# Wavelet processing of Indian Classical Music Ragas

Ankur Sharma
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
ankursharma@ucsd.edu

## Abstract

*In this project, I have aimed to analyze the efficiency of various discrete wavelet transforms for classification of Indian Classical Music Ragas. This project fits in the broader scheme of understanding how human beings produce and understand speech and whether wavelet based analysis is better for audio content analysis compared to STFT. My analysis has shown that wavelets outperform STFT by a huge margin for Hindustani Raga classification. I used a bidirectional LSTM model coupled with CNN to generate the classifier. The model performed 96%-97% for wavelets whereas only 38% for STFT based classification. This analysis was done with the same model and only the input layer was changed.*

## 1. Introduction

Audio content analysis has been at the forefront of music information retrieval. Devising efficient ways to catalog and analyze audio has been of great interest for commercial reasons and the need for complex digital signal processing required in analyzing the music has piqued the interest of the scientific community in order to simplify the process. Efficient music categorization is an ongoing endeavor. The music being analyzed varies vastly and with that, the algorithms required to process them efficiently. As one moves outside the realm of western music, into music of other cultures, a lot of the techniques need modifications as per the grammar and conventions of that music.

One such genre of "non-conventional" music is the Indian Classical Music; specifically the Indian Ragas. Indian ragas are a complex set of melodies and art pieces that have a defined set of rules, but a huge margin of improvisation during the performance. So long as the set of rules defining the raga are followed, it is not classified as a different raga, despite an incredible amount of variations in the performance. To add to the complexity, various ragas have similar tonal structures and are often very difficult to differentiate. It is reasonable to say that the Indian ragas cannot be analyzed based on the algorithms and features developed for western music.

This leads researchers to develop algorithms that can mimic the trained ear of an Indian classical musician. Hence, a very good point to begin the analysis and identification of ragas is to figure out the Mel Frequency Cepstral Coefficients and develop on them. These coefficients fall short when we take into account the transitions between notes within the raga. So, a need for a better and more robust feature set arises, perhaps beginning from the emulation of the human hearing system, i.e. the cochlea. Once the emulation of cochlea and logarithmic representation of frequency is done, we move on to develop some more complex audio features which can give us the contour of the raga and help develop that into a feature vector.

### 1.1. A note on Indian Classical Music

Indian music is a orally taught tradition. It's origins date back to sacred Vedic scriptures over 6,000 years ago where chants developed a system of musical notes and rhythmic cycles.

In this way, Indian classical music is very closely connected to nature, taking inspiration from natural phenomena including the seasons and times of the day to create 'ragas' or musical moods and many time cycles or 'taals' that have been further codified. the melodic framework of ragas is based on seven swaras (notes) or saptak.

The seven swaras in an octave are Sadja(Sa), Risabha(Re), Madhyama (Ma), Panchama(Pa), Dhaivata(Dha) and Nisada/Nishad (Ni). Each raga is made up of five or more notes woven with patterns and attributes which lay the foundation for compositions and improvisation (Manuel, P. (2001)). These attributes include the thaat (sequence of notes), the aaroh, the avaroh, the chalan or pakad (specific clusters of notes), the vadi (the note having the highest significance) and samvadi (the note having the second highest significance) swaras, the jati, the gayan samay (time of day) etc., which give the raga its own unique flavor.

Compositions are fixed but most of the music is improvised within the structure of notes and mathematics. This

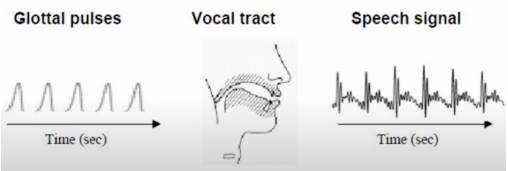


Figure 1. Speech as a convolution of vocal frequency response and glottal pulses

gives the music a spontaneous freedom where each artist and every performance is ensured to be completely unique. Indian classical music is generally passed down in an oral tradition where the student would spend many years with their 'guru', developing a very special, spiritual bond, imbibing all aspects of the music along with philosophical and moral principles that shape them for life.

Now Indian classical music can be learned in many institutions and has been heavily documented and notated but learning through observation, listening and memory is still paramount and connecting with an expert teacher is considered the most fruitful way to learn.

### 1.2. Indian Ragas and Time of Day

One of the first documented attempts at classification of ragas is by bhatkhande [9]. He tried to recognise patterns in various ragas and proposed a theory that would lay down the "Time of day" principles; these principles spoke about certain ragas to be sung at a respective time of day. His thesis pertained to the idea that each raga exhibits emotions that get highlighted when performed at certain times of the day. The theory, still holds respect among Classical Hindustani music singers and they choose a certain raga to perform depending on the time of the concert.

### 1.3. Speech generation

Fig 1 illustrates the generation of speech in humans. The glottal pulses convolved with the impulse response of vocal tracts produces speech that we hear from each other. The vocal tract as a filter on the glottal pulses and adds semitones that gives us all a unique timbre. A similar phenomenon occurs in the voice box of a guitar. The audio from the strings is amplified by the wooden box. This amplification comes with slight modification and that makes each guitar unique.

### 1.4. Speech perception

Now that we discussed speech production, we need to discuss speech perception in humans. Human ears act like a constant Q filter banks for several frequency bands, separated logarithmically. The hair cells provide slight modification to the quality factor, but that can be ignored for this discussion.

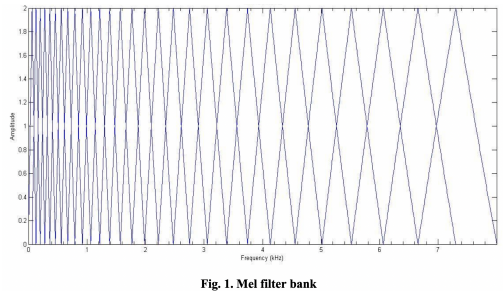


Figure 2. Mel scale filter bank

The log spaced frequency perception is close to pitch perception. Our ears are designed to recognise pitches in sounds. This understanding led to several psychological experiments that led to different models of human audio perception and a filter bank that would emulate it closely. There are two very popular filterbanks that have been developed over the years: The Mel Filter Bank and Bark Filter Bank.

### 1.5. Mel Scale

Mel scale is an experimentally derived scale that shows the perception of audio by humans. Mathematically, a frequency on mel will be proportional to the log of the linear scale frequency. The equation can be quantified as below:

$$f_{mel} = 2595 * log_{10}\left(1 + \frac{f_{linear}}{700}\right)$$

Fig 2 also shows the mel filter bank.

### 1.6. Bark Scale

Bark scale is derived on similar concepts as Mel Scale. This is also an experimental scale and has multiple representations:

$$f_{bark} = 13{*}tan^{-1}\left(0.76 * \frac{f_{Hz}}{1000}\right) + 3.5{*}tan^{-1}\left(\left(\frac{f_{Hz}}{7500}\right)^2\right)$$

or (Traunmüller, 1990):

$$f_{bark} = \left(\frac{26.81 * f_{Hz}}{1960 + f_{Hz}}\right) - 0.53$$

or (Wang, Sekey & Gersho, 1992):

$$f_{bark} = 6 * sinh^{-1}\left(\frac{f_{Hz}}{600}\right)$$

The bark filter bank is illustrated in Fig 3

### 1.7. Mel Frequency Cepstral Coefficients

The MFCCs are a set of features that collectively make up the Mel Frequency Cepstrum. These features are used to describe the timbre of an audio signal. MFCCs are extracted using the following steps:

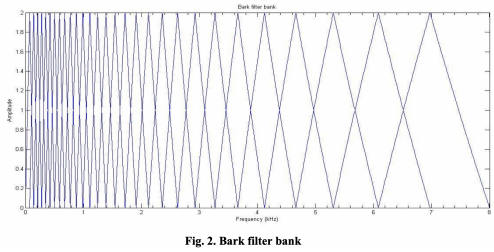


**Fig. 2. Bark filter bank**

Figure 3. Bark scale filter bank

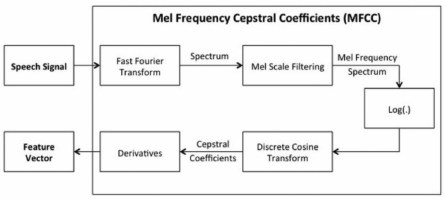


Figure 4. Steps involved in extracting MFCCs

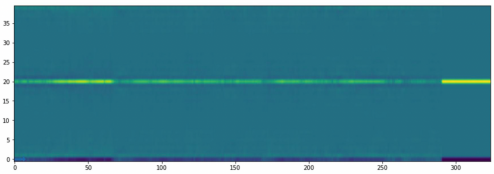

Figure 5. MFCCs of spoken "1" as a spectrum

1. Get STFT of a signal
2. Map the frequencies to the Mel scale
3. Take log of powers at each Mel frequency
4. Get the discrete cosine transform of the above powers (hence the name cepstrum)
5. MFCCs are the amplitudes of the resulting spectrum

Fig 4 shows the steps involved in extracting MFCCs. Fig 5 depicts the MFCCs of an audio signal from FSD datasets.

### 1.8. Advantage of Filter Bank based Approach

Since Indian Classical music, specifically, the Indian ragas, are extensively dependent on the contextual proceedings of the notes and transitions used, it is imperative to realize their classification using a wavelet transform rather than an STFT based approach. This approach has not been used with Indian Classical music before and it would be interesting to see the result of this assessment. This doesn't conflict with Lerch's idea of the high complexity of assessment since we can afford to do so for a small portion of audio signals.

## 2. Previous Works

Papers [7] [6] discuss the applications of DWT in the extraction and classification of musical features and contrast them with Mel coefficient-based classification. A similar paper [3] discusses how to use wavelet transform to derive coefficients that are close to log based frequency distribution and use them to identify a speaker. This is an interesting application of DWT since it aims to determine a speaker and use that methodology for biometric identification, similar to how we identify individuals over a phone call. There has also been some research on this subject by Richard Kronland-Martinet [11] [10], who has used DWT to analyze audio in great detail. He has used this process in the analysis of diverse signals, however the ones that stand out the most are in speech recognition and music analysis. Papers [4] [7] also talk about how to use audio wavelet transform and processing to figure out the phonetics and vowels in speech. The interesting thing about all the papers that have been cited here, and more, have been invested in analyzing audio based on log scale representation of frequencies which is how humans perceive pitch.

If we move to the domain of audio processing and commercial applications, people are still using the STFT based approach. As is clear in Prof. Alexander Lerch's book "An Introduction to Audio Content Analysis", the wavelet transforms based filter banks have not been successfully incorporated in music analysis because 1. a filterbank with reasonable frequency resolution is computationally intensive, and 2. the lack of perfect reconstruction ability that can convert the frequency analyzed audio back to time domain. Which is why, audio applications still use STFT based analysis methods. Numerous papers by Dr. Lerch and others [12] [13] [8] and the field of Music information retrieval is based on STFT analysis.

Moving on to Indian Classical Music analysis, the conventional characterization and identification methods are using Dynamic Time Warping, distance measure and statistical modelling [5] [14]. The analysis is still heavily reliant on deep learning models and the accuracy still lacks compared to popular music. People have also tried to use n-grams for analyzing the ragas [14].

Current state of the art: "Shazam", which is a very successful application of audio fingerprinting, uses STFT and MFCC/spectral features. It then hashes the hamming distance between the spectral peaks and matches it with the input song.

Deep learning has also been very effective in classification of music. Some of the most famous models have been the Long Short-Term Memory, the bidirectional Long Short-Term Memory, the Transformer models, Generative Adversarial Networks and Siamese Networks. These models are state of the art in achieving the highest accuracy in classification.

## 3. Proposed approach

I approached this problem with an open mind. Discrete wavelet transforms, however famous have not invoked much research interest for Indian Classical music. DWTs can prove to be better or worse for Indian Music Classification. The aim of this study is to find whether DWT matches up with STFT based audio features.

My study took the following steps:

### 3.1. Focus points

This study is performed in a very short amount of time. Hence, there have to be points that need to stay in focus and things that can be compromised on when conducting the experiments. I have decided to explore the following wavelets: bior1.1, coif10, db10, rbio1.5, sym4, bior1.5, coif3, db12, rbio2.4, sym6, bior2.4, coif6, db4, rbio6.8, sym9, bior3.5, coif8, db8, rbio1.3, sym16. I will initially test the wavelets for a decomposition to level 5. I will then assess the best 4 wavelets to a decomposition of level 3 and 7. The final discussion will be about the comparisons between STFT and wavelet based MFCC performance.

The complexity of the neural network I use has to be enough to resolve the data and learn from it properly; and at the same time the training time needs to be small enough to accommodate all the wavelets I wish to explore. A bidirectional LSTM seems like a really good option in this context.

### 3.2. Using MFCCs as features

I decided to use MFCCs to process the audio signals. Numerous python libraries like Librosa, Musi21, Pytorch have fuunctions to extract MFCC features from an audio signal. I went ahead with pytorch functions to calculate MFCCs since they are based on tensor processing and can handle audio signal of multiple channels at the same time. I am using 20 MFCCs per window of the audio clip.

### 3.3. Neural Network

As mentioned in an earlier paragraph, I am using a bidirectional LSTM model as a classifier for my ragas. I need to however use a CNN before the bi-LSTM layers to understand the 3 dimensional data as input ($levels * number\ of\ mel\ coefficients * time\ stamps$). A CNN can create coefficients from this input layer that can represent multiple layers. This also enables us to use the same network irrespective of levels of decomposition, and hence the STFT based MFCCs can also be used here.

Fig 6 shows the flow of my model.

### 3.4. Datasets

I am using three datasets in my analysis. The MNIST dataset and FSD dataset [1] from google and the saraga1.5 Hindustani music dataset [2]. The first two are to establish a sanity for my model and to see if STFTs can be used at all for classification purpose. I will then move on to the Indian Classical Music dataset and use wavelet decomposition.

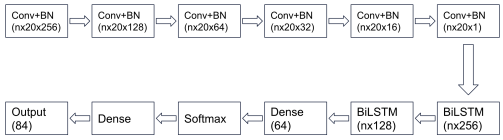


Figure 6. Bidirectional LSTM + CNN based neural network used used in my project

The Indian Classical Music dataset has several recordings of Indian ragas and 84 ragas in total. This makes this a very complex classification task.

### 3.5. Wavelet Decomposition

As mentioned earlier, I am using several wavelets that I want to test and detect which ones work best. I am using the pywavelets package to do my wavelet decomposition. I will then create a single numpy array to save all the audio channels and compute MFCCs on them.

## 4. Results and Discussion

The results from my experiments with 5 level wavelet decomposition is listed in Table 1. We can see that one wavelet of each type performs better than the others based on their shape. I then chose the best 4 wavelets from this, which are: coif10, bior1.5, rbio2.5, db4. We can note here that apart from coif10, all other wavelets are the smallest among the ones chosen in their wavelet families.

For the next experiment, I change the wavelet decomposition levels with the 4 chosen wavelets to 3 and 7. I train the network again and the observation is that the best accuracy is achieved at decomposition level = 5 and wavelet db4.

Now we put into perspective the training and testing accuracies from STFT based MFCCs. They are listed in Table 4, 5 and 6. We see a dramatic difference in accuracy for STFT based MFCCs. It is certain that the model is training well and performing well with the Google and MNIST datasets, however with the Indian music, the accuracy falls very low. This can indicate a few things: 1. that the data for Indian classical music does not suit the model that I am using, 2. that the STFT based MFCCs are very poor at classification of Indian music and wavelets do perform very well. The second option is more likely here since all the wavelets perform relatively well for the same model. This is a strong indication that the wavelets are much better at classification of Indian Classical music ragas.

| Wavelet | Training Acc | Testing Acc |
|---|---|---|
| bior1.1 | 95.44% | 95.17% |
| coif10 | 96.01% | 97.02% |
| db10 | 96.48% | 95.21% |
| rbio1.5 | 96.14% | 97.26% |
| sym4 | 95.37% | 96.57% |
| bior1.5 | 96.39% | 97.59% |
| coif3 | 96.26% | 96.23% |
| db12 | 97.37% | 96.55% |
| rbio2.4 | 96.19% | 97.07% |
| sym6 | 94.15% | 93.91% |
| bior2.4 | 89.20% | 93.99% |
| coif6 | 96.38% | 95.94% |
| db4 | 96.57% | 97.60% |
| rbio6.8 | 96.64% | 96.36% |
| sym9 | 96.01% | 96.39% |
| bior3.5 | 96.36% | 96.18% |
| coif8 | 95.85% | 96.79% |
| db8 | 97.37% | 96.55% |
| rbio1.3 | 95.36% | 96.57% |
| sym16 | 95.73% | 94.17% |

Table 1. Model performance for level=5

| Wavelet | Training Acc | Testing Acc |
|---|---|---|
| bior1.5 | 94.99% | 94.60% |
| db4 | 95.43% | 96.35% |
| coif10 | 95.26% | 94.29% |
| rbio2.4 | 96.18% | 96.48% |

Table 2. Model performance for level=3

| Wavelet | Training Acc | Testing Acc |
|---|---|---|
| bior1.5 | 96.07% | 96.03% |
| db4 | 96.12% | 96.81% |
| coif10 | 96.78% | 97.07% |
| rbio2.4 | 96.11% | 95.91% |

Table 3. Model performance for level=7

| raining Acc | Testing Acc |
|---|---|
| 36.24% | 38.13% |

Table 4. Model performance for level=7

## 5. Conclusion

The data and results conclude that wavelet based processing outperforms STFT based audio processing by a huge margin. This makes wavelet processing the best choice for audio analysis and for a silicon cochlea.

| Training Acc | Testing Acc |
|---|---|
| 99% | 97% |

Table 5. STFT based classification for MNIST dataset

| Training Acc | Testing Acc |
|---|---|
| 98% | 90% |

Table 6. STFT based classification for FSD dataset

## References


[1] Mnist audio dataset. 4

[2] Saraga1.5 hindustani music dataset. 4

[3] M. I. Abdalla and H. S. Ali. Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models, 2010. 3

[4] G. F. Choueiter and J. R. Glass. "a wavelet and filter bank framework for phonetic classification. 2005. 3

[5] A. K. Sharma et al. Classification of indian classical music with time-series matching deep learning approach, 2021. 3

[6] G. Essl G. Tzanetakis and P. Cook. Audio analysis using the discrete wavelet transform. 3

[7] M. Gupta and A. Gilbert. Robust speech recognition using wavelet coefficient features. 2001. 3

[8] S. Gururani and A. Lerch. Automatic sample detection in polyphonic music. 3

[9] Constance Jones and James D Ryan. *Encyclopedia of hinduism*. Infobase publishing, 2006. 2

[10] Jean Morlet Kronland-Martinet, Richard and Alexander Grossmann. Analysis of sound patterns through wavelet transforms. *International journal of pattern recognition and artificial intelligence*. 3

[11] Richard Kronland-Martinet. The wavelet transforms for analysis, synthesis, and processing of speech and music sounds. 1988. 3

[12] Alexander Lerch. On the evaluation of generative models in music, 2020. 3

[13] P. Seshadri and A. Lerch. Improving music performance assessment with contrastive learning. 2021. 3

[14] Prabha Kumari Yogesh H. Dandawate and Anagha Bidkar. Indian instrumental music: Raga analysis and classification. 2015. 3