Vocal Melody Extraction from Polyphonic Audio with Pitched Accompaniment

Vishweshwara Rao (05407001)
Ph.D. Defense
Guide: Prof. Preeti Rao

(June 2011)
Department of Electrical Engineering
Indian Institute of Technology
Bombay
OUTLINE

- Introduction
  - Objective, background, motivation, approaches & issues
  - Indian music

- Proposed melody extraction system
  - Design
  - Evaluation
  - Problems
    - Competing pitched accompanying instrument

- Enhancements for increasing robustness to pitched accompaniment
  - Dual-F0 tracking
  - Identification of vocal segments by combination of static and dynamic features
  - Signal-sparsity driven window length adaptation

- Graphical User Interface for melody extraction

- Conclusions and Future work
INTRODUCTION

Objective

- Vocal melody extraction from polyphonic audio
  - Polyphony: Multiple musical sound sources present
  - Vocal: Lead melodic instrument is the singing voice

- Melody
  - Sequence of notes
  - Symbolic representation of music
    - Pitch contour of the singing voice
INTRODUCTION

Background

- **Pitch**
  - Perceptual attribute of sound
  - Closely related to periodicity or fundamental frequency (F0)

- **Vocal pitch contour**

\[
F_0 = \frac{1}{T_0} = 100 \text{ Hz}
\]

\[
F_0 = \frac{1}{T_0} = 300 \text{ Hz}
\]
INTRODUCTION
Motivation, Complexity and Approaches

- Motivation
  - Music Information Retrieval applications – Query-by-singing/humming (QBSH), Artist ID, Cover Song ID
  - Music Edutainment – Singing learning, karaoke creation
  - Musicology

- Problem complexity
  - Singing – large F0 range, pitch dynamics, Diversity – Inter-singer, across cultures
  - Polyphony – Crowded signal – Percussive & tonal instruments

- Approaches
  - Understanding without separation
  - Source-separation [Lag08]
  - Classification [Pol05]
INTRODUCTION

Indian classical music: Signal characteristics

<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>15</td>
<td>1500</td>
</tr>
<tr>
<td>20</td>
<td>2000</td>
</tr>
</tbody>
</table>

- **Ghe Na Tun**
- **Tabla (percussion)**
- **Tanpura (drone)**
- **Harmonium (secondary melody)**
INTRODUCTION
Melody extraction in Indian classical music

- **Issues**
  - Signal complexity
    - Singing
    - Polyphony
  - Variable tonic
  - Non-availability of ground-truth data
    - Almost completely improvised (no universally accepted notation)

- **Example**

![Spectrogram example](image-url)
SYSTEM DESIGN
Our Approach

- Design considerations
  - Singing
  - Robustness to pitched accompaniment
  - Flexible
SYSTEM DESIGN
Signal Representation

- Frequency domain representation
  - Pitched sounds have harmonic spectra
  - Short-time analysis and DFT
  - Window-length
    - Chosen to resolve harmonics of minimum expected F0

- Sinusoidal representation
  - More compact & relevant
  - Different methods of sinusoid ID
    - Magnitude-based
    - Phase-based
  - Main-lobe matching (Sinusoidality) [Grif88] method found to be most reliable
    - Frequency transform of window has a known shape
    - Local peaks whose shape closely matches window main-lobe are declared as sinusoids

\[
X(n, \omega) = \sum_{m=0}^{M-1} x(m)w(n-m)e^{-\frac{2\pi imn}{M}}
\]
SYSTEM DESIGN
Multi-F0 Analysis

- **Objective**
  - To reliably detect the voice-F0 in polyphony with a high salience

- **F0-candidate identification**
  - Sub-multiples of well-formed sinusoids (Sinusoidality > 0.8)

- **F0-salience function**
  - Typical salience functions
    - Maximize Auto-correlation function (ACF)
    - Maximize comb-filter output
  - Harmonic sieve-type [Pol07]
    - Sensitive to strong harmonic sounds
  - Two-way mismatch [Mah94]
    - Error function sensitive to the deviation of measured partials/sinusoids from ideal harmonic locations

- **F0-candidate pruning**
  - Sort in ascending order of TWM errors
  - Prune weaker F0-candidates in close vicinity (25 cents) of stronger F0 candidates
SYSTEM DESIGN
Predominant-F0 Trajectory Extraction

- **Objective**
  - To find that path through the F0-candidate v/s time space that best represents the predominant-F0 trajectory

- **Dynamic-programming** [Ney83] based path finding
  - Measurement cost = TWM error
  - Smoothness cost must be based on musicological considerations

Cost functions

\[ W(p, p') = \text{OJC} \left| \log_2 \left( \frac{p'}{p} \right) \right| \]

\[ W(p, p') = 1 - e^{-\frac{\left(\log_2(p') - \log_2(p)\right)^2}{2\sigma^2}} \]

\( p' \) and \( p \) are F0s in current and previous frames resp.
EVALUATION
Predominant-F0 extraction: Indian Music

Data
- Classical: 4 min. of multi-track data, Film: 2 min. of multi-track data
- Ground truth: Output of YIN PDA [Chev02] on clean voice tracks with manual correction

Evaluation metrics
- Pitch Accuracy (PA) = % of vocal frames whose pitch has been correctly tracked (within 50 cents)
- Chroma Accuracy (CA) = PA except that octave errors are forgiven

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length</td>
<td>40 ms</td>
</tr>
<tr>
<td>Hop</td>
<td>10 ms</td>
</tr>
<tr>
<td>Lower limit on F0</td>
<td>100 Hz</td>
</tr>
<tr>
<td>Upper limit on F0</td>
<td>1280 Hz</td>
</tr>
<tr>
<td>Upper limit on spectral content</td>
<td>5000 Hz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genre</th>
<th>Audio content</th>
<th>PA (%)</th>
<th>CA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian classical music</td>
<td>Voice + percussion</td>
<td>97.4</td>
<td>99.5</td>
</tr>
<tr>
<td>Indian classical music</td>
<td>Voice + percussion + drone</td>
<td>92.9</td>
<td>98.2</td>
</tr>
<tr>
<td>Indian classical music</td>
<td>Voice + percussion + drone + harmonium</td>
<td>67.3</td>
<td>73.04</td>
</tr>
<tr>
<td>Indian pop music</td>
<td>Voice + guitar</td>
<td>96.7</td>
<td>96.7</td>
</tr>
</tbody>
</table>
SYSTEM DESIGN
Voicing Detection

- **Features**
  - FS1 – 13 MFCCs
  - FS2 – 7 static timbral features
  - FS3 – Normalized harmonic energy (NHE)

- **Classifier**
  - GMM – 4 mixtures per class

- **Boundary detector**
  - Audio novelty detector [Foote] with NHE

- **Data**
  - 23 min. of Hindustani training data
  - 7 min. of Hindustani testing data

### Results on testing data

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Vocal recall (%)</th>
<th>Instrumental recall (%)</th>
<th>Vocal recall (%)</th>
<th>Instrumental recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS1</td>
<td>92.17</td>
<td>66.43</td>
<td>97.86</td>
<td>61.61</td>
</tr>
<tr>
<td>FS2</td>
<td>92.38</td>
<td>66.29</td>
<td>96.28</td>
<td>68.98</td>
</tr>
<tr>
<td>FS3</td>
<td>89.05</td>
<td>92.10</td>
<td>93.15</td>
<td>96.58</td>
</tr>
</tbody>
</table>
EVALUATION
Submission to MIREX 2008 & 2009

Music Information Retrieval Evaluation eXchange
- Started in 2004
- International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL)
- Common platform for evaluation on common datasets

Tasks
- Audio genre, artist, mood classification
- Audio melody extraction
- Audio beat tracking
- Audio Key detection
- Query by singing/ humming
- Audio chord estimation

Diagram:
- Music Signal
  - DFT
  - Main Lobe Matching
  - Parabolic interpolation
- Sinusoids frequencies and magnitudes
  - DFT
  - TWM error computation
  - Sort (Ascending)
  - Vicinity Pruning
- F0 candidates and measurement costs
  - Multi-F0 analysis
  - Predominant F0 trajectory extraction
- Predominant F0 contour
  - Thresholding normalized harmonic energy
  - Grouping over homogenous segments
  - Dynamic programming-based optimal path finding
  - Vocal segment pitch tracks
EVALUATION
MIREX 2008 & 2009 – Datasets & Evaluation

Data
- ADC 2004: Publicly available data
  - 20 excerpts (about 20 sec each) from pop, opera, jazz & midi
- MIREX 2005: Secret data
  - 25 excerpts (10 – 40 sec) from rock, R&B, pop, jazz, solo piano
- MIREX 2008: ICM data
  - 4 excerpts of 1 minute each from a male and female Hindustani vocal performance. 2 min. each with and without a loud harmonium
- MIREX 2009: MIR 1K data
  - 374 Karaoke recordings of Chinese songs. Each recording is mixed at 3 different Signal-to-accompaniment ratios (SARs) {-5,0,5 dB}

Evaluation metrics:
- Pitch evaluation
  - Pitch accuracy (PA) and Chroma accuracy (CA)
- Voicing evaluation
  - Vocal recall (Vx recall) and Vocal false alarm rate (Vx false alm)
- Overall accuracy
  - % of correctly detected vocal frames with correctly detected pitch
- Run-time
# EVALUATION

## MIREX 2009 & 2010 – MIREX’05 dataset (vocal)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Vx Recall</th>
<th>Vx False Alm</th>
<th>Pitch accuracy</th>
<th>Chroma accuracy</th>
<th>Overall Accuracy</th>
<th>Runtime (dd:hh:mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cl1</td>
<td>91.2267</td>
<td>67.7071</td>
<td>70.807</td>
<td>73.924</td>
<td>59.7095</td>
<td>-</td>
</tr>
<tr>
<td>cl2</td>
<td>80.9512</td>
<td>44.8062</td>
<td>70.807</td>
<td>73.924</td>
<td>64.4609</td>
<td>-</td>
</tr>
<tr>
<td>dr1</td>
<td>93.7543</td>
<td>53.5255</td>
<td>76.1145</td>
<td>77.7138</td>
<td>66.9613</td>
<td>-</td>
</tr>
<tr>
<td>dr2</td>
<td>88.0635</td>
<td>38.0207</td>
<td>70.9258</td>
<td>75.9161</td>
<td>66.5175</td>
<td>-</td>
</tr>
<tr>
<td>hjc1</td>
<td>65.8441</td>
<td>19.8206</td>
<td>62.6594</td>
<td>73.4868</td>
<td>54.8538</td>
<td>-</td>
</tr>
<tr>
<td>hjc2</td>
<td>65.8441</td>
<td>19.8206</td>
<td>54.1294</td>
<td>69.4221</td>
<td>52.2905</td>
<td>-</td>
</tr>
<tr>
<td>jjy</td>
<td>88.8546</td>
<td>41.9813</td>
<td>76.2696</td>
<td>79.3236</td>
<td>66.3123</td>
<td>-</td>
</tr>
<tr>
<td>kd</td>
<td>82.6017</td>
<td>15.2554</td>
<td>77.4622</td>
<td>80.8218</td>
<td>76.962</td>
<td>-</td>
</tr>
<tr>
<td>mw</td>
<td>99.9345</td>
<td>99.7947</td>
<td>75.7398</td>
<td>80.3791</td>
<td>53.7323</td>
<td>-</td>
</tr>
<tr>
<td>pc</td>
<td>75.6436</td>
<td>21.0879</td>
<td>71.7068</td>
<td>72.5089</td>
<td>70.465</td>
<td>-</td>
</tr>
<tr>
<td>rr</td>
<td>92.908</td>
<td>56.3639</td>
<td>75.9506</td>
<td>79.1084</td>
<td>65.7745</td>
<td>-</td>
</tr>
<tr>
<td>toos</td>
<td>91.2267</td>
<td>67.7071</td>
<td>70.807</td>
<td>73.924</td>
<td>59.7095</td>
<td>-</td>
</tr>
</tbody>
</table>

**2010**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Vx Recall</th>
<th>Vx False Alm</th>
<th>Pitch accuracy</th>
<th>Chroma accuracy</th>
<th>Overall Accuracy</th>
<th>Runtime (dd:hh:mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HJ1</td>
<td>70.80</td>
<td>44.90</td>
<td>71.70</td>
<td>74.93</td>
<td>53.89</td>
<td>00:59:31</td>
</tr>
<tr>
<td>TOOS1</td>
<td>84.65</td>
<td>41.93</td>
<td>68.87</td>
<td>74.62</td>
<td>60.84</td>
<td>00:50:31</td>
</tr>
<tr>
<td>JJJ2</td>
<td>96.86</td>
<td>69.62</td>
<td>70.17</td>
<td>78.32</td>
<td>60.81</td>
<td>01:09:30</td>
</tr>
<tr>
<td>JJJ1</td>
<td>97.33</td>
<td>70.16</td>
<td>71.58</td>
<td>78.61</td>
<td>61.54</td>
<td>03:48:02</td>
</tr>
<tr>
<td>SG1</td>
<td>76.39</td>
<td>22.78</td>
<td>61.80</td>
<td>73.70</td>
<td>62.13</td>
<td>00:08:15</td>
</tr>
</tbody>
</table>
## EVALUATION

### MIREX 2009 & 2010 – MIREX’09 dataset (0 dB mix)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Vx Recall</th>
<th>Vx False Alm</th>
<th>Pitch accuracy</th>
<th>Chroma accuracy</th>
<th>Overall Accuracy</th>
<th>Runtime (dd:hh:mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cl1</td>
<td>92.4858</td>
<td>83.5749</td>
<td>59.138</td>
<td>62.9508</td>
<td>43.9659</td>
<td>00:00:28</td>
</tr>
<tr>
<td>cl2</td>
<td>77.2085</td>
<td>59.7352</td>
<td>59.138</td>
<td>62.9508</td>
<td>49.2294</td>
<td>00:00:33</td>
</tr>
<tr>
<td>dr1</td>
<td>91.8716</td>
<td>55.3555</td>
<td>69.8804</td>
<td>72.5138</td>
<td>60.1294</td>
<td>16:00:00</td>
</tr>
<tr>
<td>dr2</td>
<td>87.3985</td>
<td>47.3422</td>
<td>66.549</td>
<td>70.7923</td>
<td>59.5076</td>
<td>00:08:44</td>
</tr>
<tr>
<td>hjc1</td>
<td>34.1722</td>
<td>1.7909</td>
<td>72.6577</td>
<td>75.2906</td>
<td>53.1752</td>
<td>00:05:44</td>
</tr>
<tr>
<td>hjc2</td>
<td>34.1722</td>
<td>1.7909</td>
<td>51.6871</td>
<td>70.002</td>
<td>51.7469</td>
<td>00:09:38</td>
</tr>
<tr>
<td>jjy</td>
<td>38.906</td>
<td>19.4063</td>
<td>75.9354</td>
<td>80.2461</td>
<td>49.686</td>
<td>02:14:06</td>
</tr>
<tr>
<td>kd</td>
<td>91.1846</td>
<td>47.7842</td>
<td>80.4565</td>
<td>81.8811</td>
<td>68.2237</td>
<td>00:00:24</td>
</tr>
<tr>
<td>mw</td>
<td>99.992</td>
<td>99.4688</td>
<td>67.2905</td>
<td>71.0018</td>
<td>43.6365</td>
<td>00:02:12</td>
</tr>
<tr>
<td>pc</td>
<td>73.1175</td>
<td>43.4773</td>
<td>50.8895</td>
<td>53.3672</td>
<td>51.5001</td>
<td>03:05:57</td>
</tr>
<tr>
<td>rr</td>
<td><strong>88.8091</strong></td>
<td><strong>50.7595</strong></td>
<td><strong>68.6242</strong></td>
<td><strong>71.3714</strong></td>
<td><strong>60.7733</strong></td>
<td><strong>00:00:26</strong></td>
</tr>
<tr>
<td>toos</td>
<td>99.9829</td>
<td>99.4185</td>
<td>82.2943</td>
<td>85.7474</td>
<td>53.5623</td>
<td>01:00:28</td>
</tr>
</tbody>
</table>

### 2010

<table>
<thead>
<tr>
<th>Participant</th>
<th>Vx Recall</th>
<th>Vx False Alm</th>
<th>Pitch accuracy</th>
<th>Chroma accuracy</th>
<th>Overall Accuracy</th>
<th>Runtime (dd:hh:mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HJ1</td>
<td>82.06</td>
<td>14.27</td>
<td>83.15</td>
<td>84.23</td>
<td>76.17</td>
<td>14:39:16</td>
</tr>
<tr>
<td>TOOS1</td>
<td>94.17</td>
<td>38.58</td>
<td>82.59</td>
<td>86.18</td>
<td>72.23</td>
<td>12:07:21</td>
</tr>
<tr>
<td>JJJ2</td>
<td>98.33</td>
<td>70.62</td>
<td>81.29</td>
<td>83.83</td>
<td>62.55</td>
<td>14:06:20</td>
</tr>
<tr>
<td>JJJ1</td>
<td>98.34</td>
<td>70.65</td>
<td>82.20</td>
<td>84.57</td>
<td>62.90</td>
<td>65:21:11</td>
</tr>
<tr>
<td>SG1</td>
<td>89.65</td>
<td>30.22</td>
<td><strong>80.05</strong></td>
<td><strong>85.50</strong></td>
<td>73.59</td>
<td>01:56:27</td>
</tr>
</tbody>
</table>
EVALUATION
Problems in Melody Extraction

- “No bold increase in melody extraction over the last 3 years (2007-2009)” [Dres2010]

- Errors due to loud pitched accompaniment
  - Accompaniment pitch tracked *instead* of voice
    - Error in Predominant-F0 trajectory extraction
  - Accompaniment pitch tracked *along* with voice
    - Error in voicing detection

- Errors due to signal dynamics
  - Octave errors due to fixed window length
    - Error in Signal representation
ENHANCEMENTS: PREDOMINANT-F0 TRACKING

Problems

- Incorrect tracking of loud pitched accompaniment

- ICM Data – Largest reduction in accuracy for audio in which
  - Voice displays large rapid modulations
  - Instrument pitch is flat

- Predominant F0 trajectory
  - DP-based path finding
  - Based on suitably defined
    - Measurement cost
    - Smoothness cost

- Accompaniment errors
  - Bias in measurement cost: Salient (spectrally rich) instrument
  - Bias in smoothness cost: Stable-pitched instrument
ENHANCEMENTS: PREDOMINANT-F0 TRACKING

Design & Implementation

- Extension of DP to track F0 candidate ordered pairs (nodes) – called Dual-F0 tracking

- Node formation
  - All possible pairs
    - Computationally expensive ($10P_2 = 90$)
    - F0 and (sub) multiple may be tracked
  - Prohibit pairing of harmonically related F0 candidates
    - Low harmonic threshold of 5 cents
    - Allows pairing of voice F0 and octave-separated instrument F0 because of voice detuning

- Node measurement cost computation
  - Joint TWM error [Mah94]

- Node smoothness cost computation
  - Sum of corresponding F0 candidate smoothness costs

- Final selection of predominant-F0 contour
  - Based on voice-harmonic instability
ENHANCEMENTS: PREDOMINANT-F0 TRACKING

Selection of Predominant-F0 contour

- Harmonic Sinusoidal Model (HSM)
  - Partial tracking algorithm used in SMS [Serra98]
  - Tracks are indexed and linked by harmonic number

- Std. dev. pruning
  - Prune tracks in 200 ms segments whose std. dev. <2 Hz

- Mark that 200 ms segment with greater residual energy as predominant-F0

Spectrogram

HSM (before pruning)

HSM (after pruning)
ENHANCEMENTS: PREDOMINANT-F0 TRACKING

Final Implementation – Block Diagram

Music Signal

DFT → Main-lobe matching → Parabolic interpolation

Signal representation

Sinusoids frequencies and magnitudes

Sub-multiples of sinusoids $C$

$F0$ candidates

TWM error computation → Sorting (Ascending) → Vicinity pruning

Multi-F0 analysis

$F0$ candidates and saliences

Ordered pairing of $F0$ candidates with harmonic constraint

Joint TWM error computation

Nodes (F0 pairs)

Predominant-F0 trajectory extraction

Optimal path finding

Optimal path finding → Vocal pitch identification

Melodic contour

Optimal path finding

Vocal pitch identification

Melodic contour

Department of Electrical Engineering, IIT Bombay
ENHANCEMENTS: PREDOMINANT-F0 TRACKING

Experimental evaluation: Setup

- Participating systems
  - TWMDP (single- and dual-F0)
  - LIWANG [LiWang07]
    - Uses HMM to track predominant-F0
    - Includes the possibility of a 2-pitch hypothesis but finally outputs a single F0
    - Shown to be superior to other contemporary systems
  - Same F0 search range [80 500 Hz]

- Evaluation metrics
  - Multi-F0 stage
    - % presence of true voice F0 in candidate list
  - Predominant-F0 extraction (PA & CA)
    - Single F0
    - Dual-F0
      - Final contour accuracy
      - Either-Or accuracy : Correct pitch is present in at least one of the two outputs
ENHANCEMENTS: PREDOMINANT-F0 TRACKING
Experimental Evaluation: Data & Results

Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Vocal (sec)</th>
<th>Total (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Li &amp; Wang data</td>
<td>55.4</td>
<td>97.5</td>
</tr>
<tr>
<td>2</td>
<td>Examples from MIR-1k dataset with loud pitched accompaniment</td>
<td>61.8</td>
<td>98.1</td>
</tr>
<tr>
<td>3</td>
<td>Examples from MIREX’08 data (Indian classical music)</td>
<td>91.2</td>
<td>99.2</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>208.4</strong></td>
<td><strong>294.8</strong></td>
</tr>
</tbody>
</table>

Multi-F0 Evaluation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Percentage Presence of Voice-F0 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Top 5 Candidates</strong></td>
</tr>
<tr>
<td>1</td>
<td>92.9</td>
</tr>
<tr>
<td>2</td>
<td>88.5</td>
</tr>
<tr>
<td>3</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Comparison of TWMDP single-F0 tracker and LIWANG for dataset 1

(a) Pitch accuracies (%)
(b) Chroma accuracies (%)

Department of Electrical Engineering, IIT Bombay
ENHANCEMENTS: PREDOMINANT-F0 TRACKING
Experimental evaluation: Results (Dual-F0)

- TWMDP Single-F0 significantly better than LIWANG system for all datasets
- TWMDP Dual-F0 significantly better than TWMDP single-F0 for datasets 2 & 3
- Scope for further improvement in final predominant-F0 identification
  - Indicated by difference between TWMDP Dual-F0 Either-Or and Final accuracies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Single-F0 (A2)</th>
<th>Dual-F0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA (%)</td>
<td>CA (%)</td>
</tr>
<tr>
<td>1</td>
<td>88.5 (8.3)</td>
<td>90.2 (6.4)</td>
</tr>
<tr>
<td></td>
<td>89.3 (0.9)</td>
<td>92.0 (1.1)</td>
</tr>
<tr>
<td></td>
<td>84.1 (2.9)</td>
<td>88.8 (3.9)</td>
</tr>
<tr>
<td>2</td>
<td>57.0 (24.5)</td>
<td>61.1 (14.2)</td>
</tr>
<tr>
<td></td>
<td>74.2 (-6.8)</td>
<td>81.2 (-5.3)</td>
</tr>
<tr>
<td></td>
<td>69.1 (50.9)</td>
<td>74.1 (38.5)</td>
</tr>
<tr>
<td>3</td>
<td>66.0 (11.3)</td>
<td>66.5 (9.7)</td>
</tr>
<tr>
<td></td>
<td>85.7 (30.2)</td>
<td>87.1 (18.0)</td>
</tr>
<tr>
<td></td>
<td>73.9 (24.6)</td>
<td>76.3 (25.9)</td>
</tr>
</tbody>
</table>

TWMDP Single-F0 significantly better than LIWANG system for all datasets
TWMDP Dual-F0 significantly better than TWMDP single-F0 for datasets 2 & 3
Scope for further improvement in final predominant-F0 identification
- Indicated by difference between TWMDP Dual-F0 Either-Or and Final accuracies

![Graph showing performance comparison between A3, A2, and A1 across datasets D2 and D3 for PA and CA metrics]
ENHANCEMENTS: PREDOMINANT-F0 EXTRACTION

Example of F0 collisions

Contour switching occurs at F0 collisions
ENHANCEMENTS: VOICING DETECTION
Problems
ENHANCEMENTS: VOICING DETECTION

Features

- Proposed feature set – combination of static & dynamic features
- Features extracted using a harmonic sinusoidal model representation
- Feature selection in each feature set using information entropy [Weka]

<table>
<thead>
<tr>
<th>C1 Static timbral</th>
<th>C2 Dynamic timbral</th>
<th>C3 Dynamic F0-Harmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F0</strong></td>
<td>( \Delta ) 10 Harmonic powers</td>
<td>Mean &amp; <strong>median</strong> of ( \Delta F_0 )</td>
</tr>
<tr>
<td>10 Harmonic powers</td>
<td>( \Delta ) SC &amp; ( \Delta ) SE</td>
<td>Mean, median &amp; Std.Dev. of ( \Delta ) Harmonic ( \epsilon [0 \ 2 \ kHz] )</td>
</tr>
<tr>
<td>Spectral centroid (SE)</td>
<td><strong>Std. Dev.</strong> of SC for 0.5, 1 &amp; 2 sec</td>
<td>Mean, median &amp; Std.Dev. of ( \Delta ) Harmonic ( \epsilon [2 \ 5 \ kHz] )</td>
</tr>
<tr>
<td>Sub-band energy (SE)</td>
<td><strong>MER</strong> of SC for 0.5, 1 &amp; 2 sec</td>
<td>Mean, median &amp; Std.Dev. of ( \Delta ) Harmonics 1 to 5</td>
</tr>
<tr>
<td></td>
<td><strong>Std. Dev.</strong> of SE for 0.5, 1 &amp; 2 sec</td>
<td>Mean, <strong>median</strong> &amp; Std.Dev. of ( \Delta ) Harmonics 6 to 10</td>
</tr>
<tr>
<td></td>
<td>MER of SE for 0.5, 1 &amp; 2 sec</td>
<td>Ratio of mean, median &amp; Std.dev. of ( \Delta ) Harmonics 1 to 5 : ( \Delta ) Harmonics 6 to 10</td>
</tr>
</tbody>
</table>

MER – Modulation energy ratio
ENHANCEMENTS: VOICING DETECTION

Data

<table>
<thead>
<tr>
<th>Genre</th>
<th>Number of songs</th>
<th>Vocal duration</th>
<th>Instrumental duration</th>
<th>Overall duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Western</td>
<td>11</td>
<td>7m 19s</td>
<td>7m 02s</td>
<td>14m 21s</td>
</tr>
<tr>
<td>II. Greek</td>
<td>10</td>
<td>6m 30s</td>
<td>6m 29s</td>
<td>12m 59s</td>
</tr>
<tr>
<td>III. Bollywood</td>
<td>13</td>
<td>6m 10s</td>
<td>6m 26s</td>
<td>12m 36s</td>
</tr>
<tr>
<td>IV. Hindustani</td>
<td>8</td>
<td>7m 10s</td>
<td>5m 24s</td>
<td>12m 54s</td>
</tr>
<tr>
<td>V. Carnatic</td>
<td>12</td>
<td>6m 15s</td>
<td>5m 58s</td>
<td>12m 13s</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>45</strong></td>
<td><strong>33m 44s</strong></td>
<td><strong>31m 19s</strong></td>
<td><strong>65m 03s</strong></td>
</tr>
</tbody>
</table>

**Genre Number of songs Vocal duration Instrumental duration Overall duration**

<table>
<thead>
<tr>
<th>Genre</th>
<th>Singing</th>
<th>Dominant Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV. Hindustani</td>
<td>Syllabic and melismatic. Varies from long, pitch-flat, vowel-only notes to large &amp; rapid modulations.</td>
<td>Mainly flat-note harmonium (woodwind). Pitch range overlapping with voice.</td>
</tr>
<tr>
<td>V. Carnatic</td>
<td>Syllabic and melismatic. Replete with fast pitch modulations.</td>
<td>Mainly pitch-modulated violin. F0 range generally higher than voice but has some overlap in pitch range.</td>
</tr>
</tbody>
</table>
**ENHANCEMENTS: VOICING DETECTION**

**Evaluation**

- Two cross-validation experiments
  - Intra-genre – Leave 1 song out
  - Inter-genre – Leave 1 genre out

- Feature combination
  - Concatenation
  - Classifier combination

- Baseline features
  - 13 MFCCs [Roc07]

- Evaluation
  - Vocal Recall (%) and precision (%)

- Overall Results
  - C1 better than baseline
  - C1+C2+C3 better than C1
  - Classifier combination better than feature concatenation
ENHANCEMENTS: VOICING DETECTION
Evaluation (contd.)

Leave 1 Song out (Recall %)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>77.2</td>
<td>66.0</td>
<td>65.6</td>
<td>82.6</td>
<td>83.2</td>
<td>74.9</td>
</tr>
<tr>
<td>F0-MFCCs</td>
<td>78.9</td>
<td>77.8</td>
<td>78.0</td>
<td>85.9</td>
<td>85.9</td>
<td>81.2</td>
</tr>
<tr>
<td>C1</td>
<td>79.6</td>
<td>77.4</td>
<td>79.3</td>
<td>82.3</td>
<td>87.1</td>
<td>81.0</td>
</tr>
<tr>
<td>C1+C2</td>
<td>82.3</td>
<td>83.6</td>
<td>85.4</td>
<td>83.3</td>
<td>86.8</td>
<td>84.2</td>
</tr>
<tr>
<td>C1+C3</td>
<td>80.2</td>
<td>83.4</td>
<td>81.7</td>
<td>89.7</td>
<td>88.2</td>
<td>84.5</td>
</tr>
<tr>
<td>C1+C2+C3</td>
<td>81.1</td>
<td>86.9</td>
<td>86.4</td>
<td>88.5</td>
<td>87.3</td>
<td>85.9</td>
</tr>
</tbody>
</table>

**Semi-automatic F0-driven HSM**

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>77.2</td>
<td>66.0</td>
<td>65.6</td>
<td>82.6</td>
<td>83.2</td>
<td>74.9</td>
</tr>
<tr>
<td>F0-MFCCs</td>
<td>76.9</td>
<td>77.8</td>
<td>78.0</td>
<td>85.9</td>
<td>85.9</td>
<td>81.2</td>
</tr>
<tr>
<td>C1</td>
<td>79.6</td>
<td>77.4</td>
<td>79.3</td>
<td>82.3</td>
<td>87.1</td>
<td>81.0</td>
</tr>
<tr>
<td>C1+C2</td>
<td>82.3</td>
<td>83.6</td>
<td>85.4</td>
<td>83.3</td>
<td>86.8</td>
<td>84.2</td>
</tr>
<tr>
<td>C1+C3</td>
<td>80.2</td>
<td>83.4</td>
<td>81.7</td>
<td>89.7</td>
<td>88.2</td>
<td>84.5</td>
</tr>
<tr>
<td>C1+C2+C3</td>
<td>81.1</td>
<td>86.9</td>
<td>86.4</td>
<td>88.5</td>
<td>87.3</td>
<td>85.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>77.2</td>
<td>66.0</td>
<td>65.6</td>
<td>82.6</td>
<td>83.2</td>
<td>74.9</td>
</tr>
<tr>
<td>F0-MFCCs</td>
<td>76.9</td>
<td>77.8</td>
<td>78.0</td>
<td>85.9</td>
<td>85.9</td>
<td>81.2</td>
</tr>
<tr>
<td>C1</td>
<td>79.6</td>
<td>77.4</td>
<td>79.3</td>
<td>82.3</td>
<td>87.1</td>
<td>81.0</td>
</tr>
<tr>
<td>C1+C2</td>
<td>82.3</td>
<td>83.6</td>
<td>85.4</td>
<td>83.3</td>
<td>86.8</td>
<td>84.2</td>
</tr>
<tr>
<td>C1+C3</td>
<td>80.2</td>
<td>83.4</td>
<td>81.7</td>
<td>89.7</td>
<td>88.2</td>
<td>84.5</td>
</tr>
<tr>
<td>C1+C2+C3</td>
<td>81.1</td>
<td>86.9</td>
<td>86.4</td>
<td>88.5</td>
<td>87.3</td>
<td>85.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>77.2</td>
<td>66.0</td>
<td>65.6</td>
<td>82.6</td>
<td>83.2</td>
<td>74.9</td>
</tr>
<tr>
<td>F0-MFCCs</td>
<td>76.9</td>
<td>77.8</td>
<td>78.0</td>
<td>85.9</td>
<td>85.9</td>
<td>81.2</td>
</tr>
<tr>
<td>C1</td>
<td>79.6</td>
<td>77.4</td>
<td>79.3</td>
<td>82.3</td>
<td>87.1</td>
<td>81.0</td>
</tr>
<tr>
<td>C1+C2</td>
<td>82.3</td>
<td>83.6</td>
<td>85.4</td>
<td>83.3</td>
<td>86.8</td>
<td>84.2</td>
</tr>
<tr>
<td>C1+C3</td>
<td>80.2</td>
<td>83.4</td>
<td>81.7</td>
<td>89.7</td>
<td>88.2</td>
<td>84.5</td>
</tr>
<tr>
<td>C1+C2+C3</td>
<td>81.1</td>
<td>86.9</td>
<td>86.4</td>
<td>88.5</td>
<td>87.3</td>
<td>85.9</td>
</tr>
</tbody>
</table>

**Fully-automatic F0-driven HSM**

- Genre-specific feature set adaptation
  - C1+C2 – Western
  - C1+C3 - Hindustani
ENHANCEMENTS: SIGNAL REPRESENTATION

Sparsity-driven window length adaptation

- Relation between window length and signal characteristics
  - Dense spectrum (multiple harmonic sources) -> long window
  - Non-stationarity (rapid pitch modulations) -> short window

- Adaptive time segmentation for signal modeling and synthesis [Good97]
  - Based on minimizing reconstruction error between synthesized and original signals
  - High computational cost

- Easily computable measures for adapting window length
  - Signal sparsity – sparse spectrum has concentrated components
  - Window length selection (23.2, 46.4 92.9 ms) based on maximizing signal sparsity

<table>
<thead>
<tr>
<th>L2 Norm</th>
<th>Normalized kurtosis</th>
<th>Gini Index</th>
<th>Hoyer measure</th>
<th>Spectral flatness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_2 = \sqrt{\sum_k X_n^2(k)}$</td>
<td>$KU = \frac{1}{N} \sum_k</td>
<td>X_n(k) - \bar{X}</td>
<td>^4 \left( \frac{1}{N} \sum_k</td>
<td>X_n(k) - \bar{X}</td>
</tr>
</tbody>
</table>
Experimental comparison between fixed and adaptive schemes
- Fixed and adaptive window lengths (different sparsity measures)
- Sinusoid detection by main-lobe matching

Data
- Simulations: Two sound mixtures (Polyphony) and vibrato signal
- Real: Western pop (Whitney, Mariah) and Hindustani taans

Evaluation metrics
- Recall (%) and frequency deviation (Hz)
- Expected harmonic locations computed from ground-truth pitch

Results
1. Adaptive – higher recall and lower frequency deviation
2. Kurtosis driven adaptation is superior than other sparsity measures
Generalized music transcriptions system still unavailable

Solution [Wang08]
- Semi-automatic approach
- Application-specific design E.g. music tutoring

Two, possibly independent, aspects of melody extraction
- Voice pitch extraction – Manually difficult
- Vocal segment detection – Manually easier

Semi-automatic tool
- Goal: To facilitate the extraction & validation of the voice pitch in polyphonic recordings with minimal human intervention
- Design considerations
  - Accurate pitch detection
  - Completely parametric control
  - User-friendly control for vocal segment detection
Salient features
- Melody extraction back-end
- Validation
  - Visual: Spectrogram
  - Aural: Re-synthesis
- Segmental parameter variation
- Easy non-vocal labeling
- Saving final result & parameters
- Selective use of dual-F0 tracker
  - Switching between contours

A – Waveform viewer
B – Spectrogram & pitch view
C – Menu bar
D – Controls for viewing, scrolling, playback & volume control
E – Parameter window
F – Log viewer
CONCLUSIONS AND FUTURE WORK
Final system block diagram

Music Signal

Signal representation

DFT → Main-lobe matching → Parabolic interpolation

Sinusoids frequencies and magnitudes

Sub-multiples of sinusoids → F0 candidates

Multi-F0 analysis

TWM error computation → Sorting (Ascending) → Vicinity pruning

F0 candidates and saliences

Ordered pairing of F0 candidates with harmonic constraint → Joint TWM error computation

Nodes (F0 pairs and saliences)

Predominant F0 trajectory extraction

Optimal path finding

Predominant F0 contour

Optimal path finding

Voice Pitch Contour

Grouping

Classifier

Feature Extraction

Boundary Deletion

Harmonic Sinusoidal Model

Voicing Detector

Predominant F0 contour

Department of Electrical Engineering, IIT Bombay
CONCLUSIONS AND FUTURE WORK

Conclusions

- State-of-the-art melody extraction system designed by making careful choices for system modules
- Enhancements to above system increase robustness to loud, pitched accompaniment
  - Dual-F0 tracking for predominant-F0 extraction
  - Combination of static & dynamic, timbral & F0-harmonic features for voicing detection
- Fully-automatic, high accuracy melody extraction still not feasible
  - Large variability in underlying signal conditions due to diversity of music
- A priori knowledge of music and signal conditions
  - Male/Female singer
  - Rate of pitch variation
- High accuracy melodic contours can be extracted using a semi-automatic approach
CONCLUSIONS AND FUTURE WORK

Summary of contributions

- Design & validation of a novel, practically useful melody extraction system with increased robustness to pitched accompaniment
  - Signal representation
    - Choice of main-lobe matching criterion for sinusoid identification
    - Improved sinusoid detection by signal sparsity driven window length adaptation
  - Multi-F0 analysis
    - Choice of TWM error as salience function
    - Improved voice-F0 detection by separation of F0 candidate identification & salience computation
  - Predominant-F0 trajectory extraction
    - Gaussian log smoothness cost
    - Dual-F0 tracking
    - Final predominant-F0 contour identification by voice-harmonic instability
  - Voicing detection
    - Use of predominant-F0-derived signal representation
    - Combination of static and dynamic, timbral and F0-harmonic features
- Design of a novel graphical user interface for semi-automatic use of the melody extraction system
CONCLUSIONS AND FUTURE WORK

Future work

- **Melody Extraction**
  - Identification of single predominant-F0 contour from dual-F0 output
    - Use of dynamic features
  - F0 collisions
    - Detection – based on minima in difference of constituent F0s of nodes
    - Correction – allowing pairing of F0 with itself around these locations
    - Use of prediction-based partial tracking [Lag07]
  - Validation across larger, more diverse datasets
  - Incorporate predictive path-finding in DP algorithm
  - Extend algorithm to instrumental pitch tracking in polyphony
    - Homophonic music – Lead instrument (e.g. flute) with accompaniment
    - Polyphonic instruments (sitar)

- **Applications of Melody Extraction**
  - Singing evaluation & feedback
  - QBSH systems
  - Musicological studies
CONCLUSIONS AND FUTURE WORK

List of related publications

International Journals


International Conferences


CONCLUSIONS AND FUTURE WORK

List of related publications [contd.]

National Conferences


Patent

REFERENCES


