IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Automatic Carnatic Swara Transcription from Polyphonic Audio Using Pitch-Class Based Tonic Detection

[1]Shoban S, [2]Sidharth Harish M, [3]Dr S Senthamizh Selvi

[1]Student, [2]Student, [3]Associate Professor

Department of Computer Science and Engineering / ADS,

Sri Venkateswara College of Engineering, Sriperumbudur, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309021

Abstract: Automatic musical audio to symbolic representation is a challenging task, especially for Carnatic music because of microtonal variations and ornamentations. This work systematically generates Carnatic swaras from polyphonic audio recordings that are mostly based on a major scale structure. The approach combines deep learning-based source separation using Demucs and pitch estimation using CREPE, along with octave invariant pitch-class representation and histogram-based tonic detection. A gamaka-aware smoothing strategy is introduced to enhance stability and reduce frame-level noise. The system outputs swara sequences suitable for music analysis and educational applications. Experimental observations indicate that the proposed approach provides reliable tonic detection and accurate mapping of swaras under controlled conditions. Pitch Estimation, Tonic Detection, Swara Estimation, Tonic Estimation, MIR

***Index Terms* -** Carnatic Music, Automatic Swara Transcription, Tonic Estimation, Vocal Separation, Pitch Analysis, Cosine Similarity, Deep Learning, Music Information Retrieval.

I. Introduction

Transcription of music audio into a structured symbolic format is a fundamental problem in Music Information Retrieval (MIR). It enables applications such as music analysis, retrieval, education, and automatic accompaniment systems. While significant progress has been made in the transcription of Western music, the transcription of Carnatic music remains a challenging problem due to its inherent characteristics, including continuous pitch variations, non-linear melodic movements, and the presence of expressive ornamentations known as gamakas.

Pitch estimation forms the backbone of any transcription system, as it directly determines the mapping from audio to musical notes. In recent years, deep learning-based approaches such as CREPE [1] have demonstrated high accuracy in estimating the fundamental frequency from raw audio signals. However, real-world music recordings are often polyphonic, consisting of vocals along with multiple instruments. This introduces additional harmonics and interference, making direct pitch extraction unreliable. To address this, source separation techniques such as Demucs [2] are employed to isolate the vocal component, thereby improving pitch estimation quality.

In Carnatic music, musical notes (swaras) are not absolute but are defined relative to a tonic note (Sa), which varies across performers and compositions. Therefore, identifying the tonic is a critical prerequisite for accurate transcription. Prior work has demonstrated that statistical methods based on pitch-class distributions can effectively estimate the tonic in Indian art music [7]. Building upon these ideas, this work proposes a robust pipeline that integrates pitch estimation, octave-invariant representation, and histogram-based tonic detection.

The primary objective of this research is to generate time-aligned Carnatic swaras from polyphonic audio signals that follow a major scale structure. The proposed approach emphasises robustness and interpretability by combining signal processing techniques with data-driven models. Additionally, a temporal smoothing mechanism is incorporated to reduce frame-level noise and produce musically meaningful swara sequences. The system aims to bridge the gap between raw audio signals and symbolic Carnatic representations, enabling applications in music education and automated analysis.

II. Related Work

Automatic music transcription has been an active area of research in MIR, with various approaches proposed for pitch estimation, source separation, and tonal analysis. Deep learning models, particularly convolutional neural networks, have significantly improved pitch detection accuracy. CREPE [1] is a notable example that performs end-to-end pitch estimation directly from waveform data, providing high temporal resolution and robustness to noise.

Source separation is another critical component in transcription systems. Traditional methods relied on spectrogram masking and matrix factorization, but recent advancements in deep learning have led to waveformbased separation models such as Demucs [2]. These models are capable of effectively isolating vocals from polyphonic mixtures, which is essential for accurate pitch tracking in music signals.

Tonal analysis techniques often rely on pitch-class representations, which reduce pitch information into octave-invariant features. Such representations have been widely used for key detection and harmonic analysis [4, 10]. In the context of Indian classical music, tonic detection has been explored using pitch histograms and distribution-based methods [7]. These approaches leverage the statistical prominence of certain notes to identify the tonic.

Additionally, pitch-class distributions have been utilised for raga recognition tasks [6], demonstrating their effectiveness in capturing tonal characteristics. However, most existing systems focus on classification or detection tasks rather than full transcription into swaras. Furthermore, handling the continuous pitch variations and gamakas in Carnatic music remains a challenging problem.

This work extends prior research by integrating pitch estimation, tonic detection, and symbolic mapping into a unified pipeline specifically designed for Carnatic swara transcription.

III. RESEARCH METHODOLOGY

3.1 System Overview

The proposed system is comprised of the following stages: vocal separation, pitch extraction, pitch-class conversion, tonic detection, swara mapping, and temporal smoothing. The system is designed as a modular pipeline, where each stage contributes to improving the robustness and accuracy of the final swara output. By decoupling the problem into distinct stages, the system ensures better interpretability and flexibility. Each module can be independently improved or replaced without affecting the overall pipeline.

3.2 Vocal Separation

The multi-pitch polyphonic audio signal is separated using a deep learningbased vocal separation model called Demucs [2]. This helps remove the effect of instrument harmonics and enhances the accuracy of the pitch estimates.

3.3 Pitch Extraction

The vocal signal is segmented into short frames, and pitch is estimated for each frame using the CREPE model [1]. Low-confidence frames are filtered out to minimise noise. A median filter is then applied to smooth the pitch sequence and eliminate sudden pitch fluctuations.

Accurate pitch estimation is critical, as errors in this stage propagate to all subsequent steps. The CREPE model is chosen due to its ability to capture fine-grained pitch variations and its robustness across different audio conditions. However, raw pitch estimates may contain noise due to recording artefacts and model uncertainty. Therefore, post-processing techniques such as confidence thresholding and median filtering are applied to enhance stability.

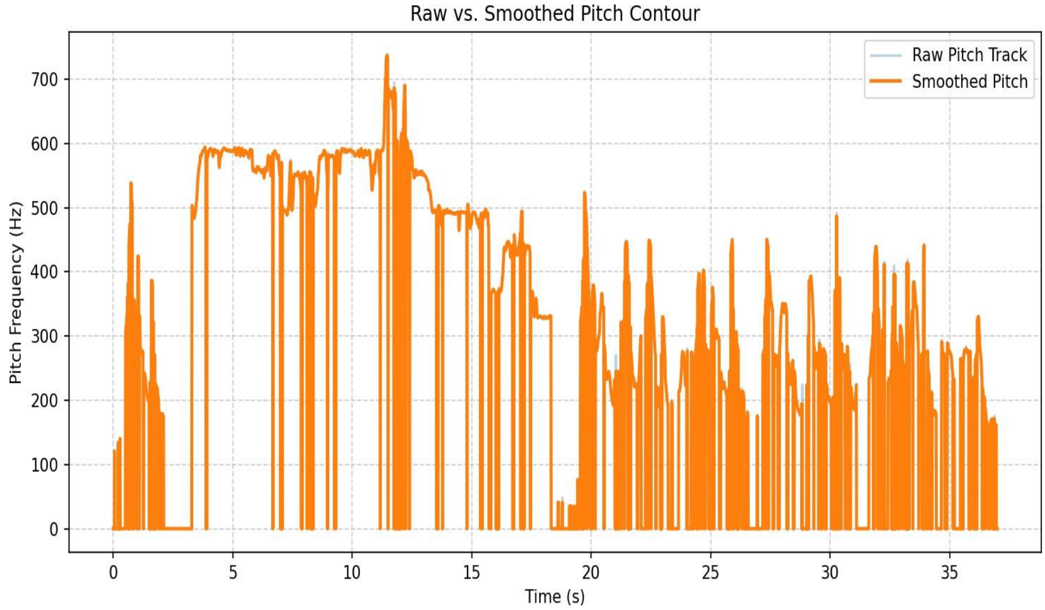

Figure 1: Raw vs. smoothed pitch contour of the vocal signal.

3.4 Pitch-Class Representation

To remove octave ambiguity, pitch values are converted into pitch classes using a logarithmic scale. This representation captures relative pitch independent of absolute frequency and is widely used in tonal music analysis [4].

$$PC = \left(\log_2\left(\frac{f}{f_{ref}}\right)\right) \bmod 1 \qquad (1)$$

where f is the detected pitch, fref is the reference frequency (typically 440 Hz), and $PC \in [0, 1)$ represents the pitch class. This transformation ensures octave invariance by mapping pitches to a single octave.

3.5 Tonic Detection

A pitch-class histogram is constructed using 12 bins corresponding to semitone intervals. The histogram captures the frequency of occurrence of each pitch class:

$$H(i) = \sum_{k=1}^{N} \delta\,(n_k = i) \qquad (2)$$

where n_k represents the semitone index of frame k, and δ is the indicator function.

A major scale template is rotated across all possible tonic candidates. The similarity between the histogram and each template is computed using cosine similarity:

$$S(k) = \frac{\sum_{i=0}^{11} H\,(i) T_k(i)}{\sqrt{\sum_{i=0}^{11} H\,(i)^2}\sqrt{\sum_{i=0}^{11} T_k\,(i)^2}} \qquad (3)$$

The tonic is selected as:

$$k^* = arg\max_{k} S(k) \qquad (4)$$

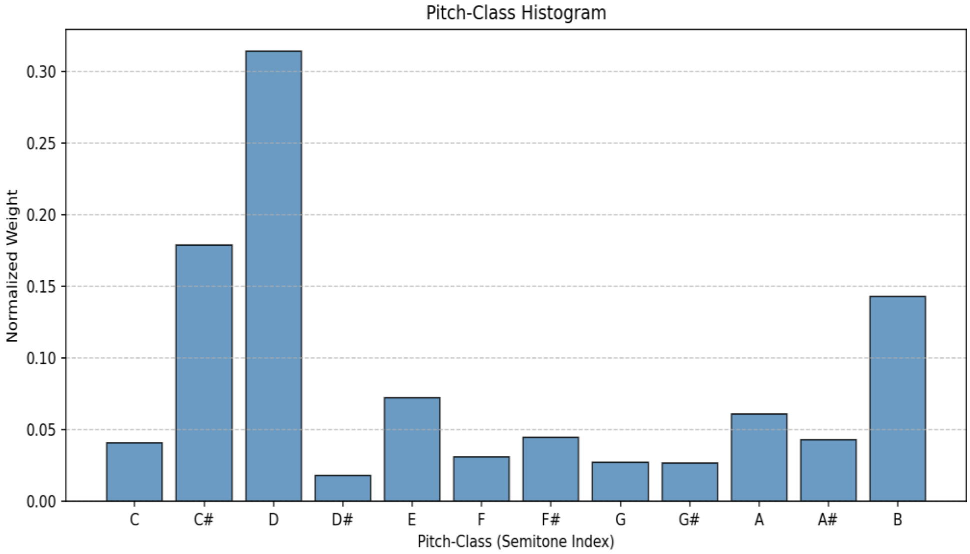

Figure 2: Pitch-class histogram constructed from the vocal pitch track.

This approach is inspired by histogram-based tonic detection methods in Indian music [7]. The use of pitch-class histograms enables a statistical representation of pitch occurrences over time, and is particularly effective in musical contexts where certain notes appear more frequently due to scale constraints.

Table 1: Tonic detection scores across all 12 pitch candidates.

Pitch Candidate	Score
C	0.12
C#	0.08
D	0.15
D#	0.10
E	0.20
F	0.05
F#	0.07
G	**0.30**
G#	0.06
A	0.11
A#	0.04
B	0.09

The selected tonic is highlighted.

3.6 Swara Mapping

Once the tonic is determined, pitch values are mapped to semitone distances relative to the tonic:

$$d = (n - k^*) \mod 12 \qquad (5)$$

These semitone values are mapped to Carnatic swaras as follows: d = 0: Sa; d = 1: Ri_1; d = 2: Ri_2/Ga_1; d = 3: Ga_2; d = 4: Ga_3; d = 5: Ma_1; d = 6: Ma_2; d = 7: Pa; d = 8: Dha_1; d = 9: Dha_2/Ni_1; d = 10: Ni_2; d = 11: Ni_3.

Mapping pitch values to discrete swaras involves quantisation, which may introduce minor deviations from the original pitch contour. However, this step is necessary to convert continuous audio signals into symbolic representations. The mapping is designed to preserve the relative structure of the scale while maintaining consistency across different tonic values.

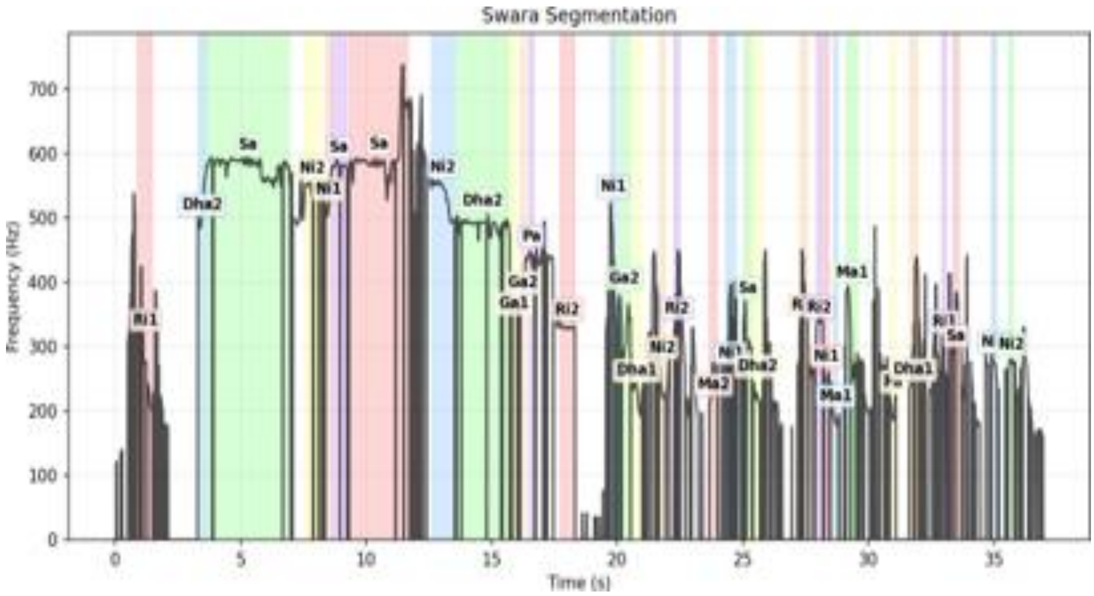

Figure 3: Swara segmentation output overlaid on the pitch contour

3.7 Temporal Smoothing

Instead of assigning swaras at the frame level, consecutive frames with stable pitch are grouped into segments using pitch continuity constraints:

$$|f_{t+1} - f_t| < \epsilon \tag{6}$$

Each segment is assigned a single swara, resulting in smoother and more musically meaningful transcription.

IV. RESULTS AND DISCUSSION

4.1 Evaluation Metrics

Tonic accuracy is defined as:

$$Accuracy = \frac{Correct\ tonic\ predictions}{Total\ samples} \tag{7}$$

Swara accuracy is defined as:

$$Swara\ Accuracy = \frac{Correct\ swara\ labels}{Total\ labels} \tag{8}$$

4.2 Observations

The experimental results demonstrate that the proposed system is capable of accurately detecting the tonic and generating consistent swara sequences for audio signals that follow a major scale structure. The use of pitch-class histograms significantly improves the robustness of tonic detection by reducing the influence of octave variations.

The temporal smoothing mechanism plays a crucial role in enhancing the quality of the output by reducing abrupt transitions between swaras. Without smoothing, frame-level predictions often exhibit high variability due to minor pitch fluctuations. By grouping stable pitch regions, the system produces more musically meaningful outputs.

It is also observed that the performance of the system is highly dependent on the quality of the vocal separation stage. Clean separation results in more accurate pitch estimation, which directly improves swara mapping accuracy.

4.3 Sample Output

Table 2: Sample swara transcription output

Time Interval (s)	Swara
0.00–0.50	Sa
0.50–1.00	Ri2
1.00–1.40	Ga3

4.4 Limitations

Despite its effectiveness, the proposed system has several limitations. First, it assumes that the input audio follows a major scale structure, which restricts its applicability to a subset of Carnatic music. Complex ragas with non-linear note progressions are not adequately handled by the current approach. Second, the system simplifies continuous pitch variations and gamakas into discrete swara labels. While this improves interpretability, it results in the loss of fine musical nuances that are essential in Carnatic music. Third, the accuracy of the system is dependent on the quality of the input audio and the effectiveness of the vocal separation model. Noisy recordings or poor separation can lead to incorrect pitch estimates and subsequently affect tonic detection and swara mapping. Finally, the system operates offline and does not support real-time transcription, which limits its usability in live performance scenarios.

4.5 Conclusion

This paper presents a structured approach for automatic transcription of Carnatic swaras from polyphonic audio signals. By integrating deep learning-based pitch estimation, octave-invariant pitch representation, and histogram-based tonic detection, the system achieves reliable swara mapping for major scale-based compositions. The inclusion of temporal smoothing further enhances the musical quality of the output. The proposed method demonstrates that combining signal processing techniques with statistical analysis can effectively address key challenges in Carnatic music transcription. The system provides a foundation for future research in symbolic music representation and automated analysis

V. Acknowledgment

The authors would like to express their sincere gratitude to **Dr. S. Senthamizh Selvi**, Associate Professor, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, for her invaluable guidance, continuous support, and constructive suggestions throughout this research. The authors also thank the Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, for providing the necessary facilities and support to carry out this work.

References

[1] Kim, J.W., Salamon, J., Li, P., Bello, J.P.: CREPE: A convolutional representation for pitch estimation. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 161–165 (2018)

[2] D´efossez, A., Usunier, N., Bottou, L., Bach, F.: Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254 (2019)

[3] McFee, B., et al.: librosa: Audio and music signal analysis in Python. In: Proc. 14th Python in Science Conf., pp. 18–25 (2015)

[4] M¨uller, M.: Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Springer, Berlin (2015)

[5] Temperley, D.: Music and Probability. MIT Press, Cambridge, MA (2007)

[6] Chordia, P., Rae, A.: Raag recognition using pitch-class distributions. In: Proc. Int. Society for Music Information Retrieval Conf. (ISMIR) (2007)

[7] Gulati, S., et al.: Tonic identification in Indian art music using pitch histograms. In: Proc. ISMIR (2014)

[8] Ganguli, K.K., Rao, P.: On the perception of raga in Carnatic music. Journal of New Music Research (2009)

[9] B¨ock, S., Schedl, M.: Enhanced beat tracking. In: Proc. Digital Audio Effects (DAFx) (2011)

[10] G´omez, E.: Tonal description of polyphonic audio. INFORMS Journal on Computing (2006)