ShrutiSense: Microtonal Modeling and Correction in Indian Classical Music

Jayanth Athipatla*, Rajarshi Ghosh*
(Equal Contribution)

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Indian classical music represents on of the world's most sophisticated microtonal systems, employing 22 distinct pitch intervals called Shrutis within each octave. Unlike Western music's 12-tone equal temperament, this system provides finer granular control over pitch relationships, enabling the subtle melodic ornamentations and emotional expressions that characterize ragas. This concept of Shruti, first formalized in Bharata's Natya Shastra, divides the octave into 22 mathematically precise intervals, each serving specific melodic and emotional functions with different raga contexts.

Mathematical Foundations

We formalize the 22-Shrut system as a logarithmic frequency division of the octave, following the theoretical framework established by Bhatkhande, with cent values. We then model each raga as a directed graph G = (S, T), where S is the set of Shruti positions and T which a subset of S X S represents permissible transitions between Shrutis. The Grammar-Constrained Shruti Hidden Markov Model (GC-SHMM) is defined over a state space S, where N represents the number of active Shrutis in a give raga. Emission probabilities are modeled using Gaussian distributions. A key feature of GC-SHMM is its enforcement of raga grammar through constrained transitions, and it uses Viterbi algorithm to identify the most probable Shruti sequence. The Shruti-aware Finite-State Transducer (FST), maps input pitch sequences to corrected or completed outputs by applying weighted edit operations. The GC-SHMM algorithm operates with a time complexity of O(TN^2), where T denotes the sequence of length and N the number of active states, and a space complexity of O(TN). In contrast, the Finite-State Transducer (FST) approach exhibits a time complexity of O(TM^2) with lattice size M.

Limitations

Despite promising results, current limitations include reliance on pre-defined raga grammars, assumption of monophonic input, limited modeling of ornamental nuances, and the requirement for tonic identification. Future work will focus on adaptive raga learning through unsupervised corpus analysis, end-to-end audio integration with robust pitch estimation for noisy inputs, expanded ornament modeling that treats gamakas and other microtonal inflections as first-class entities, and multi-voice extensions to handle drones and polyphonic textures common in Indian classical ensembles. We also envision cross-cultural adaptation to diverse microtonal traditions such as Carnatic music, Middle Eastern maqam systems, and contemporary non-Western composition practices.

Experimental Evaluation and Results

Correction Task Performance

Table 1: Pitch Correction Performance Comparison (Yaman Raga, Corruption = 0.4)

Method	Shruti Acc. (%)	Mean Error (cents)	Time (ms)
GC-SHMM	**84.0 ± 0.4**	**107.6 ± 3.6**	**12.5 ± 0.3**
Shruti FST	**91.3 ± 0.2**	**45.6 ± 1.4**	**0.1**
Nearest Cent	89.4 ± 0.3	51.8 ± 1.4	0.1
Random	12.6 ± 0.3	452.6 ± 2.4	0.0

The statistical analysis reveals strong and significant differences in accuracy across the models evaluated. Pairwise comparisons show that both FST and Nearest Cent outperform HMM, while all structured models far exceed the random baseline. The Cohen's d values outputted by the evaluation code indicate large effect sizes for each contrast, especially against random, with FST showing the greatest improvement. A one-way ANOVA further confirms significant disparities among models with an exceptionally high F-statistic. Overall, the findings suggest that FST is the most accurate model, followed closely by nearest cent, while HMM lags behind but still vastly outperforms random. Thus, when users choose to use ShrutiSense to correct an audio file, the audio-file will go through the FST pipeline.

Completion Task Performance

Table 2: Melodic Completion Performance by Missing Pattern

Missing Pattern	HMM		FST	
	Acc.	Error (cents)	Acc.	Error (cents)
Random	57.1	203.0	62.6	158.7
Clustered	40.3	344.9	26.6	317.0
Structured	82.9	48.5	70.5	228.1

As the completion task is not generally a real-world use case, we deemed it unnecessary to compare it against baseline models. The statistics for the completion task are displayed in Figure 2. Clearly, the correction task is much easier for ShrutiSense than the completion task, which is expected. Overall, the HMM performed better with a mean accuracy of 60.1 ± 30.9% vs the FST which had a mean accuracy of 48.6 ± 22.2%. That being said, the FST actually beat the HMM for the Bhairava (0.2 corruption), the Bilawal (0.2 corruption), and the Khamaaj (0.2 corruption), showing that the FST actually does as good if not better than the HMM when corruption is low. Additionally, the completion error distribution was much less spread out for the FST than the HMM. The FST was significantly faster than the HMM, as expected because of the differences in algorithmic complexity.

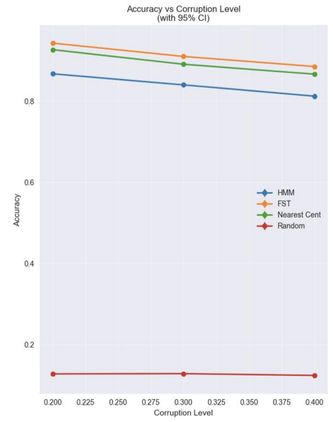

Figure 1: Correction Statistic: Accuracy vs. Corruption Level

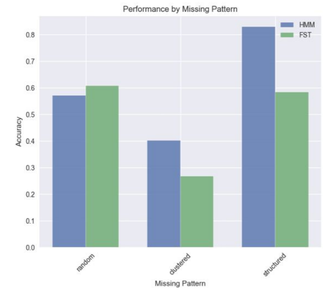

Figure 2: Completion Statistic: Performance by Missing Pattern

Robustness Analysis

Noise resilience testing demonstrates that ShrutiSense (FST model) maintains an average accuracy of 91.3% with input noise up to ±50 cents, as introduced by quantization corruption. Example sequences at corruption levels of 0.2 to 0.4 show FST accuracies ranging from 86.7% to 90.0%. Performance degrades gracefully at higher corruption levels, maintaining accuracy around 86.7% at 0.4 corruption. The system exhibits consistent performance across ragas, with minor variations: Yaman (91.1% accuracy), Bhairavi (90.7%), Bilaval (91.2%), Kalyan (91.8%), and Khamaaj (91.8%). These results, derived from 900 simulations across sequence lengths of 30, 50, and 100, validate the generalizability of the grammar-based approach.

References

S. Adhikary, M. S. M, S. S. K, S. Bhat, and K. P. L. Automatic music generation of indian classical music based on raga. In Proceedings of the IEEE International Conference for Convergence in Technology (I2CT), 2023.

V.N. Bhatkhande. A Comparative Study of Some of the Leading Music Systems of the 15th, 16th, 17th and 18th Centuries. Government Central Press, 1934.

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111(4):1917–1930, 2002.

N. Devis, N. Demerlé, S. Nabi, D. Genova, and P. Esling. Continuous descriptor-based control for deep audio synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In IEEE Workshop on Neural Networks for Signal Processing, pages 747–756, 2002.