A.T.O.M. - A TOOL FOR MUSIC TRANSCRIPTION

A Thesis
submitted in partial fulfillment of the requirements for
the award of the degree of

BACHELOR OF ENGINEERING
IN
ELECTRONICS AND COMMUNICATION ENGINEERING

SUBMITTED BY:

Tanvi Sahay	(BE/10502/2012)
Pooshkar Rajiv	(BE/10225/2012)
Arpit Aggarwal	(BE/10174/2012)

SUPERVISED BY:	Dr. Mahesh Chandra	(ECE)
	Dr. Rajeev Kumar	(ME)

Department of Electronics and Communication Engineering
Birla Institute of Technology
Mesra, Ranchi - 835215
May 4, 2016

Declaration certificate

This is to certify that the work presented in the thesis entitled "**A.T.O.M. - A TOOL FOR MUSIC TRANSRIPTION**" in partial fulfillment of the requirement for the award of the Degree of **Bachelor of Engineering** in **Electronics and Communication Engineering** of Birla Institute of Technology, Mesra, Ranchi is an authentic work carried out under my supercisino and guidance.

To the best of my knowledge, the content of this thesis does not form a basis for the award of any previous Degree to anyone else.

Date:

Dr. Mahesh Chandra

Dept. of Electronics and Comm. Engg.
Birla Institute of Technology
Mesra, Ranchi - 835215

Head of Dept.
Electronics and Comm. Engg.
Birla Institute of Technology
Mesra, Ranchi - 835215

Certificate of Approval

The foregoing thesis entitled **"A.T.O.M. - A TOOL FOR MUSIC TRANSCRIPTION"**, is hereby approved as a creditable study of research topic and has been presented in a satisfactory manner to warrant its acceptance as prerequisite to the degree for which it has been submitted.

It is uderstood that by this approval,the undersigned do not necessarily endorse any conclusion drawn or opinion therein, but approve the thesis for the purpose for which it is submitted.

(Internal Examiner)

(External Examiner)

(chairman)
Head of Department
Dept. of Electronics and Comm. Engg.
Birla Institute of Technology, Mesra

Acknowledgement

The success of an endeavor depends not only on the idea and the work but also on the people associated with it. We would like to take this opportunity to extend our heartfelt gratitude to all those without whom this project would never have been possible.

First and foremost, we would like thank our guides, **Dr. Rajeev Kumar** and **Dr. Mahesh Chandra**, whose constant guidance and support enabled us to follow all goals through and overcome the obstacles we faced. We would also like to thank **Mr. Arun Dayal Udai** for taking keen interest in the project and inspiring us to delve deeper into the available technologies to improve the quality of both our project and its documentation.

We also extend our gratitude towards **Mr. Mrinal Pathak** for motivating us to pursue a project in the field of music and for helping us understand the basics of music and the impact of the project over the music community.

Signed .... Date ....

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed . . . . Date . . . .

Abstract

Lack of proper resources and potent technology have made transcription of music pieces to its written sheet music format and its easily accessible and secure storage a major problem within the music community. Not only have insecure storage methods such as cell phones and computers led to an increasing rise in the cases of stolen or leaked music pieces, mishaps to these fragile devices have also caused a great loss to musicians and their loved ones the world over. Though applications and devices have been developed to recognize musical notes, most require an excessive amount of paraphernalia, few convert the recognized notes to sheet music and none have the ability to provide safe storage to these converted music sheets. Through this project, development of the device A.T.O.M. - A TOol for Music transcription, has been proposed, to allow automated recording, transcription and storage of music pieces in its sheet format onto a secure server, which only the particular user will have access to. The purpose of this device is to allow musicians to store and retrieve their creations easily from any part of the globe by connecting to the secure server and accessing required files by submitting the appropriate login identification and password. The device will make use of audio as well as image processing to provide dual identification of the note being played, utilize cloud computing to store the converted music onto a secure remote server along with providing an easy to use and compact user interface.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the past 70 years, inclusion of technology to different spheres of life has had a major impact on shaping the way we perceive and interact with the world today. These changes are not only limited to the way we live, commute and communicate but have spread to the field of arts as well and has, in many ways, revolutionized the artistic community. From invention of the transistor in 1947, which facilitated the development of new portable and more complex synthesizers, to development of Apple's iTunes and iPod that replaced vinyl tapes and cassettes as the accepted method of music delivery, the music industry has been influenced and improved by technology in countless ways and change is still underway.

The term *Music Technology* refers to any technology, such as a computer, an effects unit or a piece of software, that is used by a musician to help make music, especially the use of electronic devices and computer software to facilitate playback, recording, composition, storage, mixing, analysis, editing, and performance. Everything from the microphone to the accouterments associated with an electric guitar can be considered a part of music technology. However, where many inventions have taken place that change the way music is produced, recorded and modified, not much has been undertaken in the field of transcription and storage of music itself. Transcription of music refers to conversion of the notes played using an instrument to their written format, while storage means safekeeping of this sheet music in order to ensure that music piece is well documented and available for future use. With the advent of various recording devices such as cell phones and computers, the method of music transcription by hand has become long lost. However, many a times, these fragile new age devices suffer irreparable damage, which causes loss of the information stored in them, including the recorded musical sequences. This problem has plagued people involved in the music industry all over the world as they suffer a huge loss when the equipment containing their recordings is lost or damaged. Automatic music transcription comes under the branch of study known as Music Information Retrieval and this project presents a methodology to achieve this transcription using signal processing and machine learning.

1.1 Literature Review

One of the most sophisticated technologies currently in use is a device known as MIDI or Musical Instrument Digital Interface, which allows electronic instruments and other digital musical tools to communicate with each other. The device recognizes when a note is on and when it is off but it has no prospect of converting the notes to their written format or storing the recognized notes in a secure location. MIDI also has a considerable lag and requires bulky paraphernalia to fulfill its purpose. The high cost of the equipment is another factor that discourages musicians from investing in it.

Music Infromation retrieval has become a prominent area of research and with the growing influence of music in fields such as therapy, automatic transcription of music is quickly becoming one of the most sought after projects among the research community. In the past, computer vision has often been used to detect the notes being played with instruments from the guitar family as well as keyboard instruments. However, real time transcription of these recognized notes is difficult as each image needs to be processed separately before it can be determined where the finger of the player is resting. Chord estimation of guitar has been achieved using audio processing but image processing has not been employed for such cases. Isolated notes have been recognized using audio processing and image processing. However, the two have not been employed in unison to crete both sheet music as well as tablature of the notes being played.

1.2 Introduction to A.T.O.M.

A.T.O.M: A TOol for Music transcription, is a service that proposes to provide its user with the ability to automatically save the music played by him or her in the form of sheet music, store it on a remote server and access it via internet from any part of the globe. ATOM makes use of audio as well as image processing, along with cloud computing and designing and simulation for development of a user-friendly and compact interface. Appropriate audio processing is necessary for recognition of the note being played and its subsequent conversion to sheet music. However, many a times, same notes are repeated multiple times at different positions in a musical instrument. A.T.O.M. proposes to employ image processing in order to provide a confirmation of the exact location of the note being played by the user. Once the notes have been converted to sheet music, they can be stored on an online server where they will only be accessible using the appropriate password. Accessibility of the sheet music via internet as well as lack of on-site servers to store, manage and process data make cloud computing an integral part of the project. Finally, the ultimate aim of the project is to develop a device that is easy to use and has a friendly interface. Therefore, optimization of the device in term so of size and additional equipment requirements to ensure maximum utilization of the product developed play a crucial role in the completion of the project.

The need for A.T.O.M., as has aready been clarified above, is to provide a hassle-free, secure online transcription and storage platform to every musician across the world. In the following sections, the method of data acquisition and details of the prepared database have been given, followed by a brief overview of audio recognition, the techniques used for extraction of features of a music piece which are to be employed for the purpose of classification and the method of their classification. Concepts of Image Processing used for detection of fretboard and recognition of position the musician's hand have also been explained. A detailed explanation of the Internet of Things aspect involved in the project has then been provided.

1.3 Creating A.T.O.M.

A.T.O.M. can be broadly divided into four major fields as:

- Data Acquisition
- Signal Processing and Classification
- Transcription
- Server Back End

Signal Processing and Classification can further be divided into: Audio Signal Processing, Classification and Image Processing. The thesis presents detailed theory of each method employed, method of application and tools and software used in the same sequence as suggested above.

In this project, the purpose was to utilize a multimodal approach for automatically transcripting music and storing it on an online server. The approach was directed towards recognizing isolated notes obtained from electric and semi-acoustic guitars and preparing an acoustic model that would recognize new notes based on the ones it was trained with using Support Vector Machines. Images acquired for each subject were captured using a 5MP camera but only specific samples, where fretboard of the guitar was parallel to the ground were used for processing. This was done in order to prepare a flow that would automatically detect where the finger of the player was lying in order to both corroborate the result obtained from audio classification and decide the exact note being played based on the string and fret combination. Each string and fret combination of the guitar plays a unique note. However, after a certain fret, in each string, the notes start repeating. Because these notes have the exact same octave and timbre and thus the same features, determining where on the fretboard the finger lies will help deciding what the note exactly is. The language used for the purpose of feature extraction and classification of audio as well as image data was python, which being a server side scripting language, has the capability to run all the functions on the online server as well, as was the initial purpose of the project. This was done in order to ensure minimum processing on the host processor, thus removing any constraint of the processor to be used for recording the data.

Chapter 2

Data Acquisition

Before recognition of music notes can be initiated, a classifier with high degree of accuracy needs to be prepared. Appropriate database plays an instrumental role in deciding the robustness of the classifier and thus needs to be prepared meticulously. In case of music, database can either be prepared by extracting audio-visual information from pre-recorded, noiseless videos available online or it can be recorded live using standard accoudrements. However, a thorough research of online sources presents that complete data of both audio and visual formats is not readily available. Also, such a database would not ensure user-independence and would not be able to mimic real time recording, which as a result would lower the efficiency of the classifier. Consequently, a need for real time data acquisition presents itself. A.T.O.M. is a tool that can be used to transcribe music produced using any instrument. However, for the purpose of experimentation and testing, audio and image signals from a standard electric guitar have been acquired. The guitar has been chosen as the initial instrument because of its popularity in the music community and because it is the most widely known as well played musical instrument. Data from a total of **10** users was recorded using the equipment given below, connected as shown in **figure x**.

- A Fender Squire California Series Electric Guitar
- An RP155 DIGITech guitar modelling processor
- An iBall USB 2.0 8MP webcam

Data for the first 8 frets, along with the open fret were recorded for all 6 strings of the guitar. 10 samples of each note were acquired, making a total of 540 samples per user, which took 45 minutes to record. The data recorded was sequentially stored, sorted according to the user, the string number, the fret number as well as the sample number. Recording was performed in a noiseless environment, with the guitar processor acting as a standard filter.

2.1 Procedure

The Guitar's output was connected to the input pin of the RP155 Processor via a two ended 1/4" jack audio cable. The stereo output of the processor connects to the microphone input of the computer's sound card. The webcam is connected to the computer through a USB 2.0 port. Python libraries PyAudio and PyGame were used to write the data aquisition program. With each user it was observed that recording all 540 samples at one go was difficult as the users were required stay put to their positions during the entire recording. Based on their feedbacks, the record flow was divided into three as follows:

1. Frets 0-2 for all 6 strings
2. Frets 3-5 for all 6 strings
3. Frets 6-8 for all 6 strings

With the use of shutil module of python, recorded data was segregated into directories of the structure:

$|— < User >$
    $|— < User > Audio$
        $|— < User > \_s < stringno. > \_f < fretno. > \_ < sampleno. > .wav$
.
.
.
$|— < User > Image$
    $|— < User > \_s < stringno. > \_f < fretno. > \_ < sampleno. > .jpg$
.
.
.

Chapter 3

Audio Signal Processing

A sheet music consists of basic musical notes which represent a pitched sound. Any musical piece is composed of these notes and so, to achieve music transcription, recognition and conversion of these notes to their written symbols is the fundamental step. The first step towards this recognition is conversion of the audio into a format that can be recognized by a classifier i.e. extraction of the notable features of the audio [Barbancho et al., 2012].

Features can be defined as those components of an audio signal that are unique for classifying the phonological contents of a given speech sample. Among various existing feature extraction techniques for traditional speech recognition, Linear Prediction Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC),Perceptual Linear Prediction (PLP) and Linear Prediction Cepstral Coefficient (LPCC) are most frequently used. For the purpose of music recognition, the method MFCC is used as it represents the timber of a music piece, which is the quality of a musical note that distinguishes it from other notes or sounds of the same pitch and intensity. Part of what makes the timbre of a voice or instrument consistent over a wide range of frequencies is the presence of fixed frequency peaks, called formants. Thus, formants of the note are also calculated and used with MFCC to give the final set of features.

3.1 Mel-Frequency Cepstral Coefficients

The Mel-frequency cepstrum is a representation of the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up a Mel-Frequency Cepstrum. The difference between the cepstrum and the Mel-frequency cepstrum is that in MFC the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response.

MFCC is one of the best feature extraction techniques as it is less susceptible to noise and other variations. It follows the known deviation of a human ear's critical bandwidth with frequency. In this technique, two types of filters are employed. For the low frequency range, linear filters are used while for the higher frequency range, logarithmic filters are used. The speech sample is first split into various fragments, by framing and then, in order to reduce the discontinuities, windowing is done. After windowing, the samples are first converted into frequency domain using FFT (Fast Fourier transform) and then converted to Mel spectrum domain with the help of Mel-spaced filter banks. Mel scale relates frequency of the received signal of a pure tone to its actual measured frequency. It is a logarithmic scale which is linear below 1 KHz and can be stated mathematically as:

$$M(f) = 1125\ \log_e(1 + \frac{f}{100}) \tag{3.1}$$

The Mel spectrum coefcients are then transformed to time domain using Direct Cosine Transform (DCT). DCT de-correlates the features and arranges them in descending order of information they contain about speech signal. A flowchart of the process described has been given in Fig. 3.1

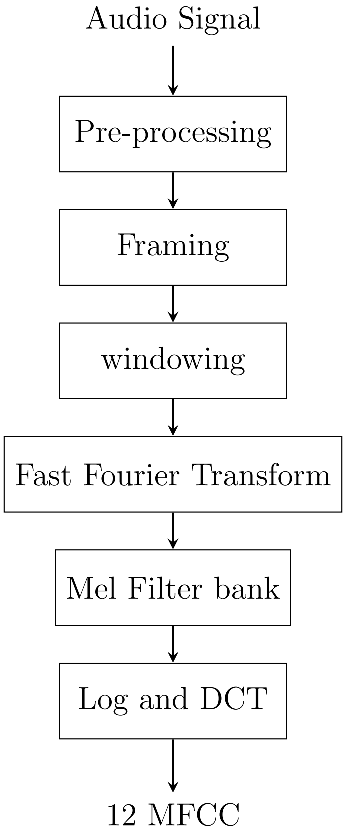


Figure 3.1: Mel-Frequency Cepstral Coefficient extraction

Pre-processing of the signal is done before feature extraction from the signal. The signal is first passed through a pre-emphasis filter to spectrally flatten the signal and make it less susceptible to finite precision effects later in the processing. Mean of the sampled signal is subtracted from the sampled signal itself to remove the DC offset present in the signal. After this, the signal is blocked into frames of 25ms or 400 samples. Once, the frames have been created, windowing of the signal is done using hamming window. An overlap of 160 samples is taken to ensure that the values at the beginning and end of a frame are not lost. No samples have been skipped in the process of windowing. Once the process is complete, Mel Frequency Cepstral technique is employed. In this method 24 features are extracted for each frame of the signal and quantized using the LBG algorithm for getting the features of an individual note. Of the obtained features, only the first 12 have been selected.

3.2 Formants

A formant, as used by James Jean, is a harmonic of a note augmented by resonance(cite wiki). The Acoustical Society of America defines a formant as "a range of frequencies [of a complex sound] in which there is an absolute or relative maximum in the sound spectrum". In music, it is defined as the range and number of partials present in a tone of a specific instrument, representing its timbre, where a partial is any one of the sine waves of which the complex tone is composed. When a guitar string is plucked, it vibrates and creates a rich spectrum of harmonic partials. The string itself is called the excitation source. The disturbed air molecules cause the guitar body to vibrate through sympathetic vibrations. The larger vibrating surface area creates higher amplitudes by causing more substantially more air to move. The guitar body does not proportionately amplify all of the frequencies of the string. Instead, some frequencies are amplified more than others. This quality is called resonance. Unless instruments are able to change their shape with each note, most exhibit a complex of many resonant frequencies that do not change, or are fixedthe specific complex of resonances are called formants and those that do not change are called fixed formants. Different notes have different fixed formants which occur at different positions in their spectrum and can thus be used as efficient features to distiguish between them.

To find out the formants of a note, the audio sample is first passed through a hamming window, given mathematically as:

$$w(n) = 0.54 - 0.46cos\left(2\pi\frac{n}{N}\right) \tag{3.2}$$

Here n lies between 0 and N, with N+1 being the window length L. This length is taken equal to the length of the sampled sound of the music note. An autoregressive model is then prepared using the output from the hamming window as its input. Order of the model can be given as:

$$o = (sampling\ frequency/1000) + 2 \tag{3.3}$$

Transfer function of the guitar string, which is the source of exitation is calculated using the autoregressive model, from which frequency response is then calculated. Peaks of this frequency response give us the formantswe require. The first three peaks are taken as features to be used in accompaniment with MFC coefficients for the purpose of classification of the isolated music notes.

Chapter 4

Support Vector Machine

Support Vector Machines are one of the latest methods that have been utilized for the purpose of classification of data [Wikipedia, 2015]. Derived from the statistical learning theory and inherently a binary classifier, the main purpose of an SVM is to minimize structural risk during classification of data points. Both linearly as well as non-linearly separable data can be classified using an SVM, the non-linear data being mapped to a higher dimension and kernel functions, instead of dot products being used as the similarity measure functions.
Support vector machines are supervised learning algorithms that perform data analysis and pattern recognition and are primarily used for the purpose of classification and regression. They emerged from the need to find a solution to the problem of over-fitting that was encountered in conventional neural networks. For a given set of p-dimensional training vectors belonging to two different classes, the SVM training algorithm prepares a model that uses a hyperplane to separate this data in such a way that the distance between the two classes, known as margin of separation, is maximized, as shown in Fig. 4.1, thus minimizing the generalization error. After preparation of the model is complete, new examples can be used to test the model by first being mapped to the same feature space as the training data and then being classified into one of the two categories depending on which side of the hyperplane they fall on. The hyperplane thus formed

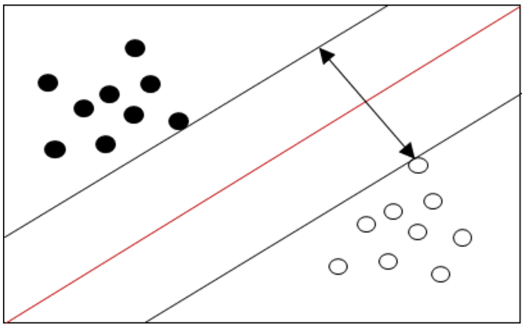


Figure 4.1: SVM for binary classification with red margin representing maximum margin hyperplane

is known as maximum-margin hyperplane and the classifier it defines is known as maximum-margin classifier. Besides binary classification, multi-class classification can be done either by one-vs-one technique or by one-vs-all technique. The library used in this study, LibSVM [Chang and Lin, 2011], uses one-vs-one classification technique, by default. Classification of data can be done linearly as well as non-linearly, as described below.

4.1 Linear SVM

For classification of linear SVM, let us take a training example $\{x_i, d_i\}$ and assume that classes represented by $d_i = +1$ and $d_i = -1$ are linearly separable. Here x represents the inputs and d represents the desired output. In this case, the decision hyperplane is given as:

$$\mathbf{w}^T\mathbf{x} + \mathbf{b} = 0 \tag{4.1}$$

where,
$\mathbf{w}^T$ gives the transpose of the weight vector
$\mathbf{b}$ gives the bias vector
$||\mathbf{w}||$ gives the Euclidean norm of $\mathbf{w}$
For linearly separable data, two hyperplanes are selected so as to separate the data such that no data points lie between them. Distance between these two planes is then maximized and the region bounded by them is called 'margin of separation'. These hyperplanes can be defined by the equations:

$$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} \geq +1 \qquad \text{for} \ \ d_i = +1 \tag{4.2}$$

$$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} \leq -1 \qquad \text{for} \ \ d_i = -1 \tag{4.3}$$

or together as:

$$d_i(\mathbf{w}^T\mathbf{x}_i + \mathbf{b}) \geq +1 \tag{4.4}$$

Here, the data points (x_i, d_i) for which the above equality holds are called support vectors. Classification of these vectors is the hardest as they lie closest to the hyperplane and hence influence the position of the hyperplane.
With these new hyperplanes, margin of separation is given as:

$$\text{margin of separation} = \frac{2}{||\mathbf{w}||} \tag{4.5}$$

Thus, the optimization problem is now given as maximization of margin of separation or minimization of $||\mathbf{w}||$. Introducing Lagrange multipliers α , this problem can be expressed in the dual form as:

$$\min_{w,b} \max_{\alpha \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{4.6}$$

Despite being of complex nature, this optimization is hardly helpful when one comes across data that is not linearly separable. Soft-margin SVM, which includes a slack variable that allows room for slight classification error, can be employed but only for cases that are still primarily linearly separable. Non-linear separation of data is done by employing the kernel trick.

4.2 Non-Linear SVM

Besides linear classification, non-linear classification for cases which cannot be spilt into two clean categories can also be performed by employing the kernel trick to map the data points to a plane where they are linearly separable [Aggarwal et al., 2015]. This trick can be explained as follows:
If a set of data points exists such that it can be classified in two classes, but not by a linear hyperplane, the common practice is to map this data to a higher dimension where a hyperplane can linearly separate this data. If the mapping function is given by $\phi(x)$, kernel $K(x, z)$ is given as:

$$K(x, z) = \phi(x)^T \phi(z) \tag{4.7}$$

Modification using the kernel trick allows the algorithm to fit the maximum margin hyperplane in a transformed feature space. The transformation may be non-linear and the transformed space high dimensional. Therefore, though the classifier is a hyperplane in the high dimensional feature space, it may be non-linear in the original input space.
Equation 4.6 is thus transformed into:

$$\min_{w,b} \max_{a \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{4.8}$$

In this way, instead of computing the dot product of data that has been transformed to a new higher dimension data, a kernel function can be used to find similarity between the training and testing data as given in the original feature space. The kernel function that has been used with the classifier in this project is the Radial basis Function or RBF kernel and can be defined as follows:

4.2.1 RBF Kernel

Radial Basis Function is one of the most widely used kernels in machine learning, primarily due to its infinite complexity. RBF kernel can be defined as:

$$exp(-gamma \ |\mathbf{u} - \mathbf{v}|^2) \tag{4.9}$$

where,
$\mathbf{u}$ = Feature vector of the hyperplane
$\mathbf{v}$ = Feature vector of the support vectors
$gamma$ = Constant parameter to be defined by the user

The free parameter gamma can take any real value. Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning far and high values meaning close. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

Chapter 5

Image processing

In a guitar, multiple notes are often repeated 2 to 3 times over the entire fretboard and audio processing is not sufficient to determine the exact position of the note on the fretboard. This problem can be solved by making use of Image Processing to determine the position of the finger on the fretboard and combine this knowledge with the labeled predicted by the classifier to determine the exact location of the note on the fretboard.

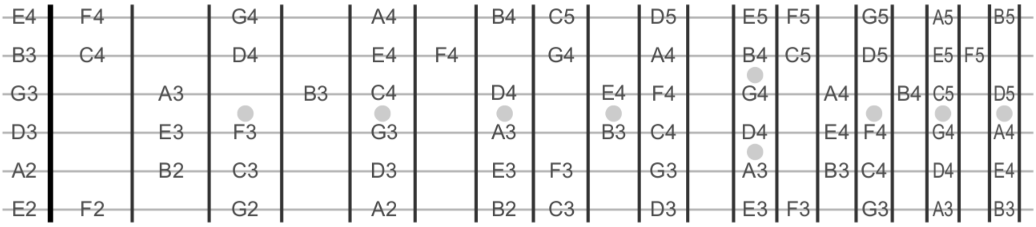

E4 F4 G4 A4 B4 C5 D5 E5 F5 G5 A5 B5
B3 C4 D4 E4 F4 G4 A4 B4 C5 D5 E5 F5
G3 A3 B3 C4 D4 E4 F4 G4 A4 B4 C5 D5
D3 E3 F3 G3 A3 B3 C4 D4 E4 F4 G4 A4
A2 B2 C3 D3 E3 F3 G3 A3 B3 C4 D4 E4
E2 F2 G2 A2 B2 C3 D3 E3 F3 G3 A3 B3

Figure 5.1: Fretboard of a guitar and notes corresponding to each fret and string

Figure 5.1 shows the image of a fretboard and the notes corresponding to each string and fret intersection. As can be seen, repetition of notes begins after the fourth fret, before which each note is unique and non-repetitive. Observing this pattern, the fretboard can be divided into two halves at the fourth fret, with the region ranging from fret one to four comprising of **Section 1** and the remaining frets i.e. those from four to eight determining **Section 2**. The problem can thus be reduced to simply detecting which section the finger of the player lies in and architecture of the guitar itself, which stays constant regardless of the type and company of the instrument, can be exploited to ease this task.

5.1 Procedure

The first step in determine the finger position of the guitarist is extracting the region of interest, which is the fretboard. However, because different guitarists have different styles of holding a guitar, preparing a generalized method that extracts the fretboard regardless of its orientation is challenging. Methods such as Canny Edge detection can be used for this purpose. However, with each changing image, various thresholds need to be tuned manually in order to obtain the required edges, which is both inefficient and time consuming. It is for this reason that the method proposed has currently been tested on ideal samples only, where the fretboard was parallel to the ground, as shown in figure 5.2.

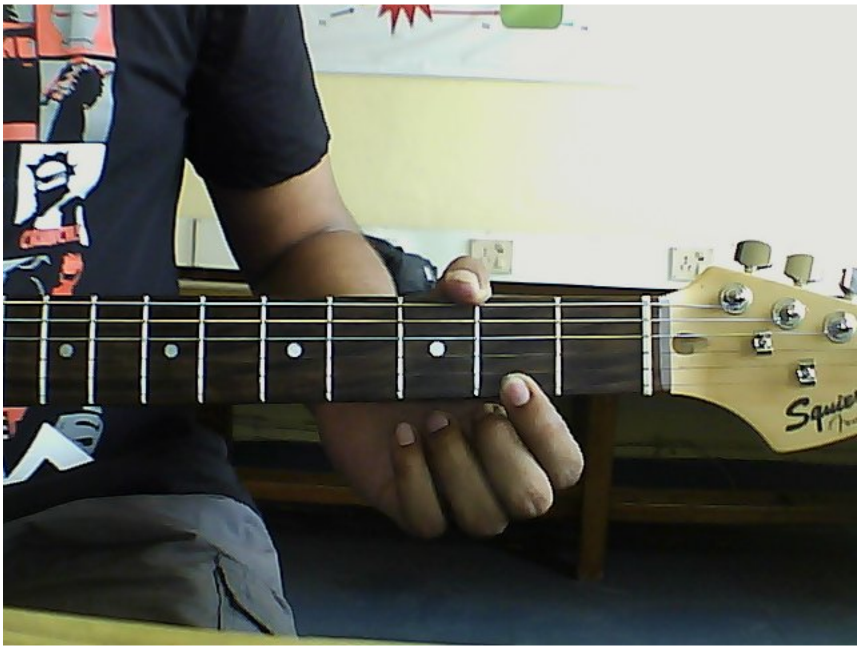


Figure 5.2: Example of the image used

Once the fretboard has been extracted, all its connected components are given separate labels. Connected components in an image refer to sections of the image which have almost same pixel intensity values and are connected in some way. Labeling connected components allows one to extract components of a specific property. This property can vary from area to intensity to orientation. At this point, it needs to be decided which characteristic of the guitar is to be used for determining finger position. Like number of strings, spacing between frets and presence of nut, position markers are another constant in the design of a standard guitar. A position marker, which is the tiny white dot in the middle of the fretboard, occurs at fret 3,5,7,9,12,15,17,19 and 21 and is used to number frets without actually counting them. Knowing the position of the finger relative to these markers is an effecive way of determining the section in which the finger lies. The labeled fretboard image can be filtered to extract only these circles and the hand. Eccentricity of an ellipse, defined as the ration of the distance between its foci to the major axis length, can be used to clearly distinguish the circular markers and hand pattern from the straight lines corresponding to frets and strings. Eccentricty values close to zero denote circular regions while those close to 1 denote straight lines. Once the position markers and finger pattern have been extracted from the original image, their centroids or centers of mass can be used as indicators of their spatial location and distances between the finger and the different position markers can be employed to decide the region in which the finger lies.

5.2 MATLAB Implementation

Images similar to figure 5.2 were used for the present analysis as they provide ideal positioning of the fretboard. The first step was cropping the image to obtain only the area containing the fretboard. This was done using the MATLAB function **imcrop()** and the result obtained was as shown in figure 5.3. This image was then converted to binary and different connected components were given different labels using **bwlabel()**. The result for graylevel labeling is shown in figure 5.4 where each pixel is colored with a different degree of gray depending on the component it was assigned to. Eccentricity of each connected component was found using **regionprops()** and components with eccentricity less than 0.9 were removed from the image. These components cover a majority of the vertical and horizontal lines. To ensure no stray dots or lines were left, erosion was applied on the image using a disk structural element of radius 3. An example of the resultant image can be seen in figure 5.5. Centroids of these elements, as represented by the green asterisk, were used to provide their coordinates and distances between them were used to determine the region in which the finger was present.

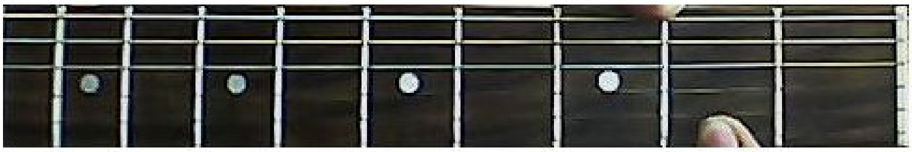


Figure 5.3: Cropped image extracted from captured image

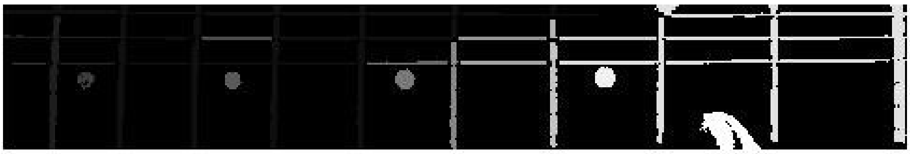


Figure 5.4: Resultant of graylevel labeling

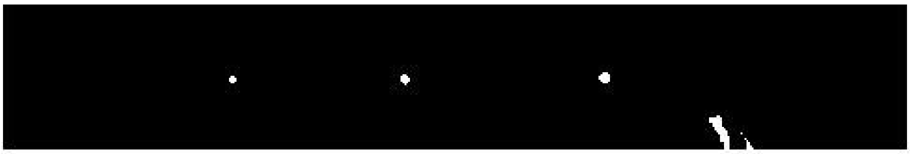


Figure 5.5: Resultant of eroded image with centroids

Images taken for the above analysis were such that only the first 8-10 frets of the guitar were visible. As we move away from the nut, size of the position marker decreases. Consequently, erosion removes the smaller markers, leaving only the prominent markers. This limits the number of position markers in the image to 3. To determine the region of finger position, the following reasoning was used:

- Markers were distinguished from the finger using their area, which will always be smaller than the finger.

- Once position of markers was known, distance of finger from the first marker was calculated by subtracting x coordinate of finger from that of the marker.

- For a negative value, the finger was concluded to lie in section 1. If the value was positive, distance from second marker was calculated in the same manner.

- For a negative value in the second case, the finger was concluded to lie in section 1. For a positive value, it was concluded to lie in section 2.

- If only three centroids were present in the image, it was concluded that the finger was lying on one of the three markers.

- If the finger was lying on marker 1, it was concluded to be in Section 1 of the fretboard. It was concluded to be in section 2 in any other case.

Once it has been determined which section of the fretboard the player is playing in, the information can be easily combined with the predicted labels from audio classification to provide the exact position of the player's finger. This information can then be saved in the form of tablature, which will be discussed later. However, due to the following constraints, this method could not be automated and could thus not be put to practical use.

1. Different users hold the instrument in different ways which makes the process of extracting the fretboard arduous and tricky.

2. The maximum eccentricity which needs to be allowed for the finger to be detected correctly while simultaneously avoiding all stright lines changes for every image.

3. Radius of the disk structural element needs to be set separately for each image in order to ensure no extra spots remain on the final image.

Chapter 6

Music Transcription

Music notations, when represented in handwritten or printed formats, making use of modern musical symbols, form what is called sheet music. A typical sheet music is as shown in figure 6.1 and the process of its generation i.e. converting music into its corresponding written format is known as music transcription. Musical notations can either be given as sheet music using musical symbols or represented as tablature, which consists of fingering information i.e. the string and fret played at a particular instance. Once we know the note being played, as predicted by the audio classifier, this information can be used to prepare the sheet music, which can then be easily stored on the online server and downloaded from any part of the globe.

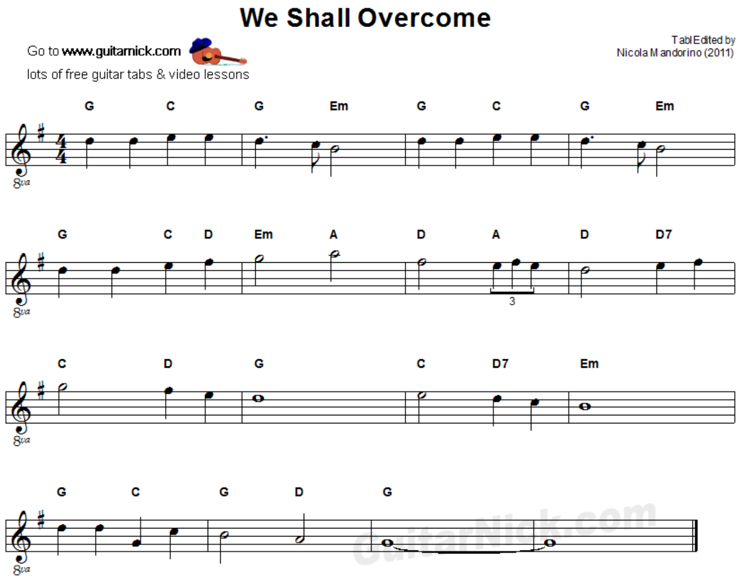
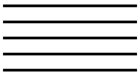
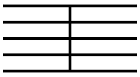


Figure 6.1: Sheet Music for Ludwig Beethoven's Moolight Sonata

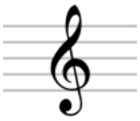
Knowledge of music symbology is critical in understanding and creating sheet music. Some of the most commonly used symbols, along with their meanings, have been shown here.



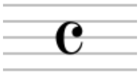
The Staff is the fundamental lattice over which music symbols are placed.



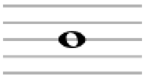
Music is cut off in uniform bars or measures and the bar line is used to separate these.



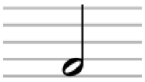
Clefs define the pitch range of the staff where it is placed and G-Clef or Treble Clef, with the center of its spiral on the last but one line assigns G above middle C or 392 Hz to it.



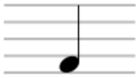
Time Signatures suggest the grouping of beats or pulses and commontime denotes a 4/4 signature.



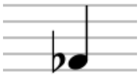
A wholenote or Semibreve is a note with length equal to 4 beats in 4/4 time.



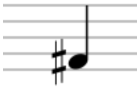
A halfnote or Minim is a note with length half that of a wholenote or equal to 2 beats in a 4/4 time.



A quarternote or Crotceht is a note with length one quarter that of a wholenote, or equal to 1 beat in 4/4 time.



A flat note reduces the pitch of a note by one semitone or half tone.



A sharp note raises the pitch of a note by one semitone or half tone.

Various software are available to achieve transcription of music both online and offline. Since the proposed system wishes to achieve transcription online, a python + LaTeX cross-platform library LilyPond was used to convert music notes to their corresponding written symbols. Python avails us with mingus, an advanced, cross-platform music theory and notation package which can be used to create sheet music with LilyPond. Other than providing functions such as comparison of notes and converting notes to MIDI events, migus allows one to group notes to form Bars, Tracks, Compositions and Suites and output png's and pdf's of the converted sheet music using LilyPond. Notation of music changes with changes in octave as well as increase and decrease in semitones. LilyPond allows inputs of notes with specific octaves and transcribes sharp as well as flat notes. Its artistic font is also approved of by musicians both seasoned and amateur. An example of the sheet music created using mingus and LilyPond has been shown in figure 6.2.

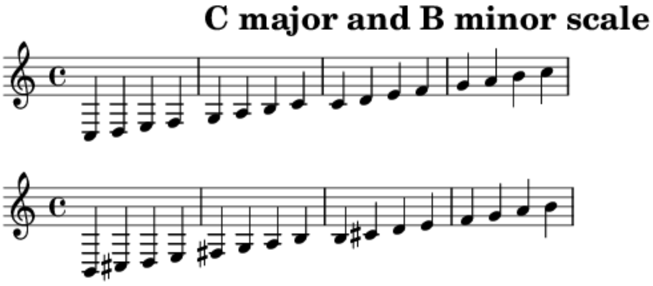


Figure 6.2: Sheet music containing notes lying on the C major and B minor scale

Chapter 7

Web Development

The need to provide a secure access to the developed sheet music over the internet requires a basic understanding of the concepts of web development and a detailed insight into what servers are, how they work and how they are created.

Web development, in its essence, refers to any work involved in developing a website or even a single web page on the internet and web server configuration is but one part of it. A web server is similar to a personal computer and stores data in the same manner as any other computer. However, the files within the server, unlike the conventional png and pdf files, are responsible for the working of a website. A server can be understood as a machine that waits for a request from another machine and upon reception of a request, send back relevant replies. All that is required to be able to host a website from a remote sever is a dedicated PC, the appropriate operating system and 24/7 internet connection. Other than these hardware requirements, various software are required for enabling the computer to work like a server of which, a web server software and a web application framework are the primary requirements.

7.1 SSH: Secure Shell

Secure Shell, or SSH, is an encrypted network protocol to allow remote login and other network services to operate securely over an unsecured network.

SSH provides a secure channel over an unsecured network in a client-server architecture, connecting an SSH client application with an SSH server. Common applications include remote command-line login and remote command execution, but any network service can be secured with SSH.

SSH uses public-key cryptography to authenticate the remote computer and allow it to authenticate the user, if necessary. There are several ways to use SSH; one is to use automatically generated public-private key pairs to simply encrypt a network connection, and then use password authentication to login.

Another is to use a manually generated public-private key pair to perform the authentication, allowing users or programs to log in without having to specify a password. In this scenario, anyone can produce a matching pair of different keys (public and private). The public key is placed on all computers that must allow access to the owner of the matching private key (the owner keeps the private key secret). While authentication is based on the private key, the key itself is never transferred through the network during authentication. SSH only verifies whether the same person offering the public key also owns the matching private key. In all versions of SSH it is important to verify unknown public keys, i.e. associate the public keys with identities, before accepting them as valid. Accepting an attacker's public key without validation will authorize an unauthorized attacker as a valid user.

7.2 Apache

One of the most popularly used web server software is the Apache HTTP server, commonly known as simply *Apache*. A web server or an HTTP server, in general, is a program that provides access to a requested website using the Hyper Text Transfer Protocol or HTTP. The features of Apache can range from server-side programming language support to authentication schemes. Programming languages commonly used are Perl, Python, Tcl and PHP while authentication schemes generally utilized are mod_access, mod_auth, mod_digest, and mod_auth_digest.Apache features configurable error messages, DBMS-based authentication databases, and content negotiation. It is also supported by several graphical user interfaces (GUIs). It supports password authentication and digital certificate authentication. Because the source code is freely available, anyone can adapt the server for specific needs, and there is a large public library of Apache add-ons. Apache features only need to be set once and once fixed, the computer can be used to receive data from other computers.

7.3 Django

A Web Application Framework (WAF) is a software framework that is designed to support dynamic website development, along with creation of web services, applications and resources. Django [Foundation, 2015] is a free and open source web application framework written in python and serves the purpose of easing the creation of complex database-driven websites. A database is nothing but a collection of information that can be manages and updated easily. With respect to servers, a database is formed when all the data received by the server is stored in the form of tables within it.

Communication between a web server and a website takes place in the form of requests and responses. The website sends a request to a server every time an activity is to take place on it. Everything from typing the URL to pressing play on an online video to submitting a form comes under a request. Django contains written protocols which direct the server on how the data and the requests are to be handled. In crude terms, a website sends a request to the web server, the web server looks for Django which then selects the appropriate file which can act as a response to the request sent by the user.

7.4 Procedure

Once a server instance is created, it is accessed using SSH protocol and the required frameworks are install and created for data in-flow and out-flow.

When a URL or Uniform Resource Locator is typed in, the browser looks for a DNS record, which points to the web host where the server is located. This server contains files that constitute the website that the user has requested. The server points the data to Apache, which only needs to be set up once and the data is then pointed to the Django project. Simply speaking, the user requests a website, this request is processed by Django or another similar WAF and a response for the request is appropriately sent.

Chapter 8

Results, Conclusions and Future Scope

The ken and scope of this project is immense. The integration of ATOM with the internet, makes it an IOT based innovation which can be used by musicians all around the world to obtain instant transcription of the music they have played. The amalgamation of audio and image components in ATOM will allow the user to listen as well as see the details of the music (chords, notes) that he has played, and find out where the played music was not good enough.

As of now, ATOM takes care of musical transcription for a guitar player. However, the same concept can be applied to other musical instruments to obtain instantaneous, easily accessible transcription of the music played. Integration of ATOM in the future vision of smart home/smart city is very easy and it can indeed help in forging a new pathway to the future.

List Of Publications

Others

Bibliography

[Aggarwal et al., 2015] Aggarwal, A., Sahay, T., Bansal, A., and Chandra, M. (2015). Grid Search Analysis of nu-SVC for Text-Dependent Speaker-Identification. In *2015 Annual IEEE India Conference (INDICON) (IEEE INDICON 2015).*

[Akbari and Cheng, 2015] Akbari, M. and Cheng, H. (2015). Real-time piano music transcription based on computer vision. *Multimedia, IEEE Transactions on*, 17(12):2113–2121.

[Barbancho et al., 2012] Barbancho, A., Klapuri, A., Tardon, L., and Barbancho, I. (2012). Automatic transcription of guitar chords and fingering from audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):915–921.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

[Foundation, 2015] Foundation, D. S. (2015). Django (Version 1.9) [Computer Software].

[Mauch and Dixon, 2010] Mauch, M. and Dixon, S. (2010). Simultaneous estimation of chords and musical context from audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1280–1289.

[Qamra et al., 2005] Qamra, A., Meng, Y., and Chang, E. (2005). Enhanced perceptual distance functions and indexing for image replica recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):379–391.

[Singh et al., 2015] Singh, B., Rajiv, P., and Chandra, M. (2015). Lie detection using image processing. In *Advanced Computing and Communication Systems, 2015 International Conference on*, pages 1–5.

[Wikipedia, 2015] Wikipedia (2015). Support vector machine - Wikipedia, The Free Encyclopedia.