tcs Research
Inventing for impact

IIT MADRAS
Indian Institute of Technology Madras

Automatic Descriptive Transcription of Carnatic Music

Presented at his Ph.D. Defense by

Venkata Subramanian Viraraghavan
(EE16D024 at IITM, ERP from TCS)
Guides:
Prof. R Aravind (EE, IITM),
Prof. Hema Murthy (CSE, IITM),
Dr. Arpan Pal (TCS)

14-Dec-2022

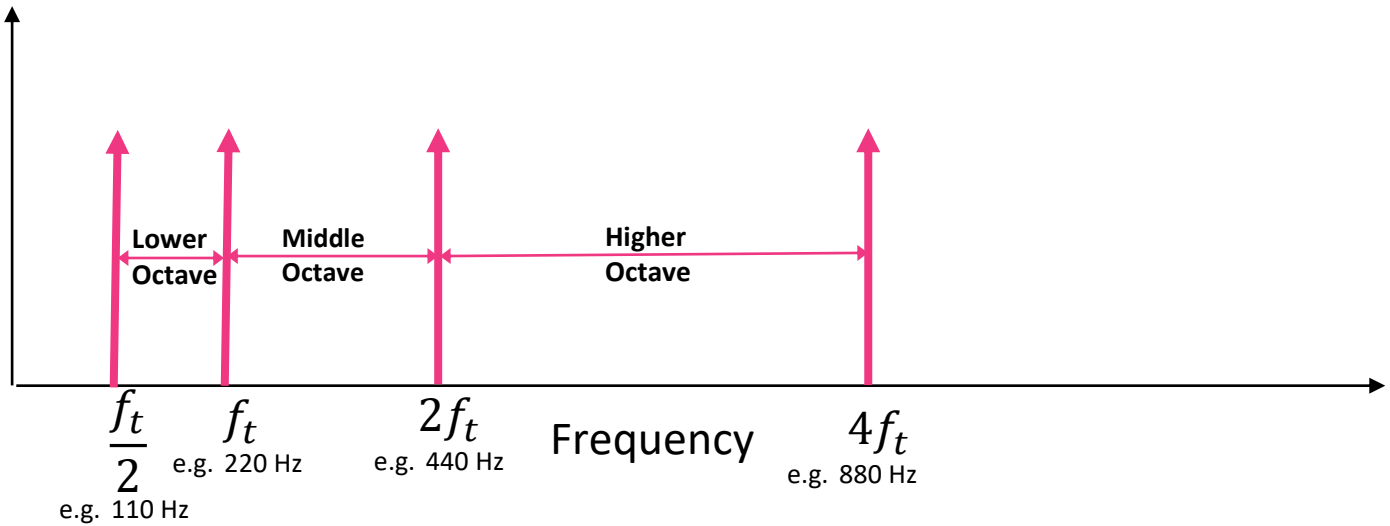
Contents

IIT MADRAS
Indian Institute of Technology Madras

- Introduction
  - Pitch, Tonic, Octave, Notes
  - Challenges in Carnatic Music Transcription
- Contributions
  - Definitions and Systematic Study of:
    - Constant-pitch notes (CPNs), Transients, & Stationary points (STAs)
  - Definitions, and Detecting Targets for:
    - Upward and Downward anchors, and Max-STAs and Min-STAs
  - Measuring Precision of CPNs and STAs
  - State-based Transcription (SBT)
    - Using Anchor-wise STA-targets and State model
- Conclusions and Future Work

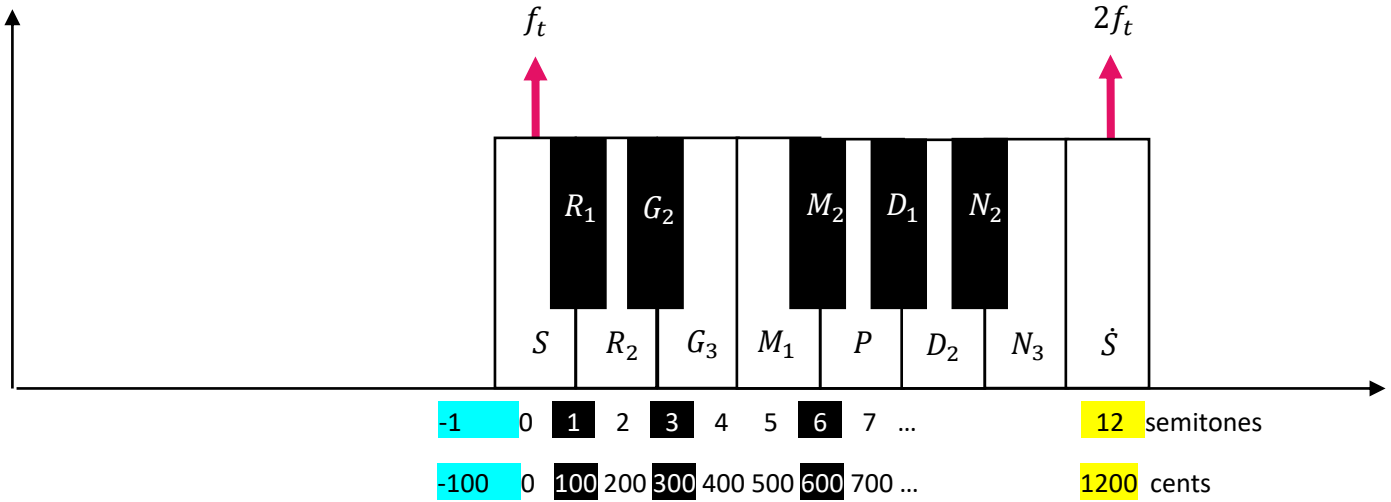
Pitch, Tonic, and Octave

- We use ‘pitch’ to signify ‘measured fundamental frequency’
- All melodies in Indian music are aligned to a *tonic note*
- If f_t is the pitch of the tonic note in Hz, range of pitch values $[f_t, 2f_t)$ is an **octave**



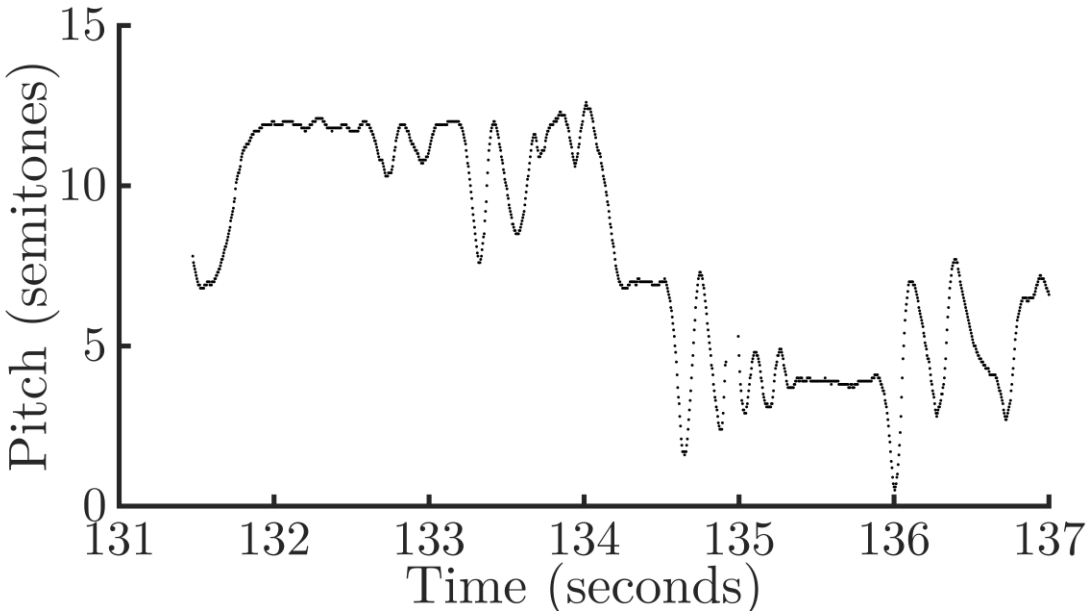
Twelve notes per octave

- Twelve notes per octave are used in several music systems
- Pitch f in Hz is converted to semitones $s = 12 \log_2 \frac{f}{f_t}$, for a *tonic* note of pitch f_t Hz



Carnatic Music (CM) Pitch curve example

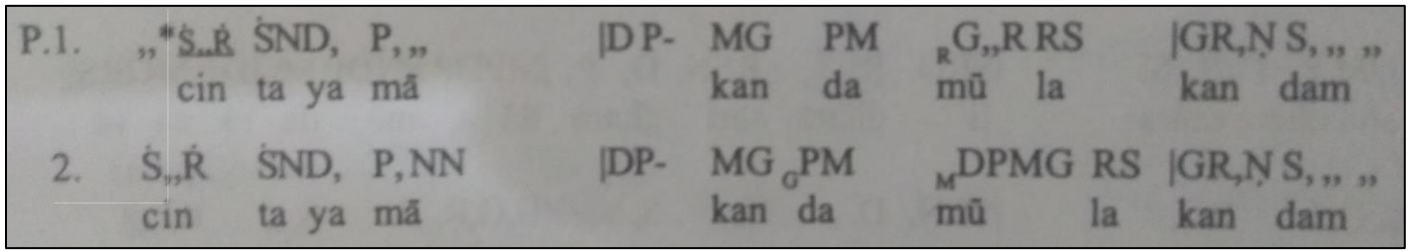
- CM uses *gamakas*, which are continuous movements of pitch between notes
- Pitch is tracked as a function of time (every $T_0 \approx 4.44$ ms), $s[n] = 1200 \log_2 \frac{f[n]}{f_t}$ cents



Challenges in Transcription of *gamakas*

- *Gamakas* are mandatory in CM
- … but are not captured in extant notation

T K Govinda Rao



- Synthesizing[1] by following the notation faithfully is clearly insufficient
- Challenge: Detailed transcription of *gamakas* is complex

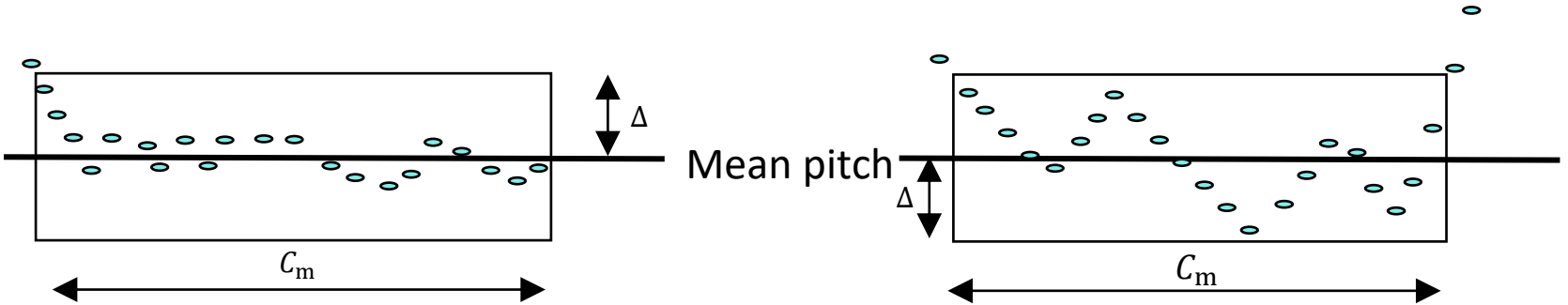
1. Kaustuv Kanti Ganguli and Preeti Rao, “Discrimination of melodic patterns in Indian classical music,” NCC 2015

Components of a CM pitch curve

- Our observation is that pitch curves in CM can be thought of as being made up of
  - **Constant-Pitch Note**s (or **CPN**s) and **Transient**s.
- Since the extent of *gamaka*/continuous pitch movements is important, we focus on
  - **Stationary Point**s (**STA**s), which are the maxima and minima of transients

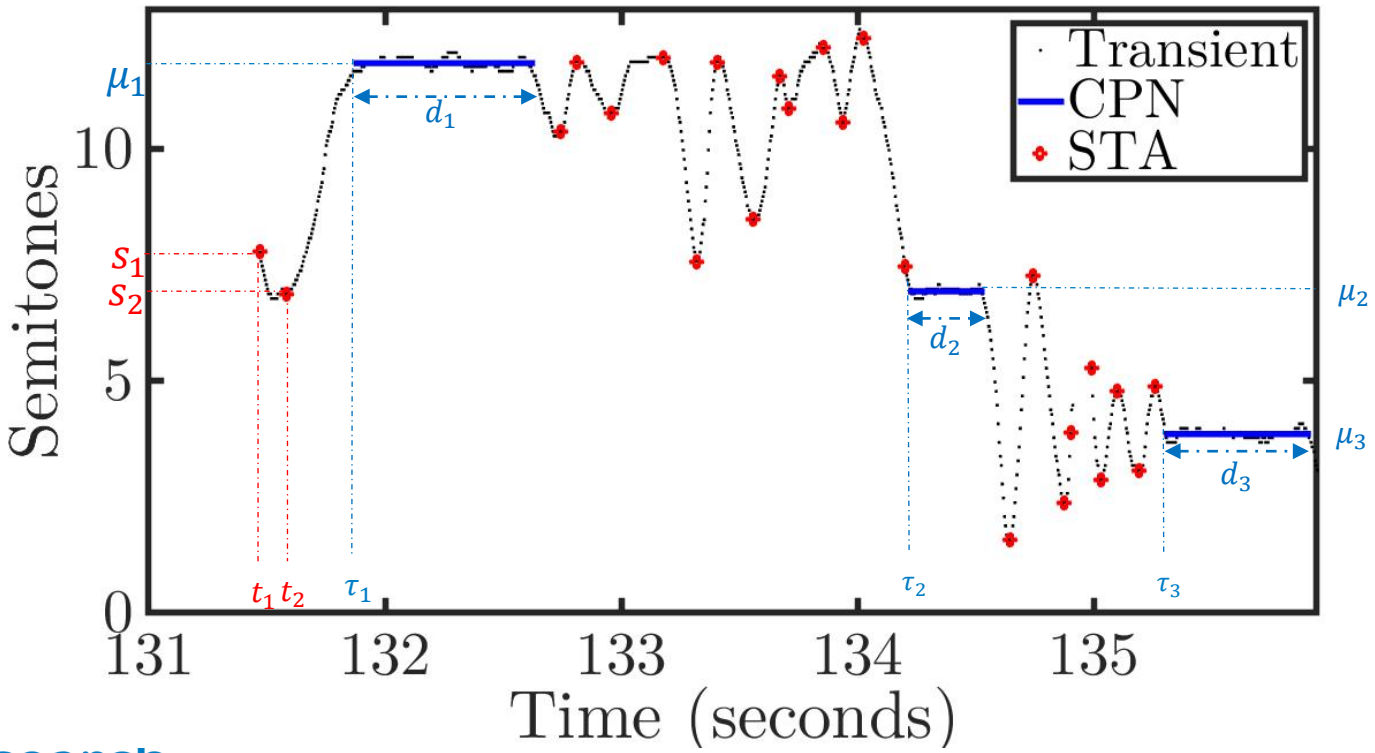
Definitions: CPN, transient, STA

- **Constant-Pitch Note (CPN):** Segment whose pitch does not vary from its mean pitch by more than $\Delta = 35$ cents and lasts for at least $C_{\mathrm{m}} = 80$ ms

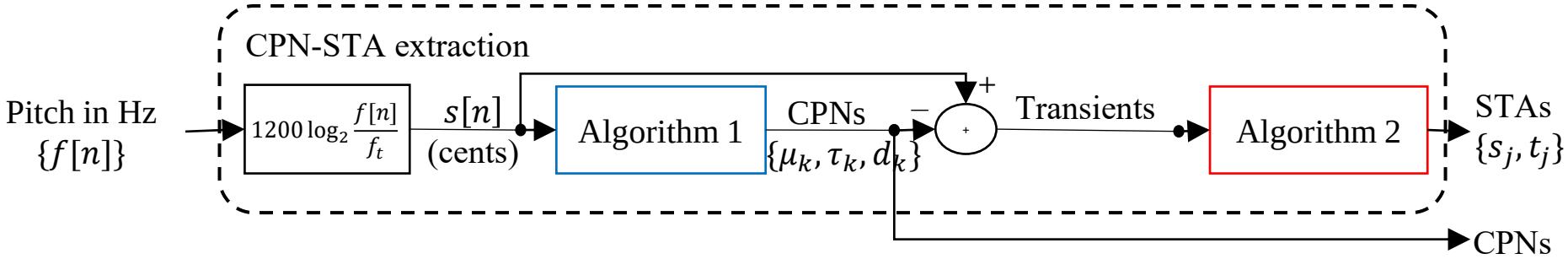


- **Transient**: Any pitch curve outside CPNs.
- **Stationary point (STA)**: Local maximum or minimum of transients

Extracted CPNs and STAs: Example



Extraction of CPNs, transients and STAs



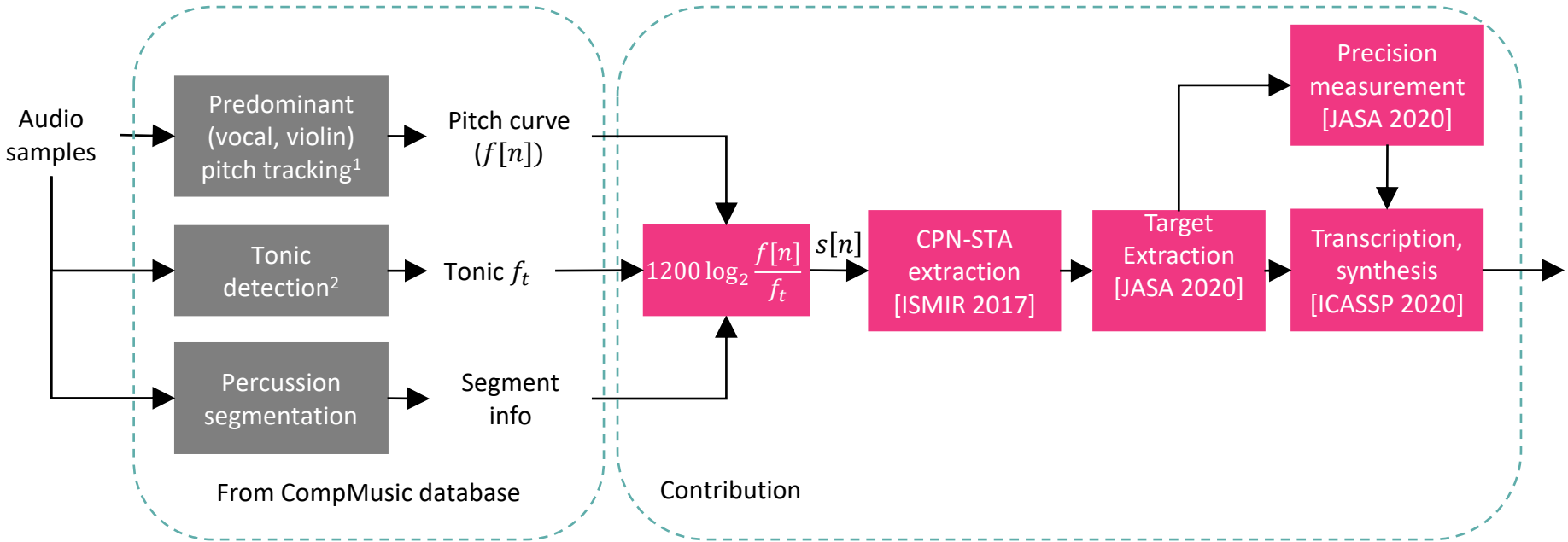
Visualization by analogy to projectile motion

IIT MADRAS
Indian Institute of Technology Madras

- CPNs and STAs are mapped to ledges and reflectors[1]

1. Visualization was done with the assistance of Rahul Gavas, TCS; presented at SMM19, Vienna

CompMusic: Dunya Corpus, Carnatic subset ("database")



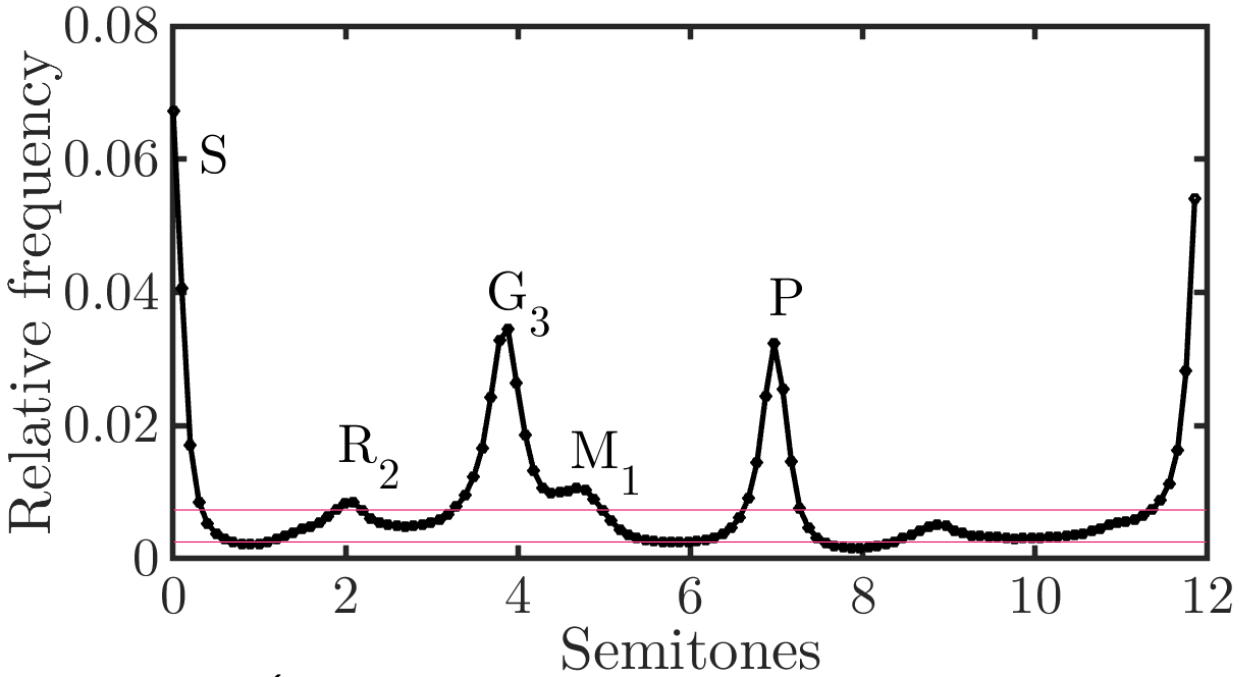
Database has 480 concert renditions in 40 *rāga*s by 64 professional musicians

- More than 200 minutes for each of 7 major ragas
- Nearly 100,000 CPNs and millions of STAs

1. Salamon and Gomez, IEEE TASLP 2015
2. Salamon, Gulati and Serra, ISMIR 2012

Histograms of pitch-values

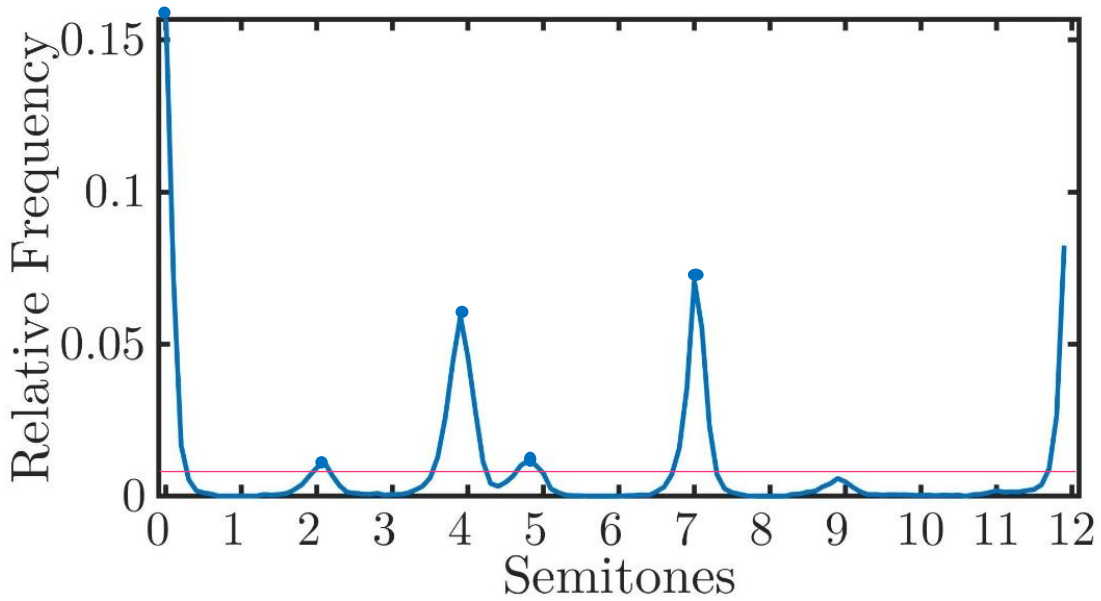
- Histogram counts the number of occurrences of pitch-values
- Wrapped around(or folded) to one octave; 120 bins of size 10 cents each



Rāga: *Śankarābharaṇam* (entire pitch curve in 12 renditions)

Detecting Raga-specific CPN Targets

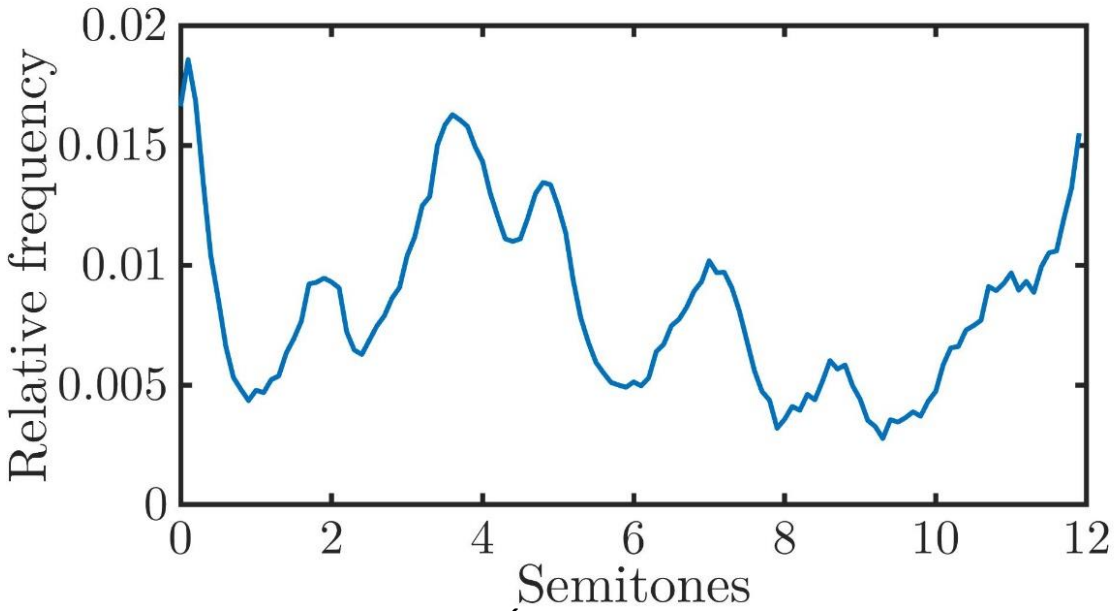
- For CPNs, detect peaks from the histogram of their mean pitch values $\{\mu_k\}$



Rāga: *Śankarābharaṇam*

Histograms of STAs

- Some peaks are visible, but are wider than in the CPN histogram
- Inconclusive about possible ‘hidden peaks’ between the visible peaks

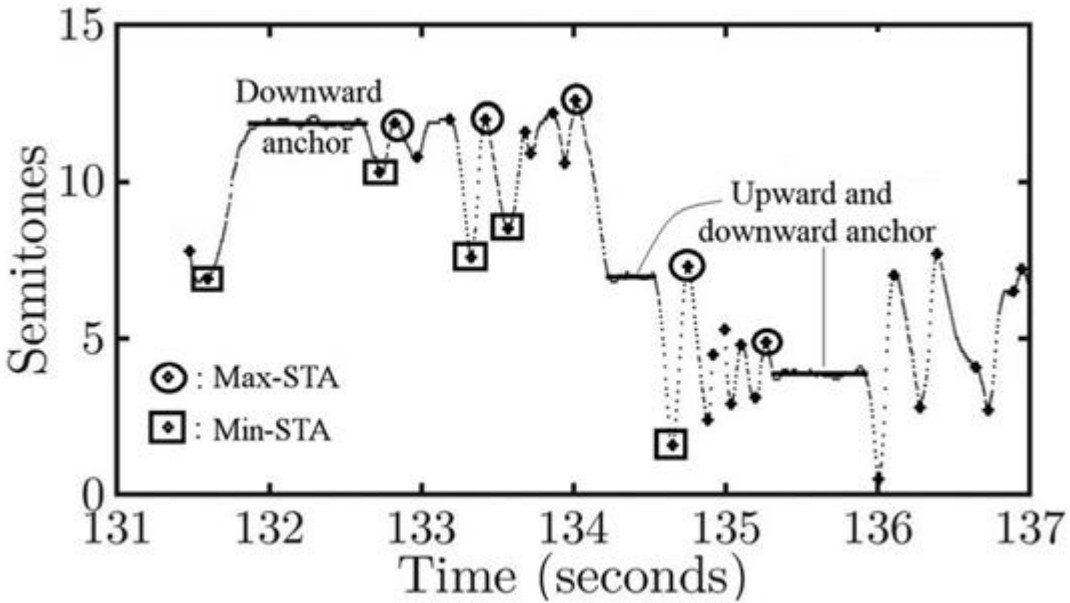


Rāga: *Śankarābharaṇam*

Anchors, Max-STAs, and Min-STAs

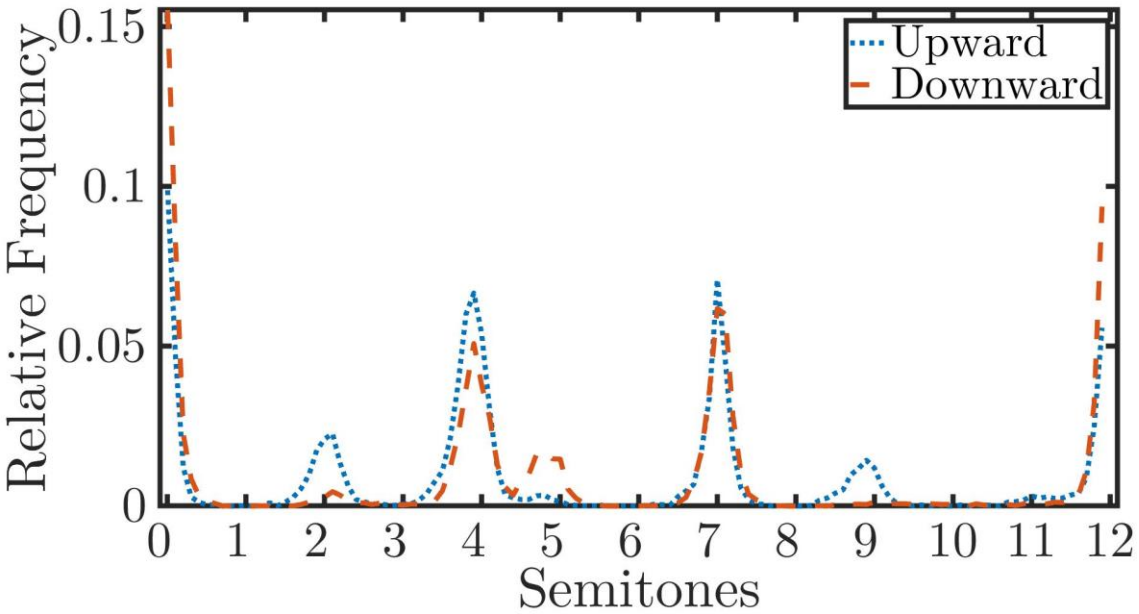
- **Anchor:** A CPN adjacent to a STA.
- An anchor can be an **upward anchor**, or a **downward anchor**, or *both*
- Each STA is *either* a **Max-STA** (local maximum) *or* a **Min-STA** (local minimum)

(Not all min-STAs & max-STAs are marked in the figure)

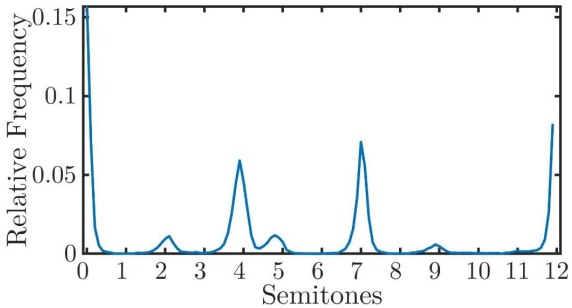


Histograms of Upward and Downward Anchors

- Peaks are as sharp as in the histogram of CPNs (reproduced on the right)
- Data-driven approach, does not use any specification in textbooks



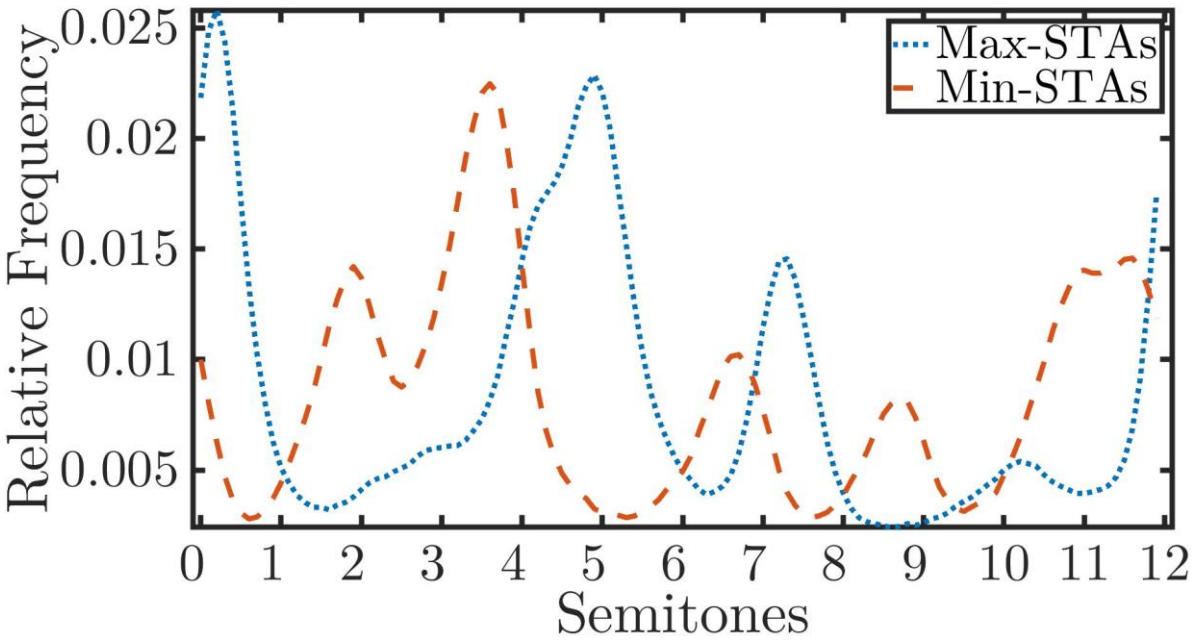
Compare: Histogram of CPNs



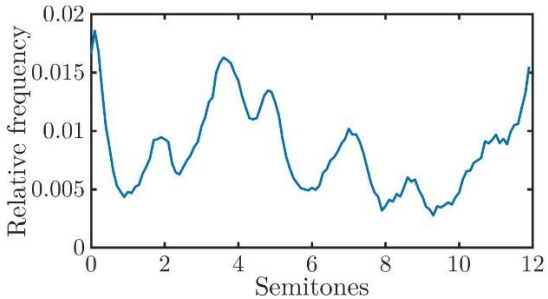
Rāga: *Śankarābharaṇam*

Detecting targets from histograms of Min-STAs and Max-STAs

- Peaks are sharper than in the histogram of all STAs, but not as sharp as for anchors
- Data-driven approach, does not use any specification in textbooks



Compare: Histogram of all STAs



Rāga: *Śankarābharaṇam*

Detecting targets for Max-STAs and Min-STAs

- Peaks are found one at a time in decreasing order of height
  - Locations of tall peaks (0.3 × max value) are always considered as targets.
  - Other peaks (0.15 × max value) farther than 110 cents from already detected targets are considered as targets
- Histogram is updated by removing detected peaks to detect smaller overlapping peaks
- Iterations halt the updated histogram has less than 10% of the original data.

Thresholds are chosen empirically, but outputs are relatively insensitive to them.

Detected Targets: Examples

IIT MADRAS
Indian Institute of Technology Madras

Śankarābharaṇam

- $\boldsymbol{\mathcal{A}_u} = \{0, 210, 390, 700, 890\}$ cents, which map to $\{S, R_2, G_3, P, D_2\}$;
- $\boldsymbol{\mathcal{A}_d} = \{0, 390, 480, 700\}$ cents, mapped to $\{S, G_3, M_1, P\}$.
- $\boldsymbol{S_{\max}} = \{20, \mathbf{250}, 370, 490, 730, \mathbf{1020}\}$ cents, mapped to $\{S, \mathbf{G_2}, G_3, M_1, P, \mathbf{N_2}\}$;
- $\boldsymbol{S_{\min}} = \{1160, 190, 360, 670, 870, \mathbf{1060}\}$ cents, mapped to $\{S, R_2, G_3, P, D_2, \mathbf{N_3}\}$

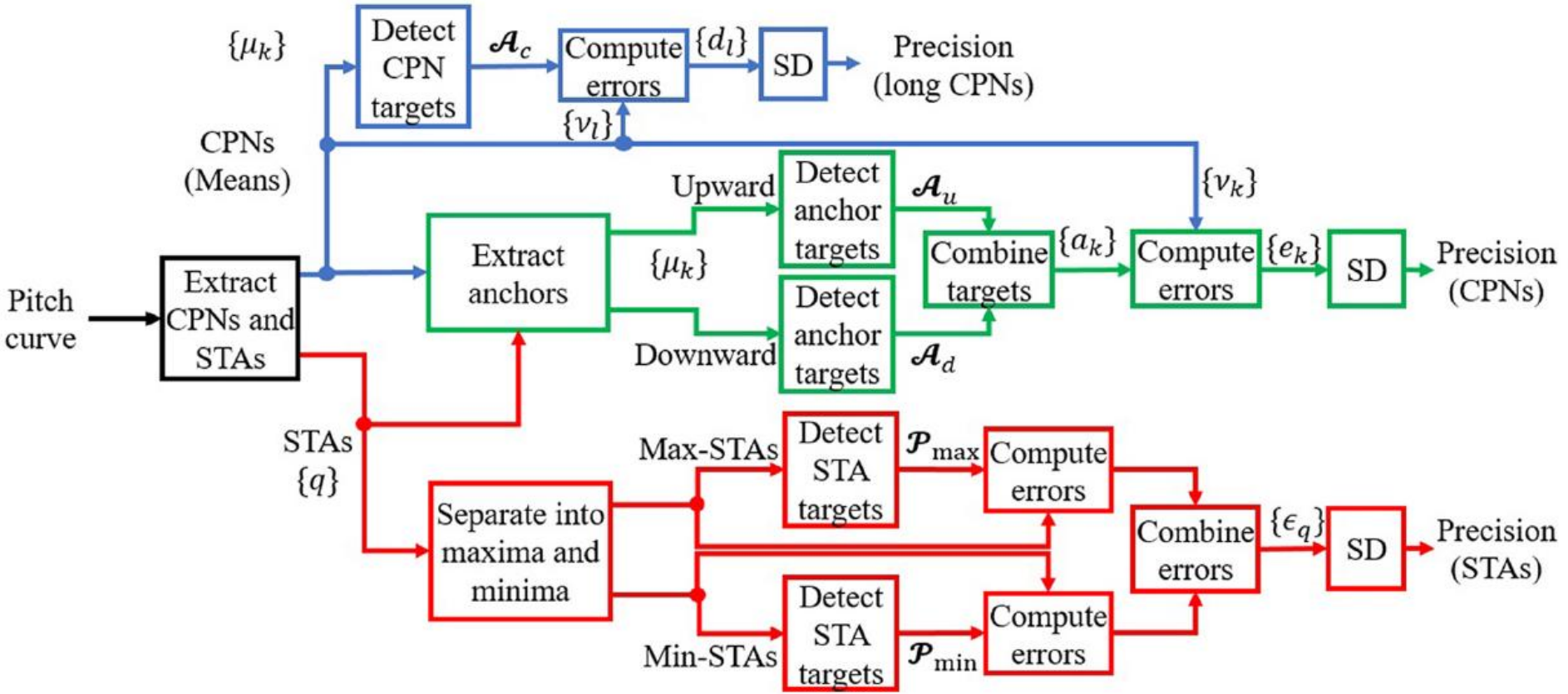
Tōḍī

- $\boldsymbol{\mathcal{A}_u} = \{0, 100, 180, 500, 700, 790, 880\}$ cents, which map to $\{S, R_1, R_2, M_1, P, D_1, D_2\}$;
- $\boldsymbol{\mathcal{A}_d} = \{0, 90, 500, 700, 790\}$ cents, mapped $\{S, R_1, M_1, P, D_1\}$.
- $\boldsymbol{S_{\max}} = \{20, 130, \mathbf{300}, 520, 700, 820, 940\}$ cents, mapped to $\{S, R_1, \mathbf{G_2}, M_1, P, D_1, D_2\}$;
- $\boldsymbol{S_{\min}} = \{1180, 170, 460, 690, 890\}$ cents, mapped to $\{S, R_2, M_1, P, D_2\}$

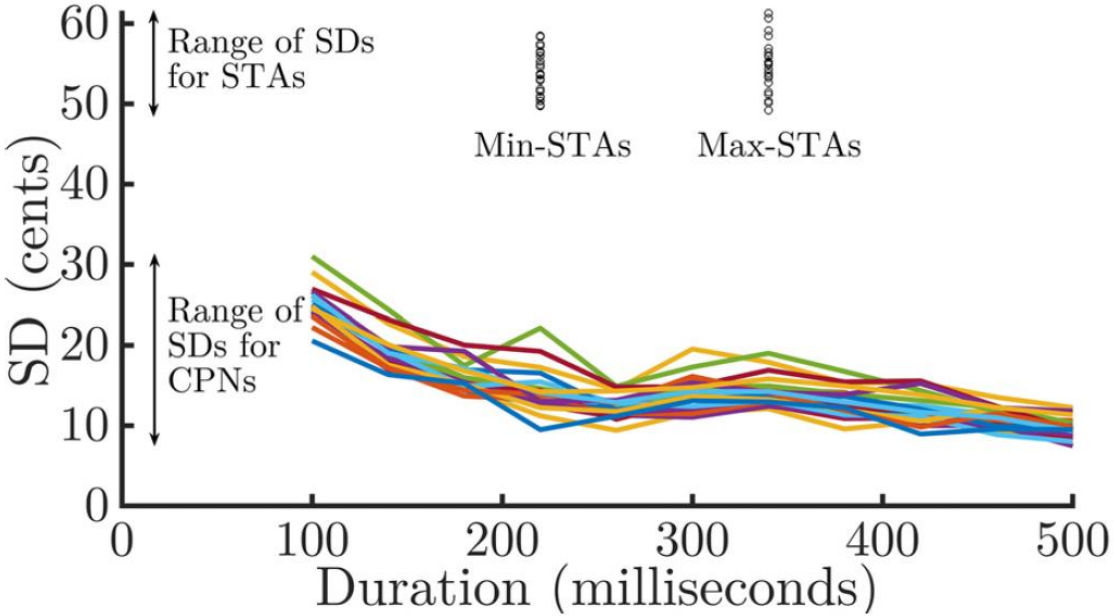
Precision Measurement

- Detected targets are quantized to integer semitones
- Precision is the standard deviation (SD) of errors with respect to the nearest target
- Measured precision guides transcription
- Precision is measured separately for CPNs, Min-STAs, and Max-STAs
- Errors with respect to targets are computed for each *rāga* separately

Precision Measurement Block Diagram



Measured Precision



Precision of professional musicians is in a range of:

~8 to ~15 cents for long CPNs

~50 to ~65 cents for STAs

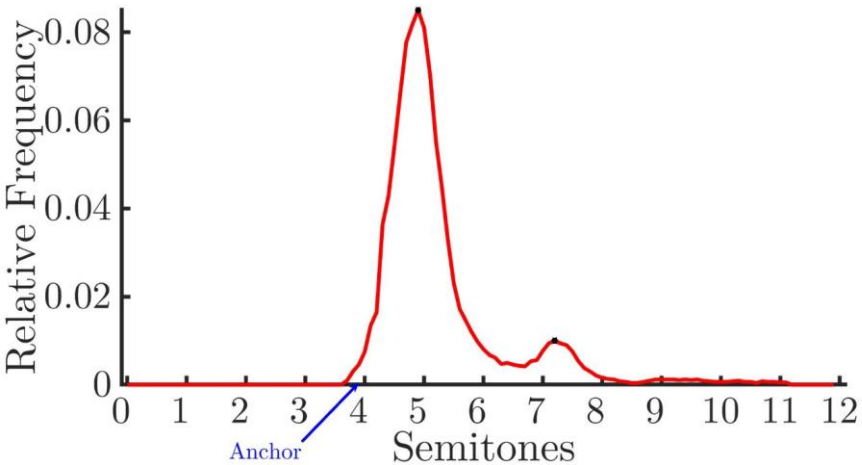
Descriptive Transcription using CPNs and STAs

IIT MADRAS
Indian Institute of Technology Madras

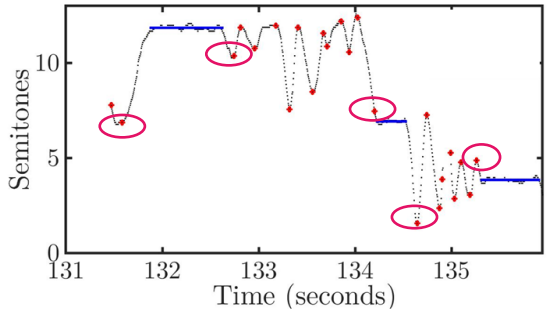
- Measured precision suggests that
  - CPNs may be quantized uniformly
  - Quantizing STAs is non-trivial
- CPNs and STAs constitute a descriptive transcription
  - Quantized pitch, location and duration of each CPN
  - Quantized pitch and location of each STA
  - Other points in a transient need not be quantized or even characterized

Anchor-specific Targets

- For each anchor $a \in \widetilde{\mathbf{A}}_u$ or $\widetilde{\mathbf{A}}_d$, obtain set of STA targets ($\tilde{\mathbf{S}}_{a,\max}, \tilde{\mathbf{S}}_{a,\min}$) from histograms of STAs adjacent to all instances of a

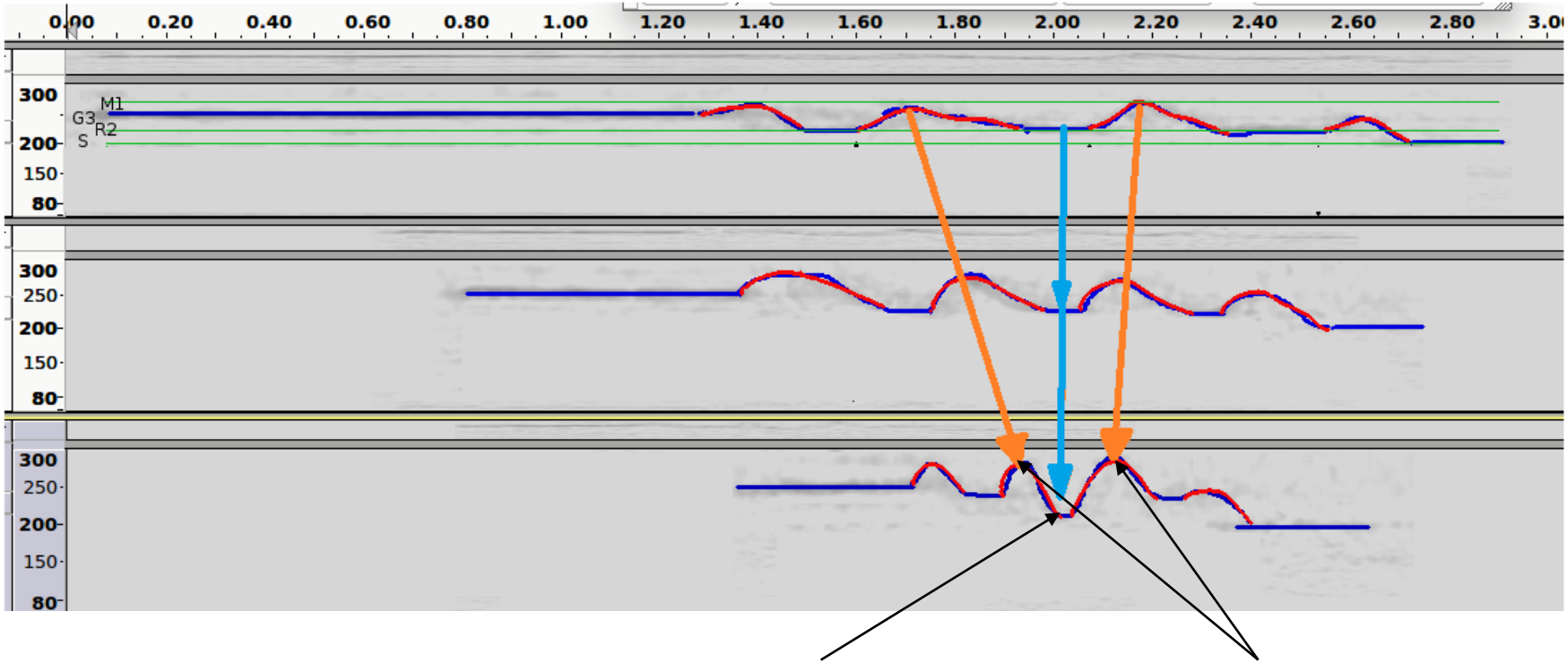


E.g., a = G_3 in *Śankarābharaṇam*



$\widetilde{\mathbf{A}}_u$	$\widetilde{\mathbf{S}}_{a,\max}$		$\widetilde{\mathbf{A}}_d$	$\widetilde{\mathbf{S}}_{a,\min}$	
S	G_2	M_1	$\dot{S}$	N_3	D_2
R_2	G_2	M_1	-		
G_3	M_1	P	G_3	R_2	
-			M_1	G_3	R_2
P	N_2	$\dot{S}$	P	G_3	R_2
D_2	N_2	$\dot{S}$	-		

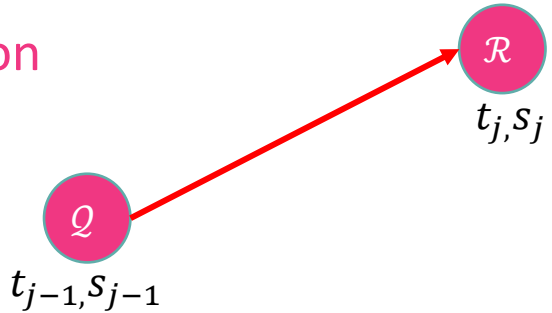
Introducing states based on properties of CPNs and STAs



STA in anchor state
$\mathcal{A}_t$ or ${}_t\mathcal{A}$

STA in transient state
${}_a\mathcal{T}$ or $\mathcal{T}_a$

Mathematical Formulation

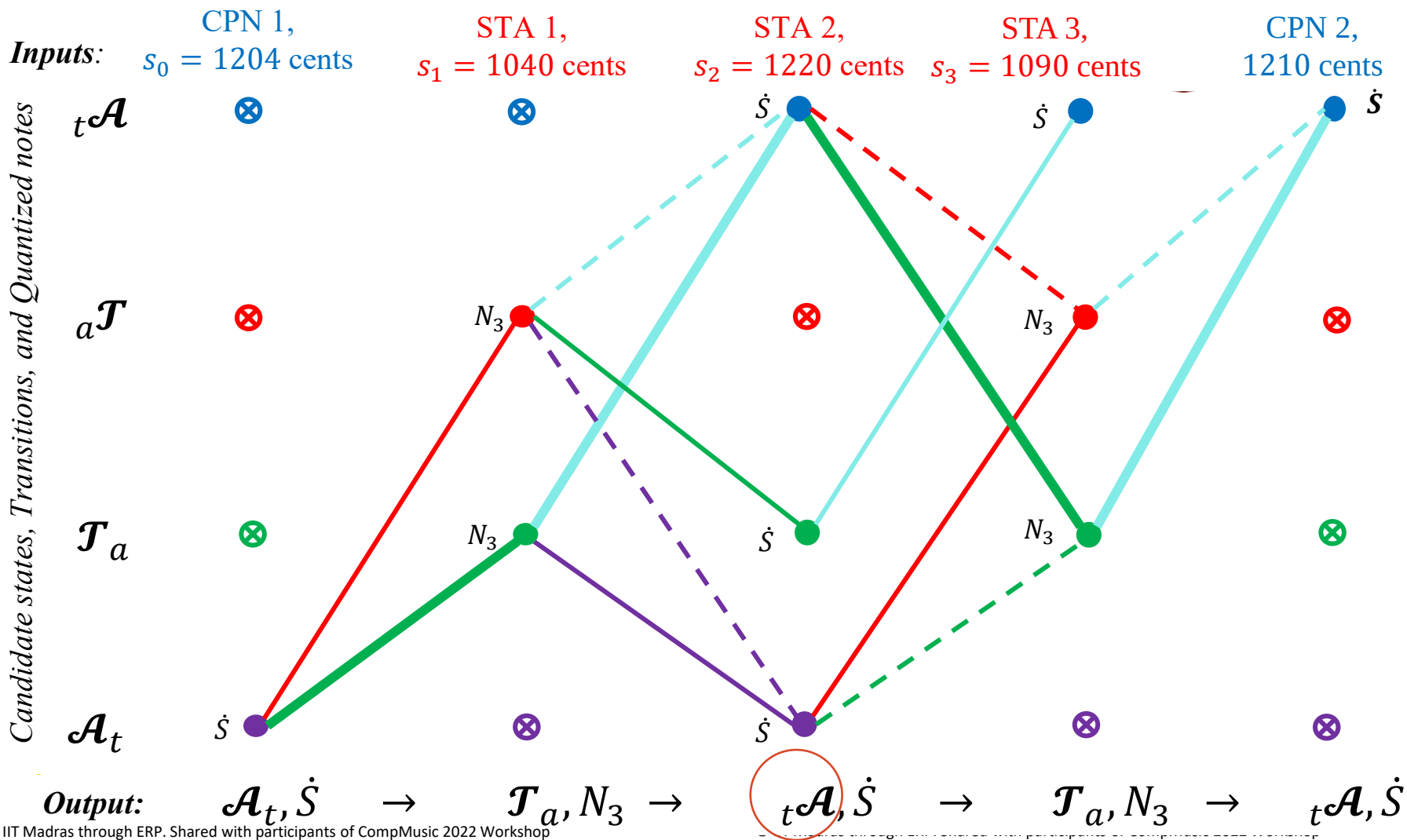


- Two basic transitions
  - Anchor $\boldsymbol{\mathcal{A}}_t$ to STA ${}_a\boldsymbol{\mathcal{T}}$: $w_{j-1} = a \in \tilde{\mathbf{A}}_u$ and $w_j \in \tilde{\mathbf{S}}_{a,\max}$
  - STA $\boldsymbol{\mathcal{T}}_a$ to anchor ${}_t\boldsymbol{\mathcal{A}}$: $w_j = a \in \tilde{\mathbf{A}}_d$ and $w_{j-1} \in \tilde{\mathbf{S}}_{a,\min}$
- Maximize probability of state transition $\mathcal{Q} \to \mathcal{R}$ given s_{j-1}, s_j

$$P\left(\mathcal{Q} \to \mathcal{R} \middle| s_{j-1}, s_j\right) = \frac{f\left(s_{j-1}, s_j \middle| \mathcal{Q} \to \mathcal{R}\right) P(\mathcal{Q} \to \mathcal{R})}{f\left(s_{j-1}, s_j\right)}$$

- Naïve Bayes formulation: Consider only the likelihood $f\left(s_{j-1}, s_j \middle| \mathcal{Q} \to \mathcal{R}\right)$
- Model s_{j-1} and s_j as independently, normally distributed around targets w_{j-1}, w_j:

$$f\left(s_{j-1}, s_j \middle| \mathcal{Q} \to \mathcal{R}\right) = \mathcal{N}\left(s_{j-1}, w_{j-1}, \sigma_{j-1}^2\right) \mathcal{N}\left(s_j, w_j, \sigma_j^2\right)$$



Example Transcription Output

IIT MADRAS
Indian Institute of Technology Madras

Example used in the previous slide

Element type	Start time*	Duration*	Quantized pitch	Notation	State
STA	0		7	P	${}_t\mathcal{A}$
STA	14		7	P	${}_t\mathcal{A}$
CPN	87	225	12	$\dot{S}$	$\mathcal{A}_t$
STA	327		11	N_3	$\mathcal{T}_a$
STA	347		12	$\dot{S}$	${}_t\mathcal{A}$
STA	383		11	N_3	$\mathcal{T}_a$
CPN	417	19	12	$\dot{S}$	$\mathcal{A}_t$
STA	464		9	D_2	${}_a\mathcal{T}$
STA	484		12	$\dot{S}$	$\mathcal{T}_a$
STA	529		9	D_2	${}_t\mathcal{A}$
STA	556		12	$\dot{S}$	${}_a\mathcal{T}$
STA	570		11	N_3	$\mathcal{T}_a$
CPN	585	19	12	$\dot{S}$	$\mathcal{A}_t$
STA	623		11	N_3	${}_a\mathcal{T}$
STA	640		12	$\dot{S}$	$\mathcal{T}_a$
CPN	686	70	7	P	$\mathcal{A}_t$

tcs Research

*Multiples of $T_0 \approx 4.44$ ms

Inventing for impact

Cosine Interpolation of transcription output

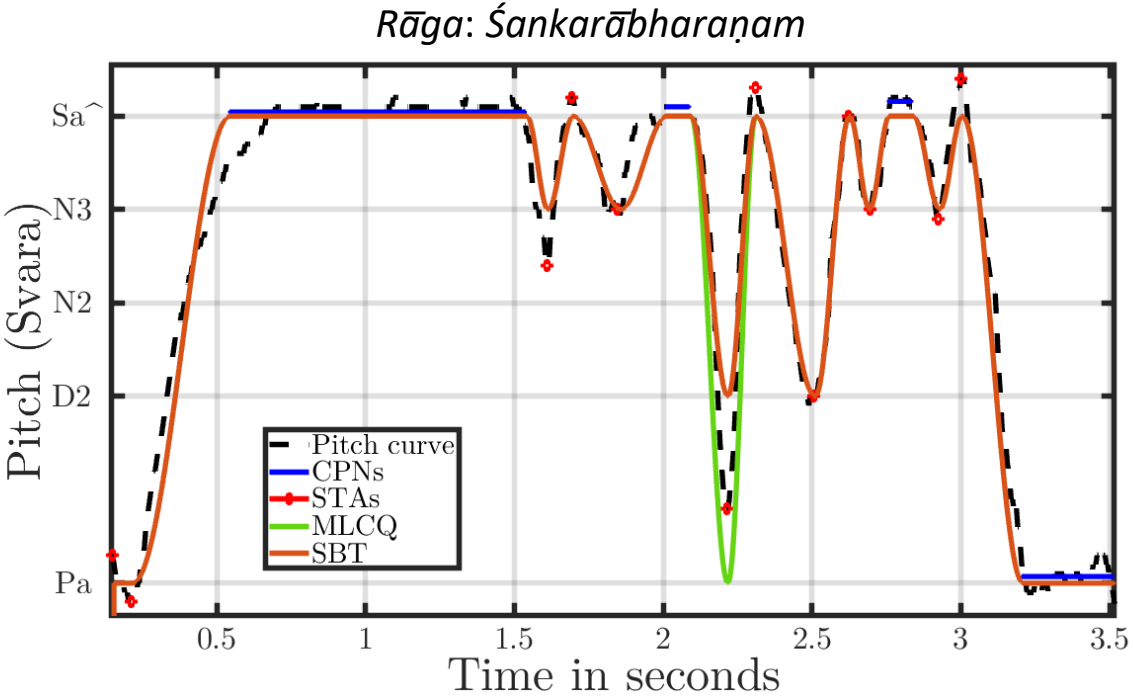
- Interpolate the transcription output to a complete pitch curve ($s[n]$ cents at nT_0)
  - Quantized CPN pitch values are used for the duration of the CPN (τ_k, d_k)
  - Quantized STA pitch values are placed at their respective locations (t_j)
- Quantized pitch-values v_1 and v_2 in semitones at time instants t_1 and t_2 are interpolated

$$\hat{s}[n] = 100v_1 + 100\frac{v_2 - v_1}{2}\left(1 - \cos\left(\pi\left(\frac{n - t_1}{t_2 - t_1}\right)\right)\right), t_1 < n < t_2$$

- Convert to frequency in Hz as

$$\hat{f}[n] = f_0 2^{\hat{s}[n]/1200}$$

Example of interpolated pitch curves



- Interpolated pitch curve in Hz is synthesized[1]

1. Kaustuv Kanti Ganguli and Preeti Rao, “Discrimination of melodic patterns in Indian classical music,” NCC 2015

Results: Perceptual Evaluation of Synthesis

IIT MADRAS
Indian Institute of Technology Madras

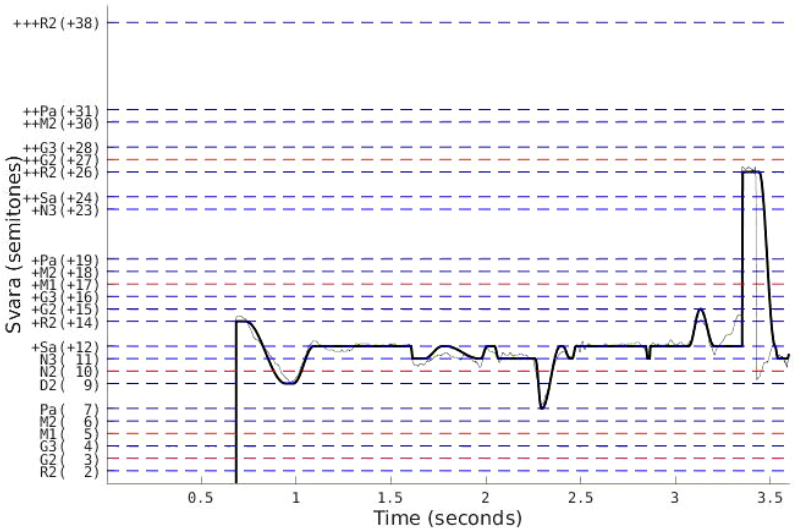
- Experiment: 27 listeners
  - Listened to original clip reference (~1 min long)
  - Rated synthesized outputs relative to the reference (0 to 100%)

Mean ratings in %		
Rāga↓ Technique→	Critical points[1] (Previous work)	SBT (Proposed)
Dhanyāsi	52	**67**
Kalyāṇī	48	**70**
Śankarābhāraṇam	47	**76**
Bhairavī	65	**74**

1. Ranjani H G et al., “A compact pitch and time representation for melodic contours in Indian art music”, JASA 2019

Transcription and synthesis of unseen renditions

- *Śankarābharaṇam*, *Tōḍī*, ***Bhairavī***, ***Rītigauḷa***
- *Yaman tān*[1] (Genre is not Carnatic)



1. Visualization was done with the assistance of Jom Kuriakose, CSE, IITM

Visualization for Correction

IIT MADRAS
Indian Institute of Technology Madras

- Transcription output still needs correction
- A website for correction[1]
- SBT offers granular correction

1. Front-end implemented by Jom Kuriakose, CSE, IITM

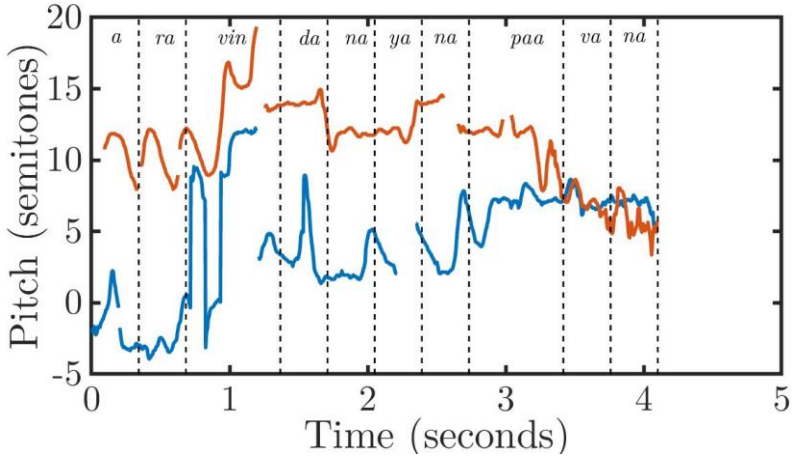
tcs Research

Inventing for impact

Conclusions

- CPN-STAs help understand CM pitch curves
- Detected targets
  - Explain properties of *rāga*s and *gamaka*s
  - Allow precision measurement in CM; CPNs are more precise than STAs
  - Constitute a descriptive transcription scheme
- Anchor-specific targets
  - Enable State-based Transcription
- Output needs correction by musicians to be used as ground truth
  - Granularity of anchor-specific targets helps correct systematic “errors”

Future Work



- First-level ground-truth for deep learning approaches (e.g., end to end)
- Syllables of the lyrics/*svara*s in CM must occur at specific points in the *tāḷa*.
  - Locations of syllables of lyrics/*svara*s should scale uniformly across rendition-speeds
  - The pitch curve between these points scales non-uniformly

Prescriptive transcription scales uniformly, descriptive transcription scales non-uniformly

- A joint study of syllables, rhythm, and melody is needed

tcs Research
Inventing for impact

IIT MADRAS
Indian Institute of Technology Madras

Thank you