International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 704-715
https://doi.org/10.70917/ijcisim-2026-5108
Article

A DEEP LEARNING BASED PITCH ANALYSIS MODEL TO IDENTIFY PITCH ANOMALIES AND ASSESSING THE HARMONY OF INSTRUMENT AND VOCAL COMPONENTS

Renju K[1,2], Ashok Immanuel V[3]

[1] CHRIST (Deemed to be University), Bengaluru, Karnataka, India.
[2] Mount Carmel College (Autonomous), Bengaluru, Karnataka, India.renju.k@mccblr.edu.in
[3] CHRIST (Deemed to be University), Bengaluru, Karnataka, India.ashok.immanuel@christuniversity.in
Corresponding Author: Ashok Immanuel V (ashok.immanuel@christuniversity.in)

Abstract: Due to its intricate pitch structures and gamakas or ornamentations and raga specific rules, Carnatic music face challenges in automated analysis of vocal performances. Researchers have proposed innovative approaches in music information retrieval on categorizing different types of music, identifying instruments and pitch extraction, little research have been made on assessing the harmony of instruments and vocals and also in automated analysis of pitch anomalies of vocal performances. An innovative pitch analysis PA deep learning model is proposed in this paper to detect pitch anomalies in vocal performances and also evaluating the harmony between the vocals and accompanying instrument violin. The proposed PA model incorporates the CREPE deep learning model for fundamental frequency (Sa) extraction directly from time-domain audio signals. The pitch values obtained are mapped to remaining fifteen notes in Carnatic music and hence pitch anomalies are identified for those notes which are not part of raga's arohana and avarohana. From the experimental results, it is observed that the proposed PA model shows good capability in accurately identifying pitch anomalies across multiple Carnatic ragas and assessing accompaniment coherence with the vocals. This research provides a comprehensive tool for music analysis, with potential applications in music education, performance evaluation, and intelligent music tutoring systems.

Keywords: Pitch Analysis, Anomaly Detection, CREPE, Harmony Assessment, Deep Learning, Carnatic Music..

1. INTRODUCTION

One of the basic distinctive feature of sound called pitch, is defined as the rate at which the chords in the vocal tract vibrate. Pitch is derived from its fundamental frequency or tonic and it exists in voiced speech only. Analyzing the pitch of unvoiced speech such as whispering, whistling, hissing need not be considered as they do not contain fundamental frequency. Fundamental frequency identification is. extremely useful in wide range of applications such as identifying male/female voice, emotion recognition, recognizing the music notes and many more. Many audio applications such as speech recognition, speaker identification, automatic music transcription, identifying the music tones have shown great impact in the field of research by analyzing the fundamental frequency of sound waves. Hence pitch analysis is an essential step in analyzing the performance of music. Researchers have discovered many pitch detection algorithms which are based on frequency domain, time domain and also based on statistical methods. Unfortunately no algorithm gives a precise result in analyzing the pitch of an audio as the complex sound waves produced by the humans are not perfectly periodic. Also, the signal may have some disturbances as the sound waves travel through the vocal tract and lips would be challenging in perfectly identifying the periodicity

For speech analysis, this periodicity may be identified easily but this task becomes more challenging in music analysis. Researchers have discovered pitch detection algorithms such as PYIN based on time domain, YAAPT based on time and frequency domain and Crepe algorithm which is based on deep neural network. Before implementing these algorithms, the signal is filtered in order to exclude low frequencies which would be very low for human auditory system and hence need not be considered. Most of the pitch detection algorithms divide the audio signal in to short frames although researchers have also proposed methods which divide the signal in to sample-by-sample [1]. The supposition here is that within short frames, the frequency is stationary ,hence it can be analyzed from each frame. The PYIN algorithm based on autocorrelation method is very effective in analyzing high pitched voices in music and also in detecting periodicity in signals [2]. The autocorrelation function would give the significant peak positions related to the period of the wave but in some cases the sound waves produced may not be perfectly periodic. It was observed that because of the complex structure of the signal, the first maximum value after applying.autocorrelation function to the signal, may not be the desired fundamental frequency. So PYIN algorithm uses cumulative mean normalized difference function which would make the peaks more clear and visible by reducing the sensitivity of the signal to amplitude modulations. This algorithm is simple and convenient because of its logic though it is not that precise in giving the result. Another pitch detection algorithm, YAAPT(Yet Another Algorithm for Pitch Tracking) based on time and frequency domain, implement normalized cross correlation function which is similar to autocorrelation function but very good at understanding the fast changes in signal . It was observed that normalized cross correlation function works better than autocorrelation function in giving peak positions with more clarity.

2. RAGA AND SWARAS IN CARNATIC MUSIC

Carnatic music follows a set of rules and structures, and each composition in Carnatic music adheres to these rules unless the composition specifically calls for some variations within the framework of the raga [14].. The foundation of Carnatic music lies in the perception of ragas and swaras. Ragas are melodic frameworks or scales that provide the quintessence and mood for a musical composition. Each raga has a unique set of notes called arohana (ascending notes) and avarohana (decending) notes, which specify the sequence of notes to be used while performing or rendering that raga. The arohana and avarohana of a raga can vary from one raga to another, creating a distinct musical identity for each raga. The seven basic swaras or notes used in Carnatic music are ‘Sa’ (Shadja), ‘R’ (Rishabh), ‘G’(Gandhar), ‘M’ (Madhyam), ‘P’ (Pancham), ‘D’(Dhaivat), and ‘N’ (Nishad). The notes ‘Sa’ and ‘P’ have fixed pitches, while the other notes have variations[3]. These variations are expressed through gamakas, which are embellishments or oscillations around the basic pitch of a swara. Gamakas add depth, beauty, and character to the melodic rendition of a composition. In Carnatic music, there are three main octaves or saptaks: the lower octave (mandra saptak), the middle octave (madhya saptak), and the higher octave (taar saptak). Each octave consists of seven notes or swaras, hence the term "saptak." The lower octave is recognized by lower pitches, the middle octave covers a wider range, and the higher octave comprises higher pitches. These octaves provide flexibility and range to explore the melodic landscape of a composition. The pitch of all the swaras in Carnatic music is relative to the base note ‘Sa’ (Shadja). Hence, the pitch of the note ‘Sa’ is fixed and is called as tonic or fundamental frequency, and all other swaras are derived based on the intervals and relationships defined within a raga. By maintaining the proper scale or pitch of Sa, musicians can accurately render the notes of a raga and maintain the melodic structure of a composition[4]. The swaras with the respective ratio is given in Table 1.

Table 1: Swaras and Ratios in Carnatic Music

Swara	Ratio	Swara	Ratio
Sa	1	Ma2	17/12
Ri1	16/15	Pa	3/2
Ri2	9/8	Dha1	8/5
Ri3	6/5	Dha2	5/3
Ga1	9/8	Dha3	9/5
Ga2	6/5	Ni1	5/3
Ga3	5/4	Ni2	9/5
Ma1	4/3	Ni3	15/8

Computing the pitch of swaras in relation to ‘Sa’ note is an important aspect of learning and performing Carnatic music. It allows musicians to understand the melodic framework, maintain tonal consistency, and navigate through the various ragas and compositions with accuracy.

3. LITERATURE REVIEW

Researchers Faghih and Timoney[1] looked at four different ways to figure out the pitch of singing phrases. They compared pYIN, Praat, PLL-based, and the Extended Complex Kalman Filter to see which one worked best. The first two methods look at the sound in small chunks, while the last two look at it one sample at a time, which is better for real-time applications. To test these methods, they used a special tool called Spear to create a standard pitch contour for 76 audio samples of solo voices. This helped them see which method was the most accurate. What they found out was that pYIN and Praat were really good at getting the pitch right, with hardly any mistakes. On the other hand, the PLL and Kalman filter methods were very sensitive to the settings used, and often got the pitch wrong, especially with male voices. They also had problems with octave errors, which means they got the pitch too high or too low. When they looked at the standard deviation, they saw that pYIN was still the most accurate. So, even though the sample-by-sample methods like PLL and ECKF could be useful for real-time applications, they have some limitations. They need to be adjusted just right for each specific input, which makes them less useful for analyzing singing voices in general. This means that pYIN and Praat are still the best choices for getting the pitch just right.

Riley and Dixon[5] present CREPE notes, a new post-processing algorithm that divides pitch contours into discrete notes by extending the state-of-the-art monophonic pitch tracker, CREPE. As opposed to conventional onset-based segmentation methods, this approach integrates CREPE's confidence measures with pitch gradient information to further improve boundary detection, even in legato sections. The system robustly processes repeated notes through an extra onset detection mechanism and uses amplitude-based filtering to remove spurious segments. Assessed on two demanding instrumental datasets—Filosax (saxophone solo performances) and ITM-Flute-99 (Irish flute recordings)—the technique performs better than prevailing models like PYIN, Basic Pitch, and MT3, but with much lower model complexity. The technique is fast, precise, and generalizable and has the potential for enhanced automatic music transcription, especially for solo instrumental music.

Kroon[6] compares three pitch detection algorithms such as pYIN,YAAPT and CREPE to determine the efficiency of the use of neural networks for pitch estimation compared to conventional approaches. pYIN employs time domain processing YAAPT integrates time and frequency domain characteristics, and CREPE is a deep convolutional neural network that estimates pitch directly from raw audio waveforms. The assessment was done with a phonetically rich speech corpus and reference pitch trajectories obtained from laryngogram signals. The algorithms were compared on voicing errors, gross pitch errors and fine pitch accuracy. Experiments indicate YAAPT produced the lowest voicing errors, but, CREPE even with training being limited to musical data competed well with pYIN and YAAPT, particularly in terms of fine pitch accuracy. The research finds that neural network driven techniques such as CREPE are promising alternatives to traditional methods, although voice specific training data would further enhance performance. Schroter et.al [7] explores deep learning-based methods for pitch estimation in automatic music transcription (AMT) with the application of convolutional neural networks (CNNs). A Constant-Q Transform (CQT) spectrograms were proposed by the authors as input into a deep CNN to predict pitch classes for polyphonic music. It has already been proved by the researchers that the deep learning is very powerful in capturing high complex spectral patterns as compared to traditional signal processing techniques which often fail to capture overlapping harmonics especially with polyphonic music The model is trained and validated on the MusicNet dataset, and the performance is compared with state-of-the-art pitch estimators like CREPE deep learning pitch estimation model. The results show that even under difficult multi-pitch scenarios, CNN-based model gains better precision and recall.. The capability of deep learning approaches in accurately identifying pitch values from an audio has been utilized in this research. The piano notes were detected accurately by convolutional neural network(CNN) model in the research conducted by Orchisama Das et.al [8] where the spectrogram images were taken as input to CNN model to predict the correct note. In contrast to conventional onset detection approaches based on hand-crafted features or energy-based heuristics, the present method automatically learns the relevant features during supervised training. Experimental results based on the MAPS dataset demonstrate that the CNN-based approach attains higher F1-scores than the conventional baseline methods. The research illustrates the capability of deep learning to enhance onset detection performance under conditions of complicated acoustic conditions with soft attacks or note overlaps.

This work [9] presents a fully-convolutional neural network (FCN) model for accurate and efficient monophonic pitch (F0) estimation in speech signals When it comes to pitch estimation in speech, earlier models like

CREPE had some limitations. But now, researchers have come up with three new models - FCN-1953, FCN-993, and FCN-929 - that make things simpler and faster without losing any accuracy. These models were trained on fake speech data created using the PaN vocoder, which gives really precise information about the pitch. Compared to CREPE and SWIPE, these new models do a better job of figuring out the pitch, and they're way faster too. They can even handle longer pieces of audio at once, which makes them more efficient. The FCN-993 model is especially good at balancing speed and accuracy, making it perfect for real-time speech processing. Another study [10] looked at using a type of neural network called CNNs to estimate pitch in speech. The researchers trained their CNN model on fake speech data created using a formant synthesizer, which ensures that the pitch is always perfect. The model takes in special kinds of sound pictures called log-magnitude spectrograms and outputs really detailed pitch information. When they tested it, they found that their CNN model did better than traditional algorithms like YIN and SWIPE, especially when it came to handling noisy audio or needing to make quick decisions. What's really cool is that even though the model was only trained on fake data, it still worked really well with real speech. This just goes to show how powerful supervised learning can be - by using data to train the model, they were able to get better results than they would have with handmade features.

Xu and Shimodaira[11] introduce a new end-to-end neural network architecture for direct F0 estimation from raw speech waveforms. In contrast to standard or classifier-based pitch trackers such as YIN, Praat, CREPE, the approach factorizes the problem into two distinct modules: a voice detector and a value estimator. The voice detector is realized with a deep feedforward network with dropout and batch normalization, while the value estimator uses a novel "value decoder" architecture that outputs distributed representations for F0 regression.. Trained on the PTDB-TUG database and tested under both clean and noisy environments (with NOISEX-92 noise files), their tracker better performs than traditional and neural network-based trackers with regard to voice decision error (VDE), gross pitch error (GPE), and fine pitch estimation accuracy. The system specially performs well under clean environments and shows comparable robustness when under noise, confirming the usefulness of decoupling detection and estimation in maximizing pitch tracking performance.

Drugman et al. [12] suggest a pitch detection framework using conventional machine learning (ML) approaches in contrast to current deep learning–based strategies. The approach discriminates between voicing detection as classification and F_0 estimation as regression, employing hand-crafted features derived from time, spectral, and cepstral domains, as well as from a mean-based signal (MS). For voicing detection, K-means and supervised models such as Multi-Layer Perceptrons (MLP) were experimented. K-means showed a 20% relative improvement in voicing decision error (VDE) over state-of-the-art baselines, whereas the MLP model showed a 45% reduction. For estimation of F_0, a basic median filter applied to multiple estimators performed better than more sophisticated models and even the deep-learning-based CREPE model with respect to Gross Pitch Error (GPE). The target ML-based pitch tracker also demonstrated substantial gains in a speech synthesis implementation, competing with and even surpassing traditional pitch trackers like RAPT and DIO. The work points out that with well-designed features and limited data, traditional ML approaches can be competitive or better than some complex deep models. Kim et al. [13] propose CREPE, a pitch estimation model based on a deep convolutional neural network (CNN) that operates directly on raw audio waveforms. Unlike traditional pitch trackers such as YIN or pYIN, which rely on time- or frequency-domain heuristics, CREPE learns to predict pitch using a data-driven approach, mapping audio frames to 360 discrete pitch bins covering six octaves. The model is trained on a large, diverse dataset (RWC-synth, MDB-stem-synth, and NSynth) and evaluated across various monophonic music and speech recordings. CREPE achieves state-of-the-art results in both clean and noisy conditions, outperforming traditional algorithms in accuracy and robustness. Additionally, it generalizes well to unseen instruments and vocal timbres. The paper establishes CREPE as a powerful and flexible tool for pitch estimation in music and audio analysis.

4. METHODOLOGY

Identifying pitch anomalies in singing is a complex task and it is subjective and require expertise with knowledge. This task can be automated by identifying incorrect notes sung by the singer and has received limited attention in the existing research. In music competitions, identifying pitch anomalies in singing is one of the criteria for judgement and it’s been done manually which sometimes may lead to unfair assessment. An innovative pitch analysis PA model is proposed in this research which would extract the pitch values from Carnatic music audio and identify the fundamental frequency. Based on the fundamental frequency, the other fifteen notes are detected according to the ratios given in Table 1. A diverse dataset gathered from different sources including CompMusic, Gana, and YouTube videos were used in this research and from these sources, a collection of selected audio recordings were compiled. A structured CSV file containing essential details about each composition was created which included

information about the composition itself, the musical raga it belongs to, and its arohana (ascending notes) and avarohana (descending notes), crucial elements for the study. Hence, the dataset comprised 250 compositions, including performances by both professional and non-professional singers. This extensive dataset served as a significant resource for executing the research and examining diverse facets of pitch tracking and musical analysis. The process flow is given in Fig 1.

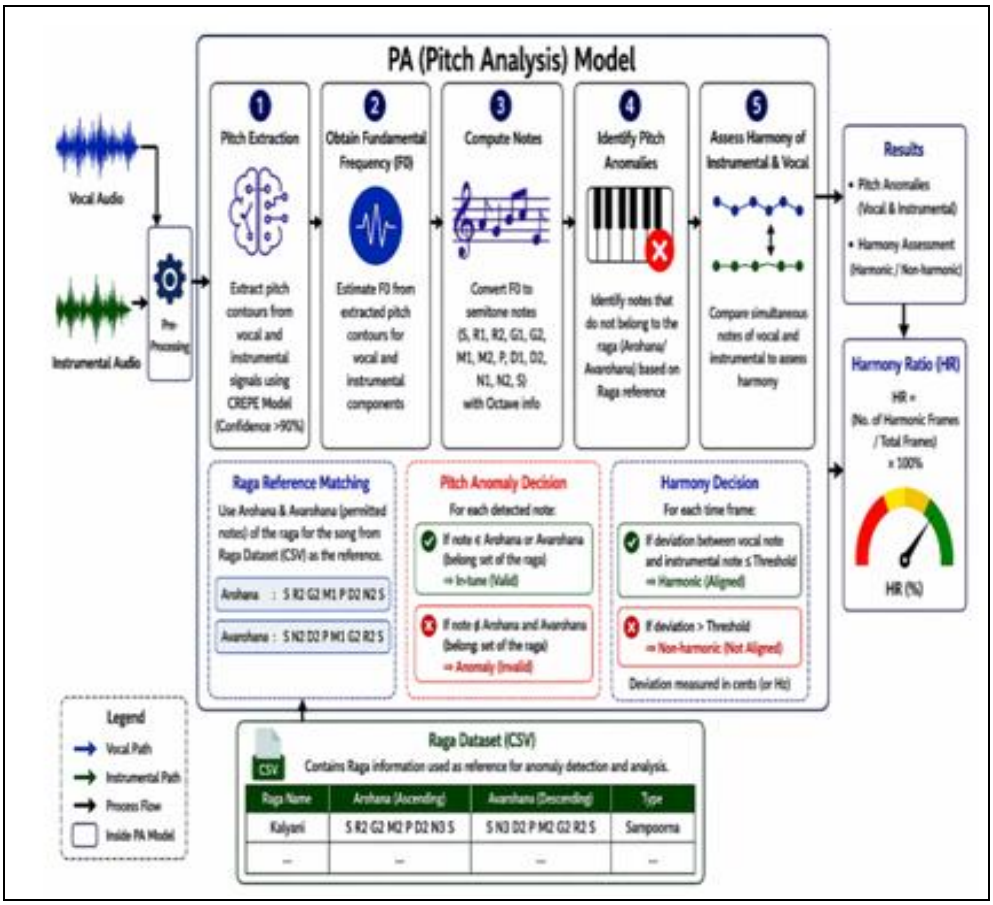

Fig 1: PA (Pitch Analysis) Model to Identify Pitch Anomalies in Carnatic Music and Assessing the Harmony between the Instrument and Vocals

A. Audio Pre Processing

This research utilizes Python library Librosa to load the audio files and for processing. As part of pre processing, all audio files which were in the mp3 format were converted to to the .wav format. For extracting the relevant audio features, appropriate functions from Librosa library were incorporated in this research. It was necessary to divide the audio signals in to smaller segments called frames for better analysis as we know that the audio signals are continuous and dynamic. Each frame contained a set of audio samples that were suitable for further processing. The audio signals were sampled at a rate of 22050Hz in this study, that means 22050 samples were recorded per second. A duration of 60 seconds were extracted from the audio file and the mentioned sampling rate ensured an accurate representation of the audio's frequency content. This extracted sample provided sufficient data for analysis while managing computational resources effectively. The audio files were also converted from stereo to mono which used only a single channel and also normalized as part of preprocessing of audio. Signal segmentation, various transformations and feature extraction techniques were incorporated from Librosa's extensive collection of functions enabling comprehensive preparation of the audio data for further analysis.

B. Vocal Instrument Separation

Working with polyphonic audio is a challenge and to overcome this complexity and also enable a deeper understanding of the individual components within a polyphonic audio, a fast and reliable music source separation tool called spleeter [15] was employed in this study. Spleeter is very useful in avoiding the interference with different sounds and hence it is chosen for this research. With its emphasis on speed, clarity, and ease of use, spleeter divided the audio signal into distinct stems, offering several options for source separation. A 4-stem separation was chosen for this study which classified the audio signal into vocals, drums, bass, and other components. For pitch analysis of the vocals and also to assess the harmony between instrument and vocals, extraction of these components separately was necessary. This allowed to have a more focused analysis on each component-vocals and instrument The decision to implement a 4-stem separation appeared fitting, considering the research objectives and the desired level of granularity. Spleeter's versatility extends to 2-stem and 5-stem separations as well, providing additional flexibility for diverse research requirements. Ultimately, spleeter's combination of speed, reliability, and user-friendliness, coupled with its ability to effectively separate polyphonic sounds, rendered it an optimal choice for this research investigation on music source separation.

C. Pitch Analysis Using PA Model

The proposed PA (Pitch Analysis) model which incorporates CREPE, a deep convolutional neural network designed for pitch estimation directly from time-domain audio signals. It processes 1024-sample segments of audio, sampled at 16 kHz, through a six-layer convolutional network to extract features. This results in a 2048- dimensional latent representation, which is then passed through a dense layer with sigmoid activations to generate a 360-dimensional output vector. Each node in this output vector corresponds to a specific pitch value defined in cents. The network estimates the pitch by calculating a weighted average of these values, with the weights determined by the activations of the nodes. Training CREPE involves minimizing binary cross-entropy loss between the predicted and target vectors The target vector is Gaussian-blurred around the true frequency to accommodate prediction tolerance. The model is trained using the ADAM optimizer with a learning rate of 0.0002, and incorporates regularization techniques such as batch normalization and dropout to enhance generalization. The architecture and training of CREPE are implemented using the Keras framework. Additionally, CREPE provides a "confidence" measure that evaluates the strength of the pitch information relative to the overall audio signal.

D. Identify Pitch Anomalies in the Audio

The audio file separated from instruments was processed using the Librosa library in Python, sampled at 48,000 Hz, starting at 60 seconds and lasting for 120 seconds. The PA model was then used to extract pitch values and their corresponding confidence levels at intervals of every 10 milliseconds. To accurately identify the fundamental frequency (Sa), only pitch values with a confidence level above 0.95 were considered, filtering out less reliable data. As the singer performs the melody, the pitch values corresponding to the notes will repeat regularly, reflecting the structure of the musical scale. The most frequently occurring pitch, identified as the fundamental frequency (Sa), was extracted from the CREPE output as shown in (1).

f_{sa}=mode(F_0) ……(1),

Where f_{sa} represents the fundamental frequency and F0 denotes the extracted pitch values from PA

model

Using this fundamental frequency, the corresponding notes—R1, R2, R3, G1, G2, G3, M1, M2, P,

D1, D2, D3, N1, N2, and N3—were computed as defined in (2) and ratios given in Table 1.

f_{note}= f_{sa} X T_{note}……(2)

Where T_{note} represents the ratio corresponding to each swara.

Next, the raga of the song, along with its arohana (ascending scale) and avarohana (descending scale) notes, were retrieved from the dataset. These raga notes were then compared with the notes derived from the fundamental frequency obtained from PA model. If the occurrence of a particular note in the melody exceeds a threshold of 100, it indicates a deviation from the raga and pitch anomaly is identified, suggesting that the singer may have sung an incorrect note. Hence pitch anomalies denoted by A is defined as the collection of detected notes that do not belong to the given raga as given in (3).

$$A=\{ n \in N_{detected} \mid n \notin N_{raga} \} \ldots\ldots\ldots\ldots\ldots(3)$$

Where $N_{detected}$ represents the set of detected notes from the audio and N_{raga} denotes the valid set of notes or swaras derived from the arohana and avarohana of the raga.

E. Assessing the Harmony of Instrument and Vocals

The PA model is used to not only identify the pitch anomalies and also in assessing the harmony between the vocal and instrument using the extracted pitch values obtained from the model. The filtered fundamental frequency sequences of vocals and instrument corresponding to high-confidence estimates with confidence greater than or equal to 95% and frequency greater than or equal to 100 Hz are considered for analysis and temporally aligned for frame-wise comparison. The absolute frequency deviation as shown in (4) is computed for every time frame and a tolerance threshold of ±5Hz is applied to account for minor pitch changes.

$$| f_{voc}(t) - f_{inst}(t) | \leq \delta \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4)$$

Where $f_{voc}(t)$ and $f_{inst}(t)$ represents the vocal and instrumental frequencies at time t and δ denotes the tolerance threshold given as ±5Hz.

The harmony is evaluated if the deviation lies within the threshold as given in (5) and a harmony ratio is computed which represents the percentage of instances where the vocal and the instrument are in harmonic alignment out of total observations.

$$H(t) = \begin{cases} 1, & | f_{voc}(t) - f_{inst}(t) | \leq \delta \ldots\ldots\ldots\ldots\ldots\ldots(5) \\ 0, & \text{Otherwise} \end{cases}$$

Where H(t) =1, indicates harmony , H(t)=0 indicates disharmony.

5. RESULTS AND DISCUSSION

The different compositions of male and female singers were fed in to the PA model. The fundamental frequency or ‘Sa’ note obtained from PA model is used to calculate the other musical notes and pitch anomalies are observed based on the arohana and avarohana of the raga extracted from the dataset.

A. Identify Pitch Anomalies

The fundamental frequency or ‘Sa’ note is identified as 396 Hz for the audio. The pitch contour obtained from PA model of the composition Sadapalaya: Raga Mohanam is given below in Fig 2 and the frequencies of each of the musical note for the raga Mohanam is given in Table 2. The interpretation of pitch anomalies obtained from the audio is given in Table 3.

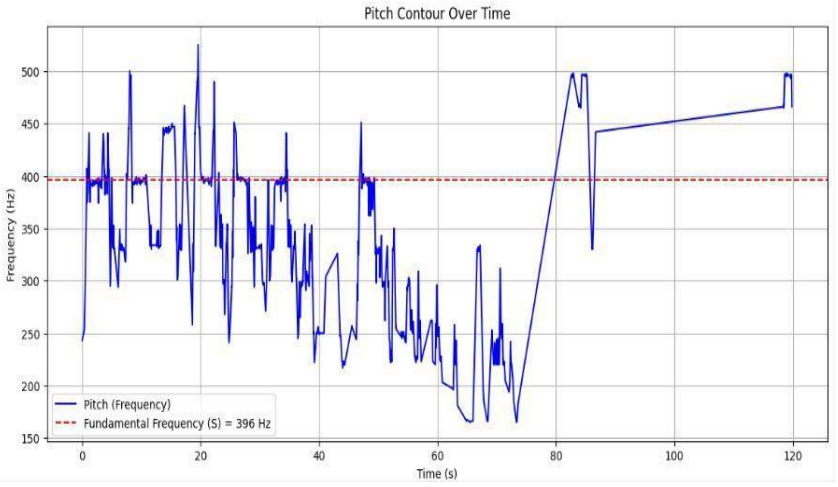

Fig 2: Pitch Contours Over Time of the Audio

Table 2: Frequency Computed For The Musical Notes Of Raga Mohanam

Fundamental Frequency:396 Raga: Mohanam	
Notes	**Frequency**
Sa	396
R2	446
G3	495
P	594
D2	660

Table 3: Interpretation Of Wrong Pitch Values Obtained From Audio

Raga: Mohanam

Wrong Pitch Notes Extracted from the Audio(Hz)	No of Occurrences	Remarks
441	24	Close to R2-Can be considered as the correct note
392	68	Near Sa note, likely an incorrect or fluctuating pitch
394	140	Near Sa note, likely an incorrect or fluctuating pitch
391	23	Near Sa note, likely an incorrect or fluctuating pitch
393	93	Near Sa note, likely an incorrect or fluctuating pitch
396	82	Near Sa note, likely an incorrect or fluctuating pitch
397	139	Near Sa note, likely an incorrect or fluctuating pitch
398	94	Near Sa note, likely an incorrect or fluctuating pitch
399	18	Near Sa note, likely an incorrect or fluctuating pitch
443	8	Slightly off from R2, possibly an incorrect pitch
445	11	Slightly off from R2, possibly an incorrect pitch
500	1	Very negligible, not a concern
498	3	Very negligible, not a concern
494	1	Very negligible, not a concern

In the analysis of the audio rendition of Raga Mohanam, which follows the arohanam (Sa R2 G3 P D2 Sa), the fundamental frequency of 396 Hz corresponding to the tonic "S" was accurately identified and predominantly maintained. However, notable pitch anomalies were observed. The presence of the 441 Hz frequency, corresponding to R1 (Shuddha Rishabha), with 24 occurrences, indicates a deviation from the raga's structure, as this note does not belong to Mohanam. Additionally, frequencies ranging from 392 Hz to 399 Hz, which are close to the tonic S, showed significant occurrences, reflecting pitch instability or inaccuracies. These deviations suggest that while the singer maintained the core notes of the raga, there were inconsistencies in pitch, particularly around the tonic, and the introduction of an extraneous note (R1), which could affect the raga's purity. These findings highlight the importance of precise pitch control in Carnatic music, especially in adhering to the raga's structure.

After obtaining the fundamental frequency for the audio as 400 Hz, the frequencies of each of the musical notes for the raga Ranjani is computed and is given in Table 4. The interpretation of pitch anomalies obtained from the audio is given in Table 5.

Table 4: Frequency Computed For The Musical Notes Of Raga Ranjani

Fundamental Frequency:400 Raga: Ranjani	
Notes	**Frequency**
Sa	400
R2	450
G2	480
M2	566
D2	666

Table 5: Interpretation Of Wrong Pitch Values Obtained From Audio

Raga:Ranjani

Wrong Pitch Notes Extracted from the Audio(Hz)	**No of Occurrences**	**Remarks**
407	16	Near S note, likely an incorrect or fluctuating pitch
404	40	Near S note, likely an incorrect or fluctuating pitch
403	89	Near S note, likely an incorrect or fluctuating pitch
402	160	Near S note, likely an incorrect or fluctuating pitch
401	83	Near S note, likely an incorrect or fluctuating pitch
405	18	Near S note, likely an incorrect or fluctuating pitch
398	28	Near S note, likely an incorrect or fluctuating pitch
399	21	Near S note, likely an incorrect or fluctuating pitch

406	30	Near S note, likely an incorrect or fluctuating pitch
397	25	Near S note, likely an incorrect or fluctuating pitch
451	8	Slightly lower than G2 - Possible anomaly
452	1	Near R2, Count is negligible, not a concern
432	1	Very negligible, not a concern
431	1	Very negligible, not a concern
400	1	Very negligible, not a concern
450	2	Very negligible, not a concern
456	1	Very negligible, not a concern

B. Assessing the Harmony of Instrument and Vocals

The pitch contours of violin and vocals are shown in Fig 3 and Fig 4 respectively which demonstrates that violin has a smoother and more consistent pitch contour compared to pitch contours for vocals which show greater pitch variability. This highlights the natural differences between violin instrument and human vocals. As seen in the Fig 4 and Fig 5, both violin and vocals share same fundamental frequency or base pitch (Sa ≈ 311–312 Hz). However, the spikes in vocals indicate occasional pitch deviations, which might cause temporary dissonance or expressive ornamentation. When the violin maintains a steady pitch around 311 Hz, it likely provides a harmonic foundation for the vocals. If the vocal pitch stays close to the violin's pitch or aligns at harmonic intervals (octaves, fifths), the result is consonant harmony. Deviations from the violin's pitch could either be deliberate like gamakas in Carnatic music or unintentional.

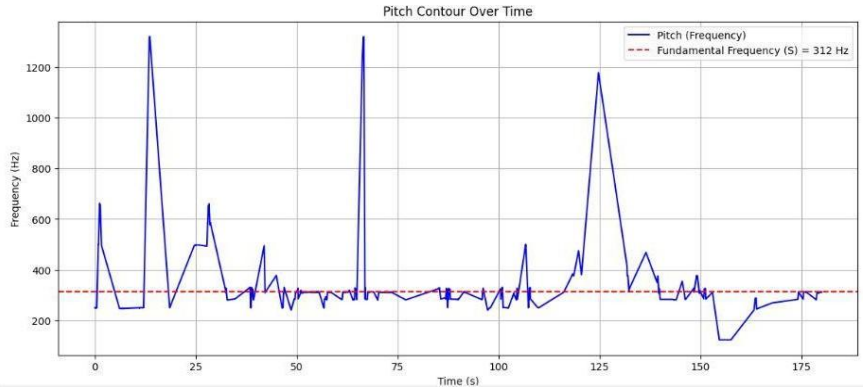

Fig 3: Pitch Contours Obtained for the Instrument Violin

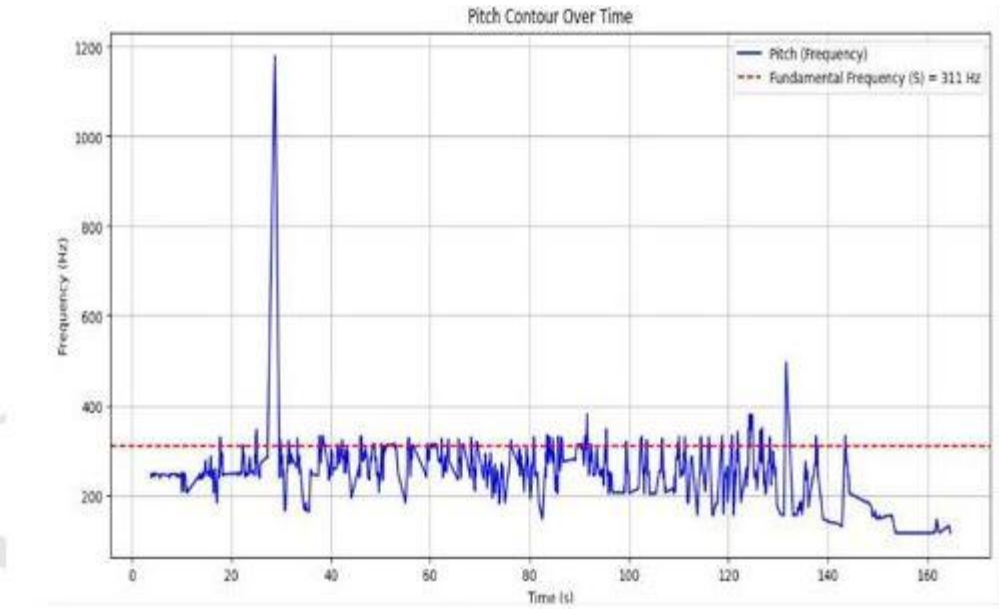

Fig 4: Pitch Contours Obtained for the Vocals

Hence, the above results shows that the harmony between the vocal and violin signals is defined as frame-level pitch compatibility, where both signals share a common tonic reference and their instantaneous fundamental frequencies remain within a perceptually acceptable tolerance range, regardless of differences in local pitch contours. Although the vocal and violin pitch contours may exhibit different local variations due to expressive mechanisms such as gamakas in singing and bow-controlled smoothing in violin performance, harmony is assessed based on pitch proximity rather than contour similarity. The violin, functioning as an accompanying instrument in Carnatic music, is expected to support the vocalist's pitch framework rather than replicate the vocal pitch trajectory exactly.

6. CONCLUSION

This work provides an original pitch analysis PA model which demonstrates the efficacy of sophisticated pitch detection methods in the interpretation of Carnatic music with emphasis on detecting pitch irregularities within the musical structure of a raga and testing the harmony between vocals and instruments. The deep learning PA model was able to identify the fundamental frequency dynamically and therefore detect pitch error that conventional techniques might have missed. There were substantial pitch anomalies or off notes extracted from the audio which are not in the raga and the PA model accurately detected these; and these were then used as input data. The findings underscore the need for accurate pitch control, and the success of AI deep learning model to facilitate accurate raga performance. It also presents a sophisticated computationally based approach to examine the harmonic alignment of vocal and instrumental components through deviation-based analytical methods and Harmony Ratio measurement. The framework will quantitatively report the amount of pitch synchronization, which in part allows precise distinction between harmonically congruent or incongruous segments. This method is validated with empirical evidence and this is reflected in the consistency of this tool when applied to a broad range of audio samples and can therefore be useful in modelling the interactions in complex musical signals. The proposed framework demonstrates significant prospects of further studies in the domain of automated performance evaluation, music pedagogy, and intelligent feedback systems to provide a framework for complex structures, notably those in Carnatic music.

References:

1. Behnam Faghih, Joseph Timoney, An investigation into several pitch detection algorithms for singing phrases analysis,IEEE, https://doi.org/10.1109/ISSC.2019.8904943, June 2019
2. Shuzhuang Xu, Hiroshi Shimodaira, Direct F0 Estimation with Neural-Network-based Regression, INTERSPEECH ,September 15–19, 2019, Graz, Austria
3. Vidya Kanthan, https://yuvasangeethalahari.com/
4. Rajshri Pendekar,S P Mahajan,Pranjali Ganoo, Harmonium Raga Recognition, International Journal of Machine learning and Computing", DOI:10.7763 / IJMLC.2013.V3.336, 2013
5. Xavier Riley, Simon Dixon, CREPE Notes: A new method for segmenting pitch contours into discrete notes, Proceedings of the Sound and Music Computing Conference 2023, Stockholm, Sweden.

6. Anja Kroon, Comparing Conventional Pitch Detection Algorithms with a Neural Network Approach, arXiv:2206.14357v1, June 2022
7. Hendrik Schroter, Tobias Rosenkranz, Alberto N. Escalante-B., Andreas Maier, LACOPE: Latency-
8. Constrained Pitch Estimation for Speech Enhancement, Research Gate 2021
9. Orchisama Das, Julius O. Smith, Chris Chafe,Improved Real-Time Monophonic Pitch Tracking with the Extended Complex Kalman Filter, J. Audio Eng. Soc., vol. 68, no. 1/2, pp. 78–86, (2020 January/February.). DOI: https://doi.org/10.17743/jaes.2019.0053
10. Luc Ardaillon, Axel Roebel, Fully-Convolutional Network for Pitch Estimation of Speech Signals,HAL Open Science January 2020
11. Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen, Max A. Little, Mads Graesboll Christensen, Robust Bayesian Pitch Tracking Based on the Harmonic Model, IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 27, No. 11, November 2019
12. Pandey, Gaurav, Chaitanya Mishra, and Paul Ipe, TANSEN: A System for Automatic Raga Identification ,IICAI. 2003.
13. Thomas Drugman, Goeric Huybrechts, Viacheslav Klimkov, Alexis Moinet, Traditional Machine Learning for Pitch Detection, IEEE Signal Processing Letters, Vol. 25, No. 11, November 2018
14. Jong Wook Kim, Justin Salamon, Peter Li, Juan Pablo Bello, Crepe: A Convolutional Representation For Pitch Estimation, arXiv:1802.06182v1 [eess.AS] 17 Feb 2018
15. Rajeswari Sridhar,T.V. Geetha, Raga Identification of Carnatic music for Music Information Retrieval, International Journal of Recent Trends in Engineering,Vol. 1, No. 1, May 2009
16. Romain Hennequin, Anis Khlif, Felix Voituret, Manuel Moussallam, “Spleeter: A Fast and State-Of-The Art Music Source Separation Tool with Pre-Trained Models”, ISMIR 2019