Naad-Nartan
Journal of Dance & music
UGC-CARE Listed ISSN: 2349-4654
Volume : 12, Issue-1, January 2024

Automatic notation generation using speech recognition

Anuvratha Narasimhan
Research Scholar, Sri Padmavati Mahila Visvavidyalayam, Tirupati

Prof. RNS Saileswari
Department of Music, Dance and Fine Arts
Sri Padmavati Mahila Visvavidyalayam,Tirupati

ABSTRACT

Notations play a pivotal role in archiving compositional forms. In carnatic music, notations are not only means to propagate music to masses and pass it down generations, but also help students of music to improve their musicianship. Reading and writing notations is an art. Transcribing a piece of melody to sol-fa (swaras in carnatic music) requires years of training and a keen ear. Apart from the musical knowledge, transcription also requires one's laborious effort on writing/typing, legibility, neat alignment etc. Thus, notation writing has 2 phases, the musical phase and the transcription phase. This paper aims at making the transcription phase almost effortless. Carnatic music has 7 sol-fa syllables used to indicate the Saptaswaras namely, Sa(Shadja), Ri(Rishabha), Ga(Gandhara), Ma(Madhyama), Pa(Panchama), Dha(Dhaivata), Ni(Nishada). Using speech recognition algorithms, this software helps convert a piece of melody, sung in sol-fa form (Swara), to text. The final output is in the traditional notation format generally used in carnatic music.

Keywords :

Research Paper

Introduction

The solfege system is a brainchild of Indian music. Assignment of syllables for different notes makes practice of music much easier. Melodies, when translated to their sol-fa counterparts, are called notations. The notation system forms the backbone of Carnatic music. Carnatic music has primarily been an oral and aural tradition. Nevertheless, preserving melodies/songs/ concepts of music in the form of notations has been given equal importance. From palm leaf manuscripts, carvings on copper plates, inscriptions on temple walls to printed papers, notations are found in every possible form. The advent of print technology was a game changer. Hundreds of books containing notations of songs have been published over the years so much so that there was a dedicated column for musical notations of carnatic songs in an entertainment magazine in Tamil Nadu.

As recording and audio/video technologies flourished, a slight decline in publication of notations has been observed. In spite of these advancements, notations remain the primary sources for learning and storing compositions for students of Carnatic music.

This paper aims at taking notation writing to the next step using the latest technologies like machine learning and speech recognition. It leverages on the fact that any piece music can be translated into solfege (the sapta svaras). A system can be trained to recognize these sol-fa syllables and transcribe any piece of music to notation.

Research Methodology

This research employs experimental and exploratory research methodologies. Data analysis has also been used.

Transcription of audio using speech recognition Technology

A. Basics of Sound

Sound is produced by vibration of materials. The sound we hear is a result of vibrating air particles. It is represented in the form of a wave. Vibrating particles create pressure differences resulting in compressions and rarefactions. Few properties of a sound wave are explained below:

1. **Amplitude:** It is measure of loudness of sound. It is represented by the height of the wave.

2. **Frequency:** It is the total number of waves produced per second. It is measured in Hertz(Hz). A human ear is capable of hearing sound in the frequency range of 20Hz-20KHz.
3. **Pitch:** It is a characteristic of sound that depends on frequency and many other factors. A higher frequency indicates a higher pitch. Simply put, pitch is the factor that enables us to differentiate between various sounds. For example, pitch is the component of sound used to differentiate between a male and a female voice.
4. **Sample Rate:** It represents the number of samples generated per second. Higher the sample rate, a better digital representation of sound wave is achieved.

B. Mel Spectrogram

Fig.1

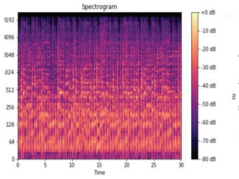


Fig.2

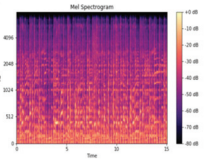


Fig.3

A sound waveform is represented as an Amplitude Vs Time graph (Fig.1). A spectrogram (Fig.2) is a better visual representation of a sound wave. It represents the spectrum of frequencies of the sound wave as it varies with time. A waveform in time domain can be converted to a spectrogram (frequency domain) by applying Fourier Transform techniques. The frequencies in a spectrogram are linear in nature.

A Mel spectrogram (Fig.3) is a visual representation of sound in frequency domain where the frequency is in logarithmic scale. Humans can hear sounds from 20Hz to 20KHz.(Joshi) However, the distinction between 100Hz and 200Hz is much more evident than the distinction between 1000Hz and 1100Hz (a same difference of 100Hz in both cases). Hence, a logarithmic frequency scale (Mel scale) mimics the human perception better. Mel Spectrograms are thus used extensively to train machine learning systems for image recognition.

C. The art and science of writing notation

Notation writing is a benchmark for testing the musical capabilities of a carnatic musician. The skill of transcribing melody to solfege is often referred as 'Svara Gnyana', roughly translating to 'Knowledge of notes'.

Notation writing has 2 phases:

1. **Musical phase:** In this phase a musician translates a song/melody into sol-fa syllables in mind or orally.
2. **Transcription phase:** In this phase, the translated song/melody is written/typed with required details such as sthayi (octave) representations, Tala (rythmic cycle) demarcations etc.

The musical phase is an intellectual process and requires years of traning and practice for an artist to master. Whereas, the transcription phase can painstakingly laborious and time consuming. A machine learning system for automatic notation generation on dictation can simplify this process.

D. Role of Machine Learning and speech recognition

Machine learning is a branch of computer science based on artificial intelligence which depends on large amounts and data and algorithms to mimic the learning capabilities of humans. Machine learning is one of the workable approaches for problems with no definitive logical solutions.(Ravikoti) Speech recognition, also known as Automatic Speech Recognition(ASR) or Speech to Text(STT) uses machine learning to transcribe speech to text format. It is a state of art technology. The system must be trained with billions of data over several years to master the language and give results with high accuracy.

There can be many approaches for designing a system for automatic notation generation. Two of them being- Pitch based sound recognition and Speech recognition. Pitch based sound recognition and music transcription is a state of art technology. Though it proves to be effective for transcription of instrumental music, it is yet to achieve similar results with vocal music.

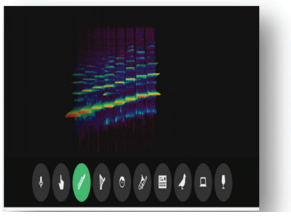


Fig.4

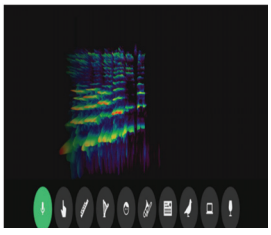


Fig.5

Fig.4 and Fig.5 are spectrograms of the seven notes in the same pitch played on flute and sung by a human respectively. The distortions in the spectrogram of human voice are very evident. Details about Mel Spectrogram are discussed later in the paper.

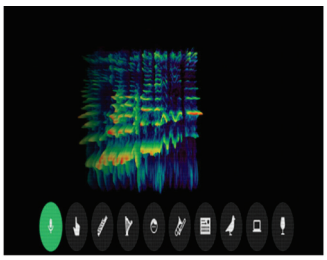
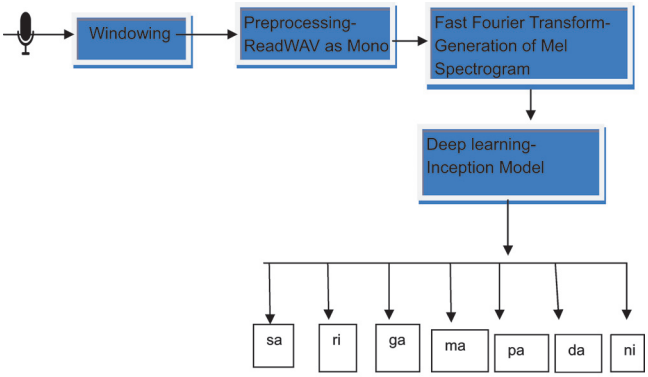


Fig.6

Fig.6 is the spectrogram of the same 7 seven notes sung in the same pitch but with Gamakas (oscillations). (Chandrasekharan) Given the numerous varieties of oscillation of notes giving rise to different varieties of each note, establishing a pitch based system for sound recognition for carnatic music would be a herculean task.

Hence, speech recognition could be a more appropriate approach for transcription of a piece of music sung in sol-fa form to text. A deep learning model developed by Google has been trained to identify the seven sol-fa notes when uttered.

E. Block Diagram of the proposed solution:



The microphone is used as the input device. A continuous piece of melody sung in the sol-fa form is received by the microphone as input. The input is in the stereo format. The waveform (time-domain) of the input is then windowed to obtain the audio waveform for every beat. Number of beats per minute is a user input variable. In the preprocessing step, the audio in stereo format (WAV) is converted to Mono. The waveform is then converted to Mel Spectrogram using Fast Fourier Transform techniques. A deep learning system is a multi layered neural network trained on a huge amount of data. This technique is used here for speech recognition using the Inception Model engineered by Google for image classification(Tiwari). The Inception model used in this research paper is trained to recognize the Mel spectrogram of the seven Sol-fa syllables namely Sa, Ri, Ga, Ma, Pa, Dha and Ni and make appropriate prediction to generate corresponding notation for the melodic input.

F. Software Implementation

The software is developed using C# programming language. It comprises of two modules:

1. Training Module

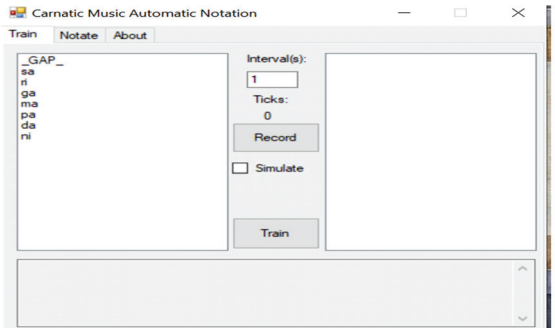


Fig.7

The training module consists of a user interface input source (Fig.4) to train the software to recognize sol-fa notes. The user can record each note multiple times and train the inception model. Once the required data is recorded and collected, on clicking the 'Train' button, the solftware converts each input audio wave to corresponding Mel Spectrogram image as shown in Fig.5. An option to simulate data is enabled. The software generates additional simulated data by changing the pitch and shifting the waveform of each user input.

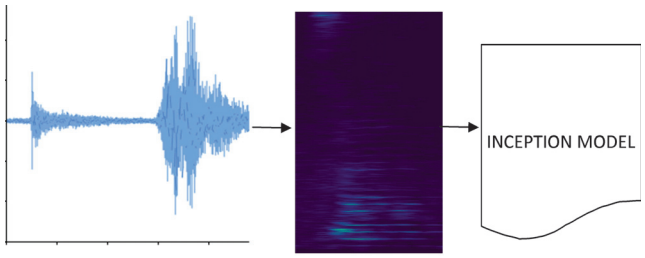


Fig.8

Prediction Module

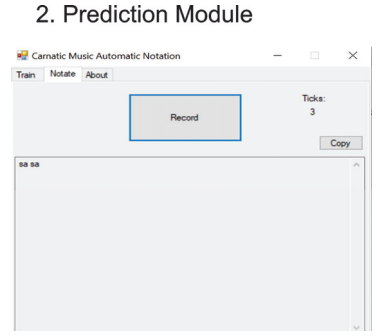


Fig.9

The prediction module consists of a user interface for the user to record the piece of melody intended to be notated. Once the user stops recording, the software uses the inception model generated using the training module and makes predictions of sol-fa notes for each beat.

In addition to the Tensorflow Inception Model the following C# libraries were used:

- ML.Net- Machine learning library
- Spectrogram.Net- For generating spectrograms from audio waveforms.
- NAudio- Audio processing library

G. Testing & Results:

Machine learning works on huge data sets. For instance, a speech recognition system for linguistic purposes has dedicated data sets like Kaggle, MNIST etc that are available to everyone. Application of machine learning techniques to carnatic music is a nascent development. There are limited data sets readily available for use. As a temporary solution to the dearth of data sets, this system uses simulated data for training and prediction. When a note is given as input for training, the software is capable of generating multiple inputs by changing the pitch and introducing time offset. Black box and white box testing methodologies have been used for testing the software.

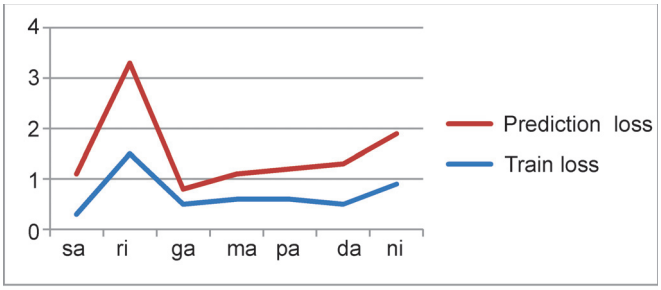


Fig.10

For training purposes, the software divides the data into 2 parts. One part used to train and the other is tested using trained model. For prediction, known melodies sung in sol-fa form were given as input and response of the software was noted and compared with the expected response. Fig.10 shows the loss function for the training and prediction model for each sol-fa syllable. The accuracy of this trained model has been observed to be 68% approximately.

This software and related code has been shared under Apache 2.0 license at https://github.com/anuvratha/CarnaticMusicAutomaticNotation/.

Libraries

1. https://musiclab.chromeexperiments.com/Spectrogram/
2. https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet
3. https://github.com/swharden/Spectrogram
4. https://github.com/naudio/NAudio

References

1. Tiwari, Rajpriya. "How to use Inception Model for Image Recognition."www.indusmic.com, 5th August 2021, https://www.indusmic.com/post/how-to-use-inception-model-for-image-recognition.
2. Ravikoti, Sridhar, "Identifying Ragas carnatic music machine learning." www.linkedin.com, 21st September 2020, https://www.linkedin.com/pulse/identifying-ragas-carnatic-music-machine-learning-sridhar-ravikoti/.
3. Athreya, Srinivas, et al. "Deep learning based Raga classification in Carnatic Music." www.medium.com, 16th July 2020, https://medium.com/@blogsupport/deep-learning-based-raga-classification-in-carnatic-music-e499018ea1b7
4. Chandrasekaran, J et al, "Spectral Analysis of Indian musical notes" Indian Journal of Traditional Knowledge, Vol.4, Issue 2, April 2005, pp. 127-131
5. Joshi, Dipti et al. "Comparative Study of Mfcc and Mel Spectrogram for Raga Classification Using CNN" Indian Journal of Science and Technology, Vol 16, Issue 11, March 2023, p: 816-822. https:// doi.org/10.17485/IJST/v16i11.1809