# GROUP DELAY BASED MELODY EXTRACTION FOR INDIAN MUSIC

December 21, 2013

**Rajeev Rajan and Hema A. Murthy**
Department of Computer Science and Engineering
Indian Institute of Technology,Madras
e-mail:rajeevrajan002@gmail.com


# Outline

- Introduction
- Related Work
- Proposed method
  - Group delay function and Modified Group delay function
  - Melodic Pitch Extraction Using Modified Group Delay Function
  - Transient analysis by Multi-resolution Framework
  - Pitch Consistency by Dynamic Programming
  - Voicing Detection
  - Evaluation metrics
  - Results
- Conclusion

# Introduction

- Melody-The single (monophonic) pitch sequence that a listener might reproduce - when asked to hum
- Extract the pitch of leading instrument/singing voice in the presence of orchestral background.
- Melody pitch extraction polyphonic music- music with accompaniments
- Techniques:-Goto's PreFEst algorithm, Subharmonic summation spectrum, pitch contours using contour feature distributions. [1]

[1] Graham E.Poliner, Daniel P. W. Ellis, Andreas F. Ehmann, Emilia Gomez, Sebastian Strich and Beesuan Ong,*Melody Transcription From Music Audio :Approaches and Evaluations"*,IEEE Transactions on Audio, Speech, and Language Processing ,pp-1247–1256,Vol-15,No-4,May 2007

# Related Work

- Goto's PreFEst algorithm [1]- Frequency components are treated as a weighted mixture of all possible harmonic structure tone models
- Cao *et al* .[2]- Subharmonic summation spectrum and the harmonic structure tracking strategy
- Justin Salamon and Emilia Gomez[3] -using pitch contours characteristics
- V. Rao and P. Rao[4] -the temporal instability of voice harmonics to detect voice pitch

---

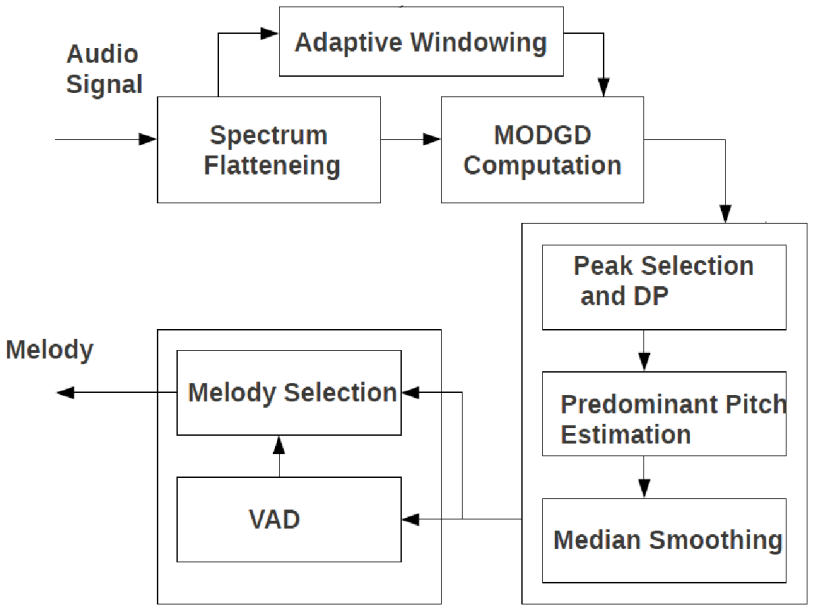
[1] M. Goto and S. Hayamizu, *"A real-time music scene description system: Detecting melody and bass lines in audio signals"*, Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis, pp 31-40

[2] C. Cao, M. Li, J. Liu, and Y. Yan, *"Singing melody extraction in polyphonic music by harmonic tracking,"* Proc. International Society for Music Information Retrieval (ISMIR),No.4, 2007.

[3] Justin Salamon and Emilia Gomez, *"Melody extraction from polyphonic music signals using pitch contours characteristics,"* In IEEE Trans. on Audio Speech and Language Processing, vol. 20, no. 6, pp. 1759- 1770, August 2012.

[4] V. Rao and P. Rao,*"Vocal melody extraction in the presence of pitched accompaniment in polyphonic music"* In Proc. of the IEEE Int. Conf. on Audio, Speech and Language Processing, no. 6, pp. 2145-2154, January 2010.

# Proposed method




- extract melodic pitch using Fourier transform phase
- The power spectrum of the music signal- flattened-MODGD
- Multi-resolution technique to capture the dynamic variation
- Dynamic programming to ensure consistency across frames

# Group delay function

- Group delay function $\tau(e^{j\omega})$ of a discrete time signal $x[n]$

$$\tau(e^{j\omega}) = -\frac{d\theta(e^{j\omega})}{d\omega} \tag{1}$$

- the group delay function can be computed directly from the signal by

$$= \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|\ X(e^{j\omega})\ |^2} \tag{2}$$

  $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of $x[n]$ and $nx[n]$ respectively.

- the group delay function is noisy–caused by the zeros of the source and convolution with the finite window length.

# Modified Group delay function

- To overcome the effects - the group delay function is modified

$$= \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|\ S(e^{j\omega})\ |^2} \tag{3}$$

where $S(e^{j\omega})$ is the cepstrally smoothed version of $X(e^{j\omega})$.[1]

| Steps | Algorithm |
|---|---|
| 1 | Let $x[n]$ be the given sequence. |
| 2 | Compute the DFT $X[k]$ , $Y[k]$, of $x[n]$ and $nx[n]$ respectively |
| 3 | Group delay function is $\tau_x[k] = \frac{X_R[k]Y_R[k]+X_I[k]Y_I[k]}{\|X[k]\|^2}$ <br> $R$ and $I$ represents real and imaginary respectively. |
| 4 | Modified group delay$\tau[k] = \frac{X_R[k]Y_R[k]+X_I[k]Y_I[k]}{\|S[k]\|^2}$, <br> where $S[k]$ is the smoothed version of $X[k]$ |
| 5 | Two new parameters $\alpha$ and $\gamma$ are introduced in Equation of $\tau[k]$ <br> $\tau_m[k] = \frac{\tau[k]}{\|\tau[k]\|}(\|\tau[k]\|)^\alpha$ <br> $\tau_m[k] = \frac{X_R[k]Y_R[k]+X_I[k]Y_I[k]}{\|S[k]\|^{2\gamma}}$ |

[1] Hema A. Murthy, Algorithms for Processing Fourier Transform Phase of Signals, PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, December 1991.

# Theory of Melodic Pitch Extraction Using Modified Group Delay Function

- Source-system model of music-Melody-The periodicity and amplitude of the source –Timbre information- the instrument or vocal tract.
- The periodicity of the source manifests as picket fence harmonics in the power spectrum.
- The timbral information can be suppressed- the picket fence harmonics-sinusoids
- The modified group delay function to resolve sinusoids in noise - in the context of extraction of melody for music.
- The $Z$-transform of two impulses separated by $T_o$.

$$E(z) = 1 + z^{-T_o} \tag{4}$$

- Fourier transform magnitude spectrum

$$\mid E(\omega) \mid^2 = \mid 2 + 2cos(\omega T_o) \mid^2 \tag{5}$$

- Replace $\omega$ by n and $T_0$ by $\omega_o$ and remove the dc component.

$$s[n] = cos(n\omega_o), n = 0, 1, 2, 3.....N - 1 \quad (6)$$

- Apply MODGD algorithm

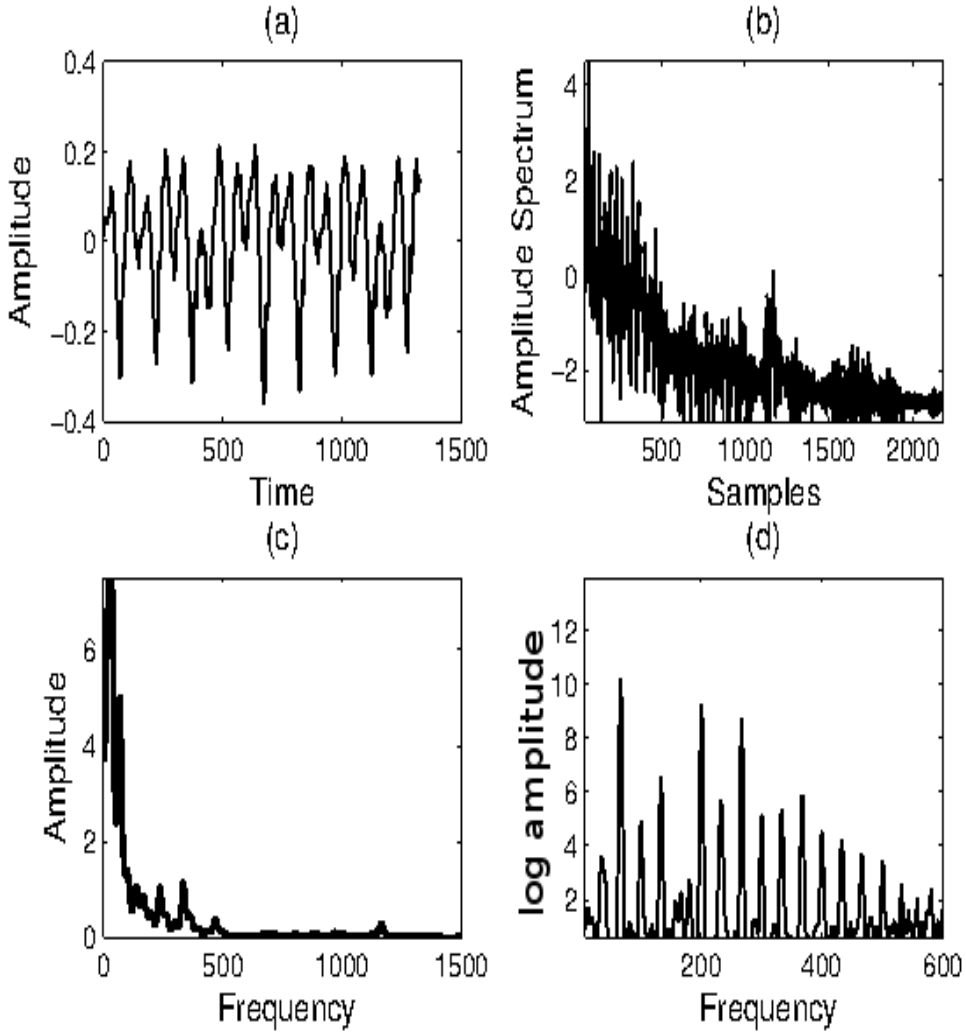



Figure: (a) Frame of music signal. (b) Magnitude spectrum. (c) Spectral envelope. (d) Flattened spectrum.

- Prominent peaks at multiples of the pitch period–reinforce the estimate of the pitch by folding over.
- Dynamic programming-consistency across frames in the pitch tracking.
- Adaptive Windowing

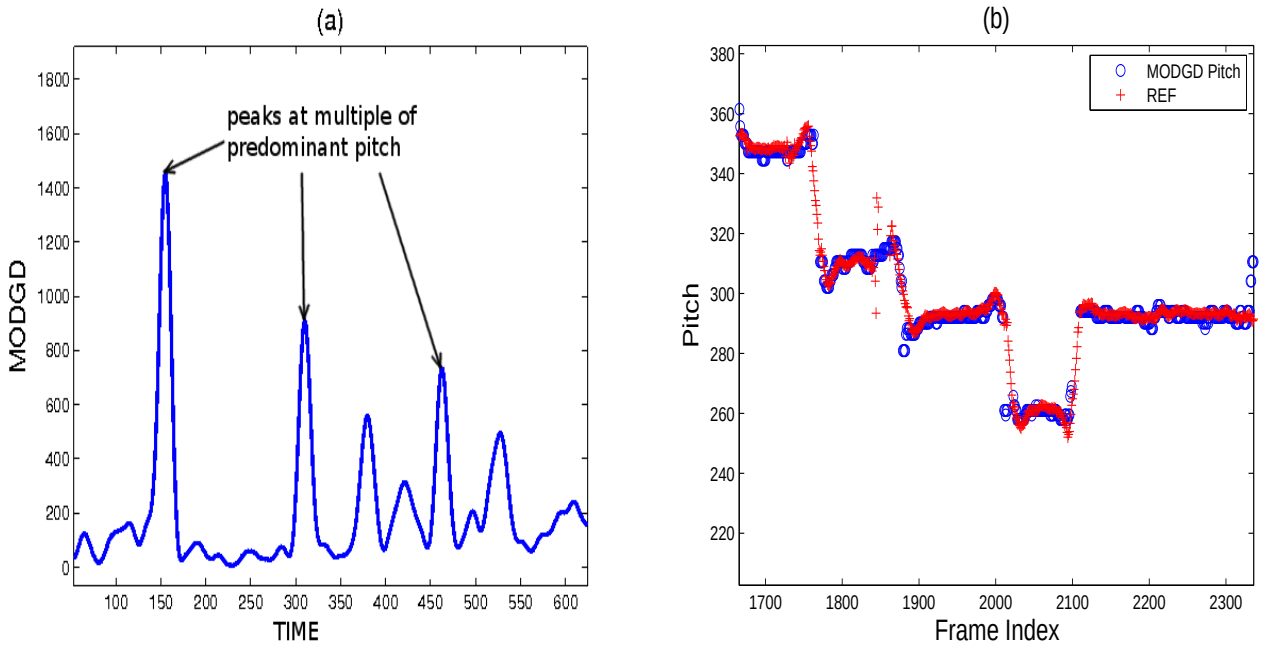



Figure: (a)MODGD plot for a frame. (b) Melody Pitch extraction for *'daisy2.wav'* using MODGD.

# Transient analysis by Multi-resolution Framework

- Transients-Variation in energy inputs,fast transitions.
- Multi-resolution framework in which shorter windows are used for transient segments and longer window otherwise
- low autocorrelation coefficient-transient

$$\rho(X,\tau,l) = \frac{\sum_k \mid X(k,l) \mid\mid X(k,l+\tau) \mid}{\sqrt{\sum_k \mid X(k,l) \mid^2 \mid X(k,l+\tau) \mid^2}} \tag{7}$$

where $X(k,l)$ denotes the $k^{th}$ coefficient of the discrete Fourier transform of the $l^{th}$ frame. $\tau$ corresponds to the autocorrelation lag.

# Pitch Consistency by Dynamic Programming

- combines local information and transition information
- local cost-pitch salience, transition cost- the relative closeness of the distance between locations of peaks in two consecutive frames.
- local cost -

$$C_l(c) = 1 - \frac{F(c)}{F_{max}}; \tag{8}$$

where $F(c)$ is the value of peak at the pitch candidate $c$ and $F_{max}$ is the maximum value of the peak

- Transition cost $C_t(c_j/c_{j-1})$ is the distance between the pitch candidates

$$C_t(c_j/c_{j-1}) = \frac{\mid L_j - L_{j-1} \mid}{l_{max}} \tag{9}$$

- Total cost(TC) = Local Cost + Transition  Cost
- optimal path
- pitch sequence starting from candidate $c$ followed by $d$

$$TC_{min} = C_1(c) + \min(C_{min}(d) + C_t(c/d)) \qquad (10)$$

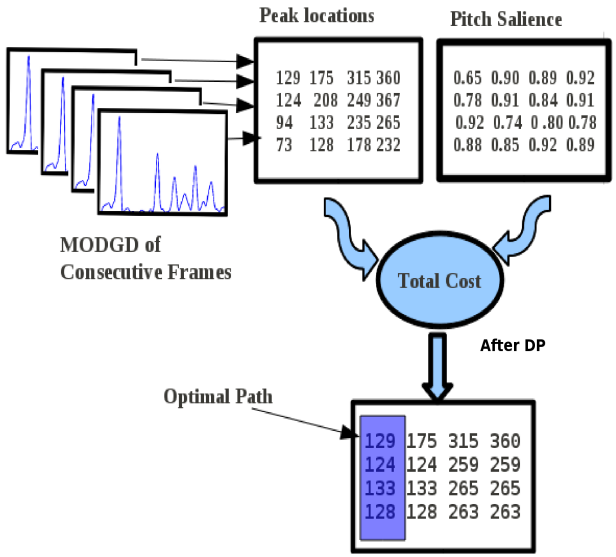



Figure: Computation of optimal path by dynamic programming

# Evaluation

- Evaluation Data set
  - MIREX2008 (North Indian Classical Music dataset): 4 excerpts of 1 min long each from north Indian classical vocal performances. -total of 8 audio clips.
  - Carnatic dataset :14 Carnatic alaapanas are used for evaluation purpose.
  - ADC-2004 dataset : 20 audio clips, styles : daisy, jazz, opera, MIDI and pop.
- Evaluation method

  The estimated pitch of a voiced frame will be considered correct when it satisfies the following condition:

$$\mid F_r(l) - F_e(l) \mid \leq \frac{1}{4} tone(50cents) \tag{11}$$

  where $F_r(l)$ and $F_e(l)$ denote reference frequency and estimated pitch frequency on the $l^{th}$ frame respectively.

# Voicing Detection

- Frame wise normalized harmonic energy
- Multiples of fundamental frequency are found out by searching the local maxima with $3\%$ tolerance.
- Harmonic energy of a signal $x[n]$ is computed by

$$E_n = \sum_{k=k_{F0}}^{K_{NF0}} \mid X[k] \mid^2 \tag{12}$$

Where $X[k]$, $k$, $F_o$ represent the Fourier transform magnitude, bin number, fundamental frequency respectively

# Evaluation Metrics

- Voicing Recall Rate (VR): the proportion of frames labeled voiced in the ground truth that are estimated as voiced by the algorithm.
- Voicing False Alarm Rate (VF): the proportion of frames labeled unvoiced in the ground truth that are estimated as voiced by the algorithm.
- Raw Pitch Accuracy (RPA): the ratio between the number of the correct pitch frames in voiced segments and the number of all voiced frames.
- Raw Chroma Accuracy (RCA) : same as raw pitch accuracy,- ignoring octave errors
- Overall Accuracy (OA) : this measure combines the performance of the pitch estimation and voicing detection

- The Standard deviation of the pitch detection $\sigma_e$: it is defined as:

$$\sigma_e = \sqrt{(\frac{1}{N}\sum(p_s - p'_s)^2 - e^2} \tag{13}$$

where $p_s$ is the standard pitch, $p'_s$ is the detected pitch, N is the number of correct pitch frames and e is the mean of the fine pitch error. e is defined as:

$$e = \frac{1}{N}\sum(p_s - p'_s) \tag{14}$$

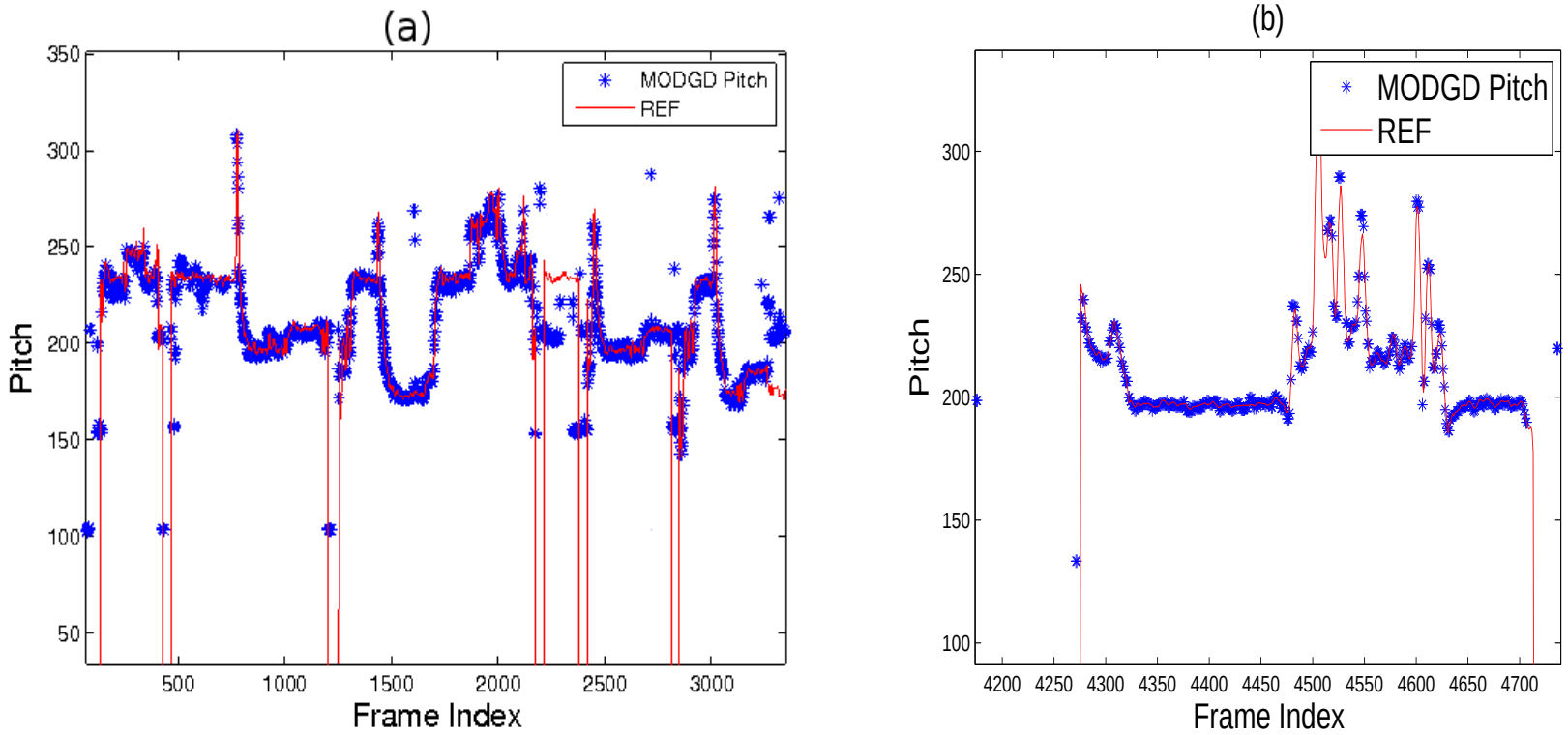



Figure: (a) Pitch extracted for MIREX2008 audio segment(b) Pitch extracted for a Carnatic segment

Original audio-GNB Kamboji4b 039GNB.wav Synthesized audio

# Results

Table: Comparison of five metrics for ADC dataset submitted in 2012/2011 evaluation.

| **Method** | **OA** | **RPA** | **RCA** | **VR** | **VF** |
|---|---|---|---|---|---|
| V.Arora et al | 69.06 | 81.41 | 85.92 | 76.51 | 23.56 |
| Sam Meyer | 60.34 | 64.23 | 71.21 | 77.36 | 32.96 |
| Bin Liao *et al*(1) | 46.24 | 55.87 | 66.71 | 99.98 | 97.76 |
| Bin Liao *et al*(2) | 41.54 | 48.32 | 59.90 | 99.96 | 95.37 |
| Bin Liao *et al*(3) | 41.54 | 48.32 | 59.90 | 99.96 | 95.37 |
| Salamon *et al* | 73.55 | 76.34 | 78.71 | 80.55 | 15.09 |
| Liao *et al* | 73.05 | 84.50 | 86.16 | 98.60 | 87.37 |
| Tachibana *et al* | 59.62 | 73.03 | 81.43 | 74.98 | 29.37 |
| **MODGD** | **60.78** | **67.80** | **75.95** | **82.26** | **26.00** |

Table: Comparison of five metrics for MIREX-2008 dataset submitted in 2012/2011 evaluation.

| | **OA** | **RPA** | **RCA** | **VR** | **VF** |
|---|---|---|---|---|---|
| V.Arora *et al* | 67.95 | 85.85 | 86.79 | 70.76 | 15.58 |
| Sam Meyer | 50.06 | 49.31 | 59.48 | 63.52 | 30.23 |
| Bin Liao *et al(1)* | 70.25 | 81.94 | 82.17 | 100.00 | 100.00 |
| Bin Liao *et al(2)* | 51.21 | 59.59 | 67.95 | 100.00 | 100.00 |
| Bin Liao *et al(3)* | 51.51 | 59.59 | 67.95 | 100.00 | 100.00 |
| Salamon *et al* | 82.78 | 87.55 | 88.02 | 89.26 | 17.86 |
| Chien *et al* | 68.88 | 71.75 | 74.67 | 89.60 | 44.81 |
| Stacy *et al* | 63.57 | 67.64 | 73.20 | 78.69 | 34.25 |
| **MODGD** | **58.21** | **64.44** | **66.05** | **82.88** | **27.94** |

Table: Comparison of $\sigma_e$, $RPA$, $RCA$ for Carnatic dataset [1]

| Method | $\sigma_e$ | $RPA$ | $RCA$ |
|---|---|---|---|
| YIN | 2.94 | 74.20 | 85.00 |
| **MODGD** | **2.67** | **75.16** | **80.49** |

[1] Ref: Melodia

# Conclusion

- An algorithm for extracting melody from music using modified group delay function.
- neither requires any substantial prior knowledge of the structure of musical pitch nor any classification framework.