely replaced the DFT in audio signal processing. Firstly, it is computationally more intensive than the DFT. Secondly, the CQT lacks an inverse transform that would allow perfect reconstruction of the original signal from its transform coefficients. Thirdly, CQT produces a data structure that is more difficult to work with than the time-frequency matrix (spectrogram) obtained by using short-time Fourier transform in successive time frames. The last problem is due to the fact that in CQT, the time resolution varies for different frequency bins, in effect meaning that the ”sampling” of different frequency bins is not synchronized [20].

3.2 Log-Frequency Spectrogram

We now derive some audio features from the STFT by converting the frequency axis (given in Hertz) into an axis that corresponds to musical

pitches. In Western music, the equal-tempered scale is most often used, where the pitches of the scale correspond to the keys of a piano keyboard. In this scale, each octave (which is the distance of two frequencies that differ a factor of two) is split up into twelve logarithmically spaced units. In MIDI notation, one considers 128 pitches, which are serially numbered starting with 0 and ending with 127. The MIDI pitch p = 69 corresponds to the pitch A4 (having a center frequency of 440 Hz), which is often used as standard for tuning musical instruments. In general, the center frequency $F_{pitch}(p)$ of a pitch $p \in [0 : 127]$ is given by the formula

$$F_{pitch}(p) = 2^{(p-69)/12} .440$$

The logarithmic perception of frequency motivates the use of a time-frequency representation with a logarithmic frequency axis labeled by the pitches of the equal-tempered scale. To derive such a representation from a given spectrogram representation, the basic idea is to assign each spectral coefficient X (m, k) to the pitch with center frequency that is closest to the frequency $F_{coef}(k)$. More precisely, we define for each pitch $p \in [0 : 127]$ the set

$$P(p) := \{k \in [0 : K] : Fpitch(p - 0.5) \leq F_{coef}(k) < F_{pitch}(p + 0.5)\}.$$

From this, we obtain a log-frequency spectrogram $Y_{LF} : Z \times [0 : 127] \rightarrow R_{\geq 0}$ defined by

$$\mathcal{Y}_{\mathrm{LF}}(m,p) := \sum_{k \in P(p)} |\mathcal{X}(m,k)|^2.$$

By this definition, the frequency axis is partitioned logarithmically and labeled linearly according to MIDI pitches [8].

3.3 Chroma Features Extraction

The human perception of pitch is periodic in the sense that two pitches are perceived as similar in "color" (playing a similar harmonic role) if they differ by one or several octaves (where, in our scale, an octave is defined as the distance of 12 pitches). For example, the pitches p = 60 and p = 72 are one octave apart, and the pitches p = 57 and p = 71 are two octaves apart. A pitch can be separated into two components, which are referred to as tone height and chroma. The tone height refers to the octave number and the chroma to the respective pitch spelling attribute. In Western music notation, the 12 pitch attributes are given by the set
{C, C, D, . . . , B} [8].

Enumerating the chroma values, we identify this set with [0 : 11] where c = 0 refers to chroma C, c = 1 to C], and so on. A pitch class is defined as the set of all pitches that share the same chroma. For example, the pitch class that

corresponds to the chroma c = 0 (C) consists of the set
{0, 12, 24, 36, 48, 60, 72, 84, 96, 108, 120}
(which are the musical notes
{. . . , C0, C1, C2, C3 . . .}).

The main idea of chroma features is to aggregate all spectral information that relates to a given pitch class into a single coefficient. Given a pitch-based log-frequency spectrogram Y_{LF} :
$Z \times [0 : 127] \rightarrow R_{\geq 0}$, a chroma representation or chromagram
$Z \times [0 : 11] \rightarrow R_{\geq 0}$
can be derived by summing up all pitch coefficients that belong to the same chroma:

$$\mathcal{C}(m,c) := \sum_{\{p \in [0:127] \mid p \bmod 12 = c\}} \mathcal{Y}_{\mathrm{LF}}(m,p)$$

4. Applications

Identifying pitches that differ by an octave, chroma features show a high degree of robustness to variations in timbre and closely correlate to the musical aspect of harmony. This is the reason why chroma features are a well-established tool for processing and analyzing music data. For example, basically every chord recognition procedure relies on some kind of chroma representation. Also, chroma features have become the de facto standard for tasks such as music alignment and synchronization as well as audio structure analysis. Finally, chroma features have turned out to be a powerful mid-level feature representation in content-based audio retrieval such as cover song identification or audio matching [5].

Example

Chord recognition:

We did a Machine Learning project called Guitar chord recognition that recognizes any guitar chord. For chord recognition, chroma feature extraction was used in the guitar chords dataset. In this work, we investigated Convolutional Neural Networks (CNNs) for learning chroma features in the context of chord recognition.

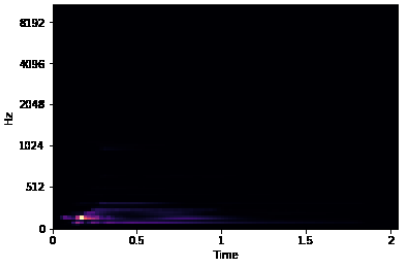
A large set of data was collected and metadata Chords.csv was created which contained information such as class_id, classname, file_name of all the audio guitar chords datasets.

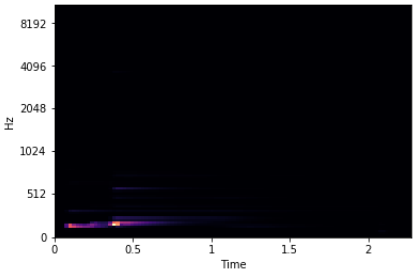
Librosa library was used to extract chromagram from the audio datasets. A new dataset consisting of mel spectrograms of the chroma features of all the audio files in the guitar chord datasets as input and corresponding class_id as output was built. Since the audio files in the dataset are of varying duration (up to 4 s), th fixed the size of the input taken was to 2 seconds (128 frames), i.e. $X \in \mathbf{R}^{128 \times 87}$

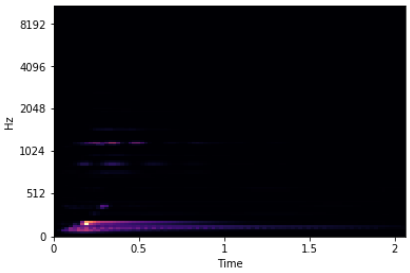
The new dataset was then shuffled and divided to training and test

datasets and was further preprocessed, fed to a convolution neural network, trained and performance metrics were evaluated on the basis of classes of chords predicted from the chroma features of the test audio dataset.

Hence, chroma feature extraction was useful to train the audio dataset for guitar chord recognition since audio dataset cannot be trained directly.







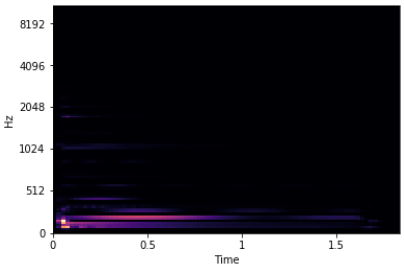


Fig 4.1: Mel Spectrogram Chroma features of audio files am89.wav, em69.wav, g200.wav and dm100.wav respectively from guitar chord dataset

The above screenshots show the results obtained from the audio data available after extracting the chroma features of the audio files using librosa.

5. Acknowledgement

This paper was written to fulfill the course requirement of COMP-407 Digital Signal Processing(DSP) offered by the Department of Computer Science and Engineering (DOCSE).The authors, would like to thank Mr. Satyendra Nath Lohani, Assistant Professor of Department of Computer Science and Engineering for providing this wonderful opportunity to explore and gain new ideas and experience in a new topic.

6. Conclusion

This paper presents the details of chroma feature extraction from any audio files and the different types of extraction methods of the chroma feature

are explained.The short term fourier transform proved to be better than the constant Q transform under chroma feature extraction mainly because it does not have inverse transform so that the original audio form could be regained.This being the main reason for the use of short term fourier transform in the chord recognition project. Experimental results using STFT chroma feature extraction is presented.

References:

[1] "Pitch (music)", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Pitch_(music). [Accessed: 23- Jan- 2019].

[2]"Chroma", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Chroma. [Accessed: 23- Jan- 2019].

[3]"chroma", Musicinformationretrieval.com, 2019. [Online]. Available: https://musicinformationretrieval.com/chroma.html. [Accessed: 23- Jan- 2019].

[4]L. Revolvy, ""Harmonic pitch class profiles" on Revolvy.com", Revolvy.com, 2019. [Online]. Available: https://www.revolvy.com/page/Harmonic-pitch-class-profilesAW. [Accessed: 23- Jan- 2019].

[5]"Chroma feature", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Chroma_feature. [Accessed: 18- Jan- 2019].

[6]Cs.uu.nl, 2019. [Online]. Available: http://www.cs.uu.nl/docs/vakken/msmt/lectures/SMT_B_Lecture5_DSP_2017.pdf. [Accessed: 19- Jan- 2019].

[7]Pdfs.semanticscholar.org, 2019. [Online]. Available: https://pdfs.semanticscholar.org/6432/19014e8aa48dda060cecf4ff413dd3ee1e3a.pdf. [Accessed: 23- Jan- 2019].

[8]Audiolabs-erlangen.de, 2019. [Online]. Available: https://www.audiolabs-erlangen.de/content/05-fau/professor/00-mueller/02-teaching/2016s_apl/LabCourse_STFT.pdf. [Accessed: 23- Jan- 2019].

[9]2019. [Online]. Available: http://mac.citi.sinica.edu.tw/~yang/tigp/lecture02_stft_yhyang_mir_2018.pdf. [Accessed: 18- Jan- 2019].

[10]Mirlab.org, 2019. [Online]. Available: http://www.mirlab.org/conference_papers/international_conference/ICASSP%202014/papers/p5880-kovacs.pdf. [Accessed: 19- Jan- 2019].

[11]Scholarship.claremont.edu, 2019. [Online]. Available: 6 ttps://scholarship.claremont.edu/cgi/viewcontent.cgi?referer=https://www.google.com.np/&httpsredir=1&article=1575&context=cmc_theses. [Accessed: 18- Jan- 2019].

[12]Atkison.cs.ua.edu, 2019. [Online]. Available: http://atkison.cs.ua.edu/papers/FT_as_FE.pdf. [Accessed: 17- Jan- 2019].

[13]"Advances in Music Information Retrieval", Google Books, 2019. [Online]. Available: https://books.google.com.np/books?id=gy5qCQAAQBAJ&pg=PA340&lpg=PA340&dq=Chroma+Feature+Extraction+using+Short+Time+Fourier+Transform(STFT)&source=bl&ots=JSxWIB9vm9&sig=ACfU3U1oZHnD8Ohu-YGLaTycJXZLB895DA&hl=ne&sa=X&ved=2ahUKEwj0uY-B6YDgAhVZaCsKHXBYAhcQ6AEwCXoECAEQAQ#v=onepage&q=Chroma%20Feature%20Extraction%20using%20Short%20Time%20Fourier%20Transform(STFT)&f=false. [Accessed: 17- Jan- 2019].

[14]2019. [Online]. Available: https://www.researchgate.net/figure/Feature-extraction-process-using-short-time-Fourier-transform-STFT-Spectrograms-of-A_fig1_236264209. [Accessed: 22- Jan- 2019].

[15]"The Short-Time FourierTransform | Spectral Audio Signal Processing", Dsprelated.com, 2019. [Online]. Available: https://www.dsprelated.com/freebooks/sasp/Short_Time_Fourier_Transform.html. [Accessed: 22- Jan- 2019].

[16]Kingma, D., & Ba, J. (2019). Adam: A Method for Stochastic Optimization. Retrieved from https://arxiv.org/abs/1412.6980v8

[17]Arxiv.org, 2019. [Online]. Available: https://arxiv.org/pdf/1811.01222.pdf. [Accessed: 22- Jan- 2019].

[18]2019. [Online]. Available: https://www.researchgate.net/publication/290632086_Evaluation_and_comparison_of_audio_chroma_feature_extraction_methods. [Accessed: 22- Jan- 2019].

[19]Iem.kug.ac.at, 2019. [Online]. Available: https://iem.kug.ac.at/fileadmin/media/iem/projects/2010/smc10_schoerkhuber.pdf. [Accessed: 22- Jan- 2019].

[20]Doc.ml.tu-berlin.de, 2019. [Online]. Available: http://doc.ml.tu-berlin.de/bbci/material/publications/Bla_constQ.pdf. [Accessed: 22- Jan- 2019].