An Introduction to Music Signal Processing

Preeti Rao

Department of Electrical Engg., IIT Bombay

SAMP 2020
NITK
February 1 2020

Music Information Retrieval

Extraction of semantically meaningful information from the audio waveform (alternately, from symbolic scores) -> "content-based MIR".

Why?

Digital multimedia repositories abound and are growing rapidly….

YouTube has 86% searches related to music (Google has 70%!)

Can we improve the quality of search and navigation tools for music?

How?

Computing on music, as a structured acoustic signal with high-level musical attributes serving as the bridge to useful MIR tasks.

2

Making Music

Classifying instruments (*Natyashastra*) by the mode of sound production

- Stretched strings
- Membranes
- Solids
- Hollow (air)

Above all, is the "Human Body" instrument

- the most expressive instrument

Supremacy of the singing voice is characteristic of many cultures.

Vocal Music Diversity

3

Overview of the tutorial

I. Musical attributes and singing signal properties
II. Signal synthesis exercise*
III. Polyphonic music processing
IV. Vocal MIR applications
V. Review of MIR datasets and tools

*See:

- https://www.ee.iitb.ac.in/course/~daplab/resources/Readme
- https://www.ee.iitb.ac.in/course/~daplab/resources/Music%20synthesis%20tutorial.ipynb

4

Decoding music

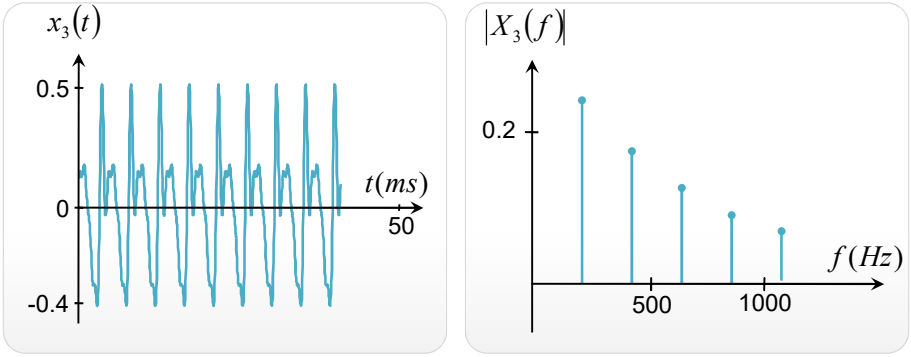
Interesting sounds are typically coded in the form of a temporal sequence of "atomic sound events".

speech -> a sequence of phones
music -> an evolving pattern of notes

An atomic sound event, or a single gestalt, can be a complex acoustical signal described by a set of temporal and spectral properties => an evoked sensation.

A "musical note": complex tone signal

F0 = 200 Hz



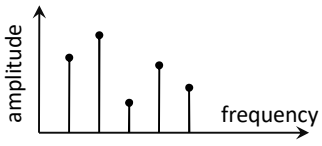
Sound and Sensation

Primary sensations	Physical correlates
○ ***loudness***	○ ***intensity***
○ ***pitch***	○ ***fundamental frequency***
○ ***timbre ("quality")***	○ ***Spectro-temporal properties***

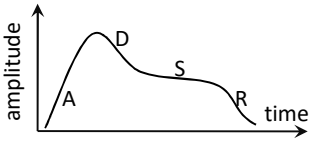
7

Timbre

- Tone "quality" or colour is linked with the identification of the instrument producing the tone. Influenced by the acoustics of the instrument + environment.
- A multidimensional attribute. Important dimensions are:
  - spectral envelope
  - time envelope "ADSR" => Attack, Decay, Sustain, Release
  - presence of irregularity and noise



spectral envelope

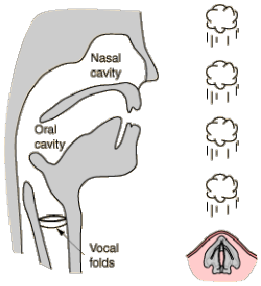


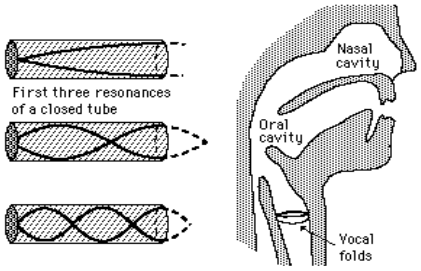
temporal envelope

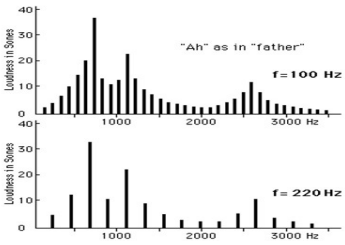
8

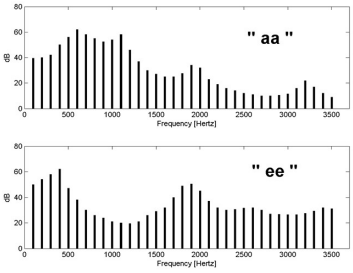
Human voice: vocal source and tract*

*HyperPhysics, Sound and Hearing, Georgia State University



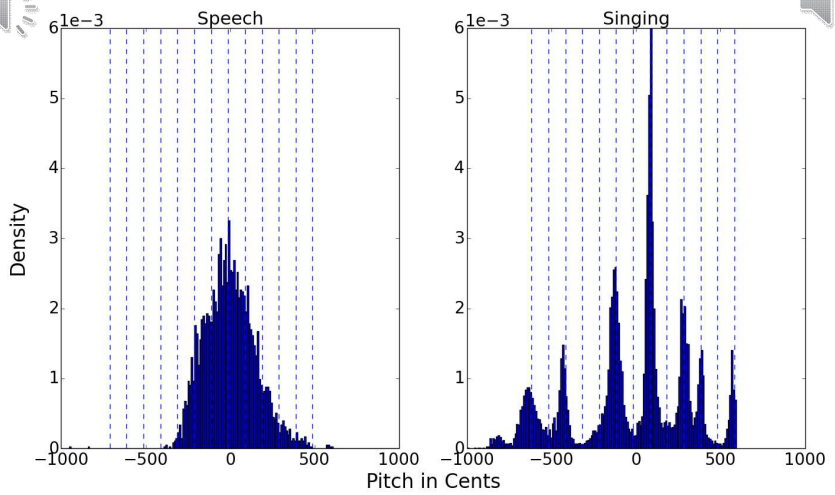






9

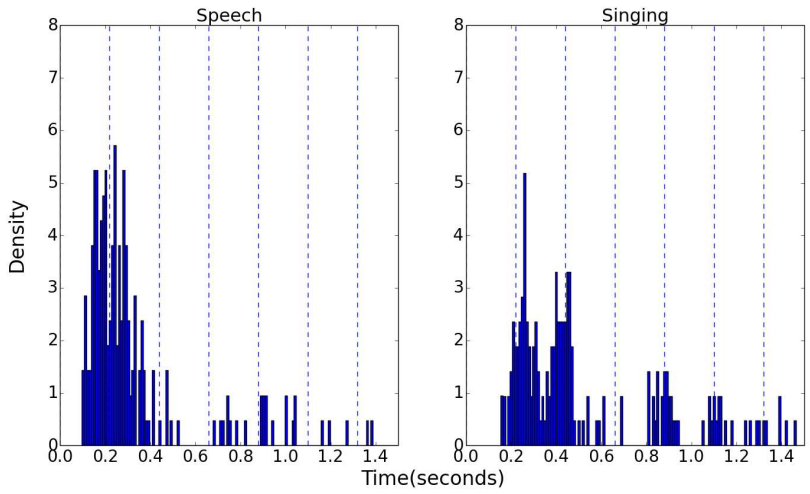
Speech vs Singing: pitch distribution



From NUS data corpus

10

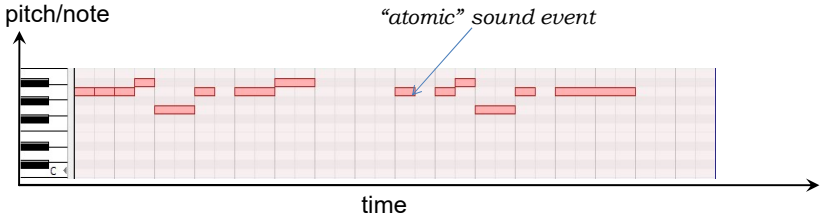
Speech vs Singing: syllable durations



11

Music

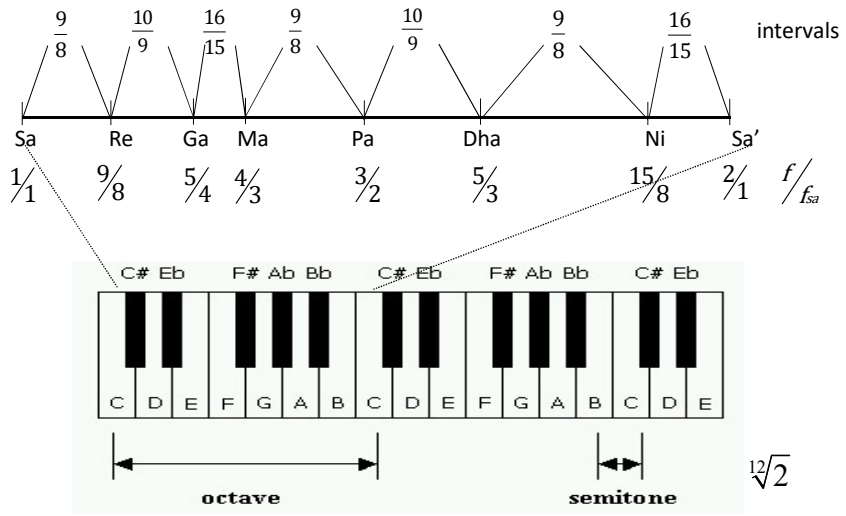
A sequence of tones each characterised by a **pitch** and **duration**.



"Universal language of music"

Music tends to use discrete **pitches** and **time intervals**.

A common musical scale: the "just intonation" scale



Returning to MIR … How to improve search?

By designing automatic systems that "understand" the audio signal.

The audio processing must be based on

- the characteristics of music (the acoustic signal properties important for perception and cognition)
- an appreciation of end applications that are relevant to listeners (since music similarity is multi-dimensional)

14

Motivation

Defining **Similarity**

An inherently ambiguous task

15

"Information Retrieval" from Speech

- Speech to text: Similarity in sequence of phones
- Language/accent id: Similarity in phone classes and articulation
- Speaker id: Similarity in physiology and articulation

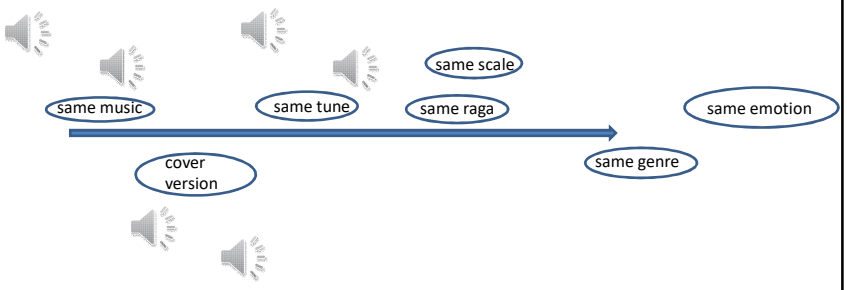
Perceived similarity -> acoustic features (read, mfcc!)

- We understand above notions of similarity. Next, let us consider similarity for music signals

16

Similarity is central

Querying for music… using audio examples



So, what is common across a pair?

17

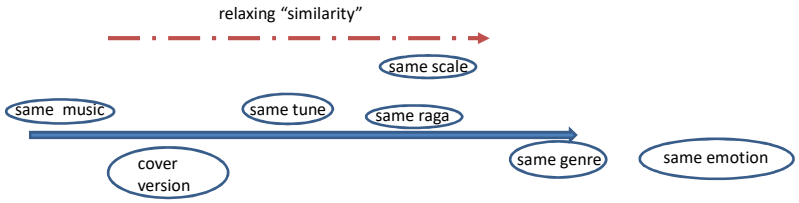
Audio waveforms look different

So, what is common across a pair?

- Not the waveform
- Frequency spectrum?
- Temporal variations?
- Mid-level musical dimensions including melody , rhythm, timbre …

18

Q: What is the property underlying the "match" in each case?



- Higher-level musical dimensions including melody , harmony, tempo, rhythm, timbre, lyrics, mood …..

19

Music concepts

Major dimensions of music for retrieval are melody, harmony, rhythm and timbre.

- Melody, harmony -> based on pitch content
- Rhythm -> based on timing information
- Timbre -> relates to instrumentation, texture

A representation of these high-level attributes can be obtained from pitch and timing information extracted by audio signal analysis.

Representations are then compared via a similarity measure to achieve retrieval.

20

Melody representation

- Melody is the temporal sequence of notes usually played by a single instrument (fixed timbre).
- **Implementation**:
  - Pitch detection is carried out on the audio signal at uniformly spaced intervals
  - Pitch sequence is segmented into notes (regions of relatively steady pitch)
  - Notes are labeled
  - Note patterns are matched to determine melodic similarity
- **Challenges**:
  - Note segmentation can be a difficult task
  - Melody detection in polyphonic music is tough

21

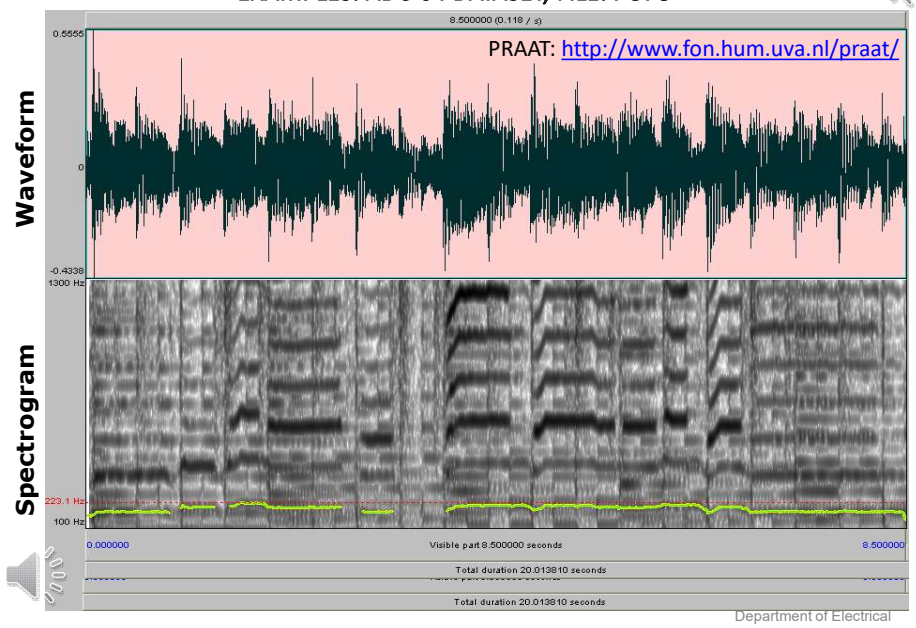
Defining melody

- Melody => predominant pitch trajectory (main voice)
- Pitch
  - Perceptual attribute related to the fundamental frequency (F0) of the voice
  - Can be detected via the periodicity of the waveform or the harmonic structure of the spectrum
- Time-lag, frequency-domain methods

ACF: $$r(n,k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]x[m+k]w[n-(m+k)]$$

FT: $$X(e^{j\omega}) = \sum_{k=0}^{N-1} 2\pi C(k)\delta\left(\omega - \frac{2\pi k}{N}\right) \quad \ldots(3.5)$$

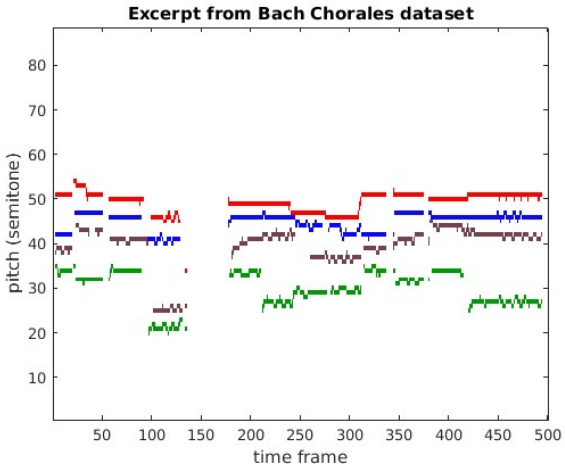
EXAMPLES: ADC'04 DATASET, FILE: POP3



Department of Electrical Engineering , IIT Bombay

Harmony

- The combination of simultaneously sounded musical notes to produce a pleasing effect. E.g. chords
- Choral singing: SATB are voices of different pitch ranges, timbre.



24

Rhythm

- Rhythm refers to the periodic and hierarchic framework that embeds the timing of events (onsets) within the audio signal.

- Rhythm detection involves detecting events at each **metrical level.**
  - Tatum
  - Tactus
  - Measure

- Onsets detected via abrupt increases in loudness, or abrupt energy changes within frequency bands.

- Rhythm is represented by the detected periodicities of the sequence of onsets.

25

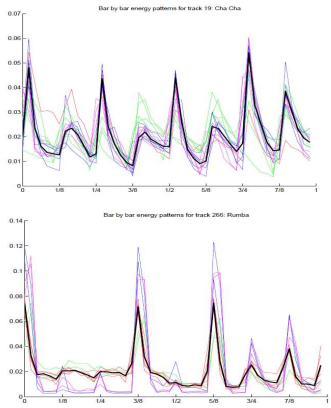
Rhythmic Patterns*

- Are important to music similarity

- IOI histogram and periodicity representations provide information about the relative frequency of various time intervals between events.

- Bar-length patterns of vectors of loudness, spectral centroid and MFCCs can be used with dynamic time warping to measure similarity.

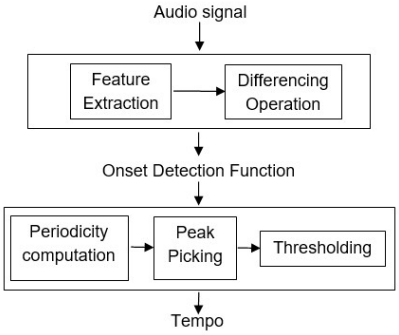
Top: Cha-Cha; Bottom: Rumba

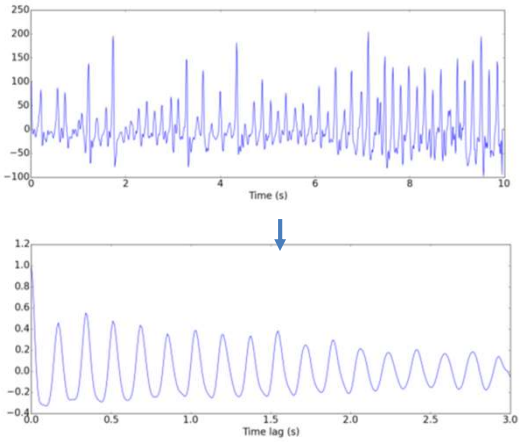


*From: Dixon, S., Gouyon, F., & Widmer, G., Proc. ISMIR 2004

26

Tempo Detection





27

Timbre

Timbral content of music changes with the instrument mix. E.g. the insertion of a new instrument.

It is represented by the frame-level spectrum shapes, and the temporal and spectral dynamics.

28

Multiple sources

Multiple sources (instruments) in an auditory scene may be visualized in a time-frequency representation. Sources are "sparse", and vocals dominant.

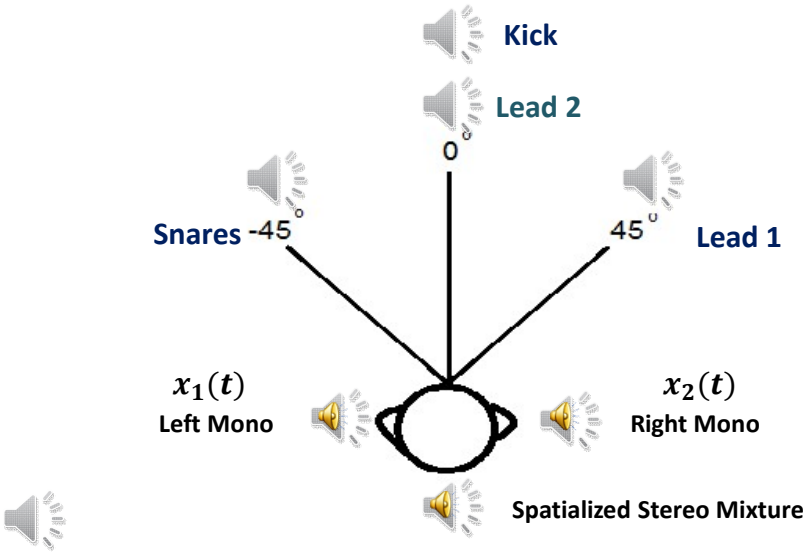
Challenges arise however due to:

- quality of t-f representation governed by t-f trade-off
- applying gestalt/grouping principles computationally

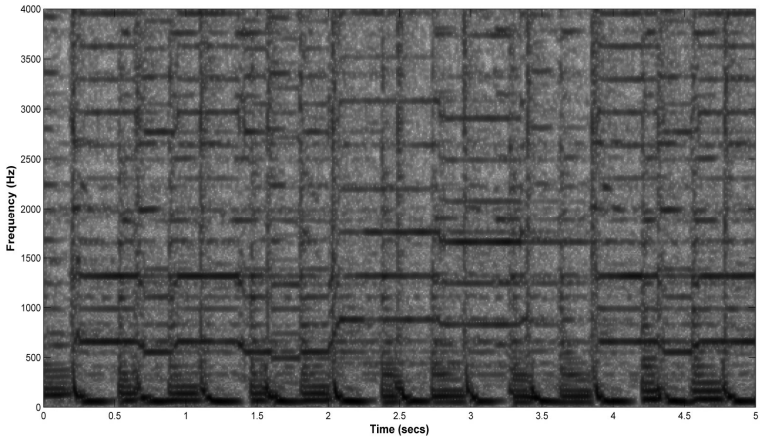
A number of methods of analyzing the t-f representation have emerged in order to:

- estimate the properties of individual sources, or
- achieve the separation of sources

Commercial audio recording



Single-channel mix



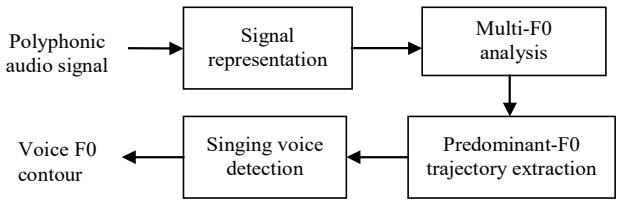
MIR task: Find the high-level attributes (melody, rhythm, etc.) of each source.

Spectrograms are with PRAAT: http://www.fon.hum.uva.nl/praat/

Melody extraction

- Detect and track the (harmonic) components of the melody voice in the (sparse) representation.
- Challenges
  - **Polyphony** => finding the dominant voice in a crowded spectrum
  - **Rapidly varying pitch** => detecting and localising harmonics within short windows
- Approach
  - Obtain a signal representation with the required frequency and time resolutions
  - Means to identify the dominant voice

Predominant melody detection in polyphony*



A. Signal representation: sinusoidal peak detection with signal-driven window length adaptation

B. Multi-F0 detection: exhaustive multiple harmonic-spectrum fits

C. Predominant-F0 trajectory: maximise pitch salience + smoothness

D. Singing voice detection: use timbral properties to distinguish main voice

*From V. Rao & P. Rao, IEEE TASLP, 2010

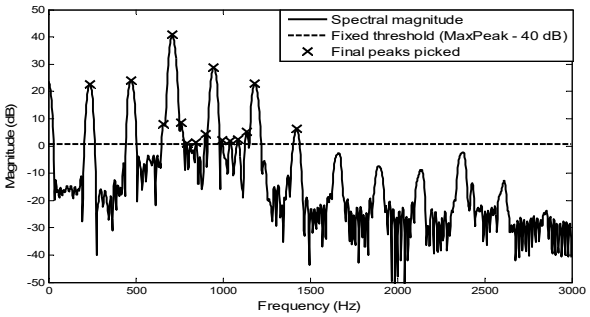
A. Signal representation

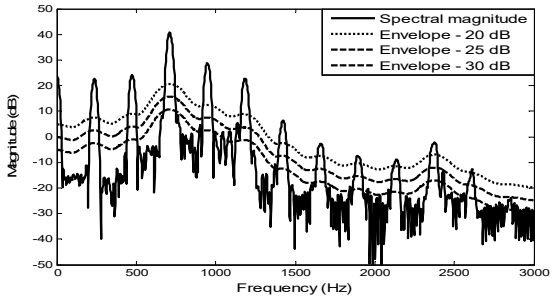
Sinusoidal model versus STFT

- STFT is a very high dimensional representation. The harmonic spectrum is sparse. Each harmonic is an amplitude- and frequency-modulated sinusoid.
- Post-processing the STFT in an optimised t-f trade-off to obtain the sinusoidal representation and detect the sinusoid parameters.
- Sinusoidal representation is compact and facilitates extraction of perceptual attributes such as pitch and timbre.
- The extracted attributes can also be used, possibly with an instrument model, to modify the sound or to separate concurrent sounds.

34

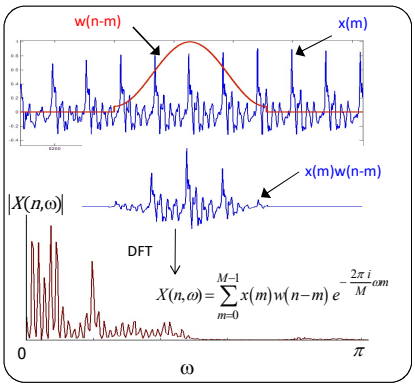
Magnitude spectrum peak detection: thresholding



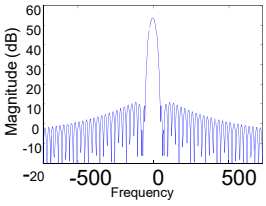


Detecting instrument harmonics: sinusoidal peak picking

Short-time Fourier analysis



Main lobe shape based matching



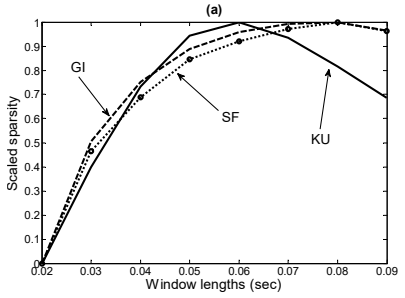
Choosing the window length

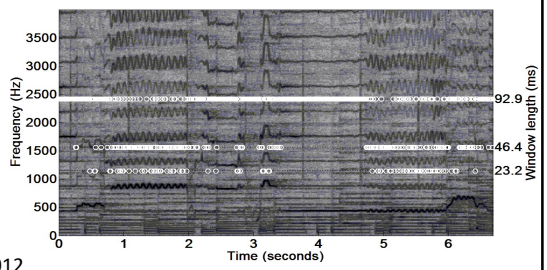
Closely-spaced components in the polyphony => need for higher frequency resolution = ***longer*** windows

Pitch variation with time can be rapid in ornamented regions => need for better time resolution = ***shorter*** windows

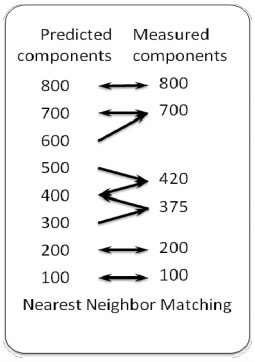
Signal driven window length adaptation*

- Adapting window length based on spectrum sparsity:
  - a sparse spectrum is more "concentrated" => better represented sinusoidal components
  - Window length selection (20, 30, 40 ms) based on maximizing signal sparsity

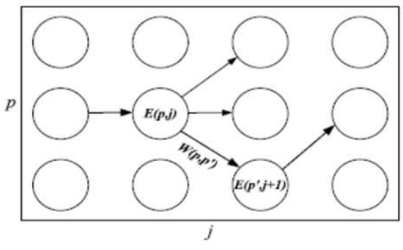




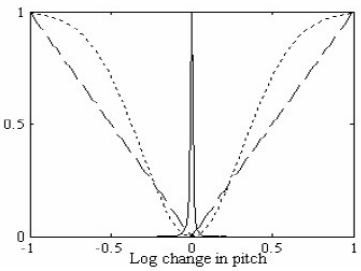
*From: Rao , Gadipatti & Rao, IEEE TSALP, 2012



B. Multi-F0 analysis (emulates central auditory processing)

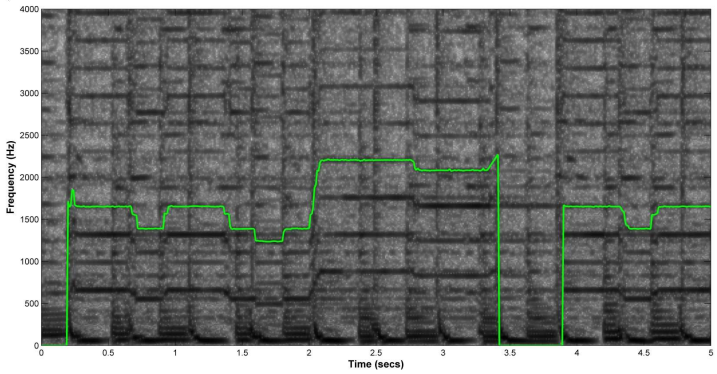


C. Predominant F0 detection using DP path optimization (emulates auditory grouping)

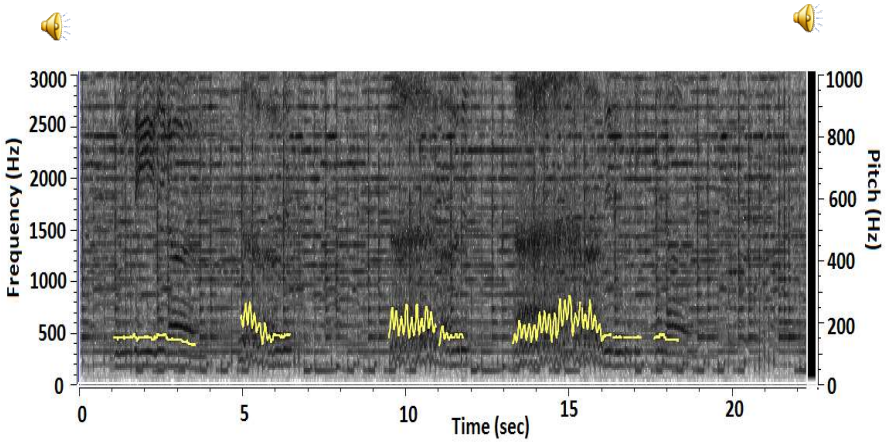


Smoothness cost

Single-channel signal with detected predominant pitch (vocal)

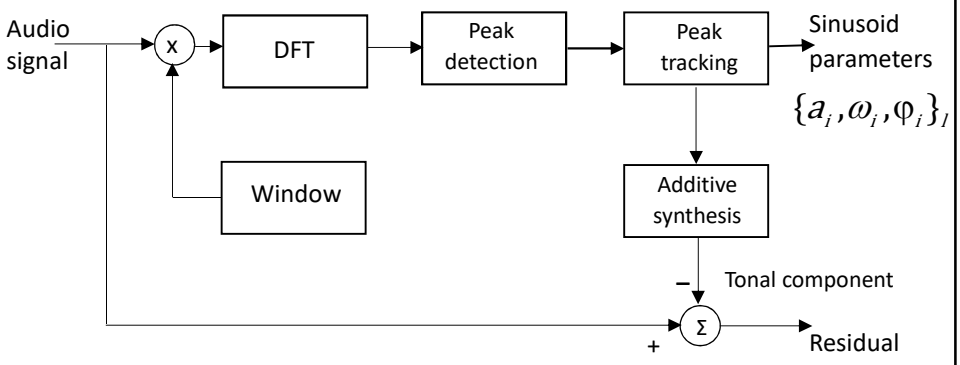


Ornamented singing



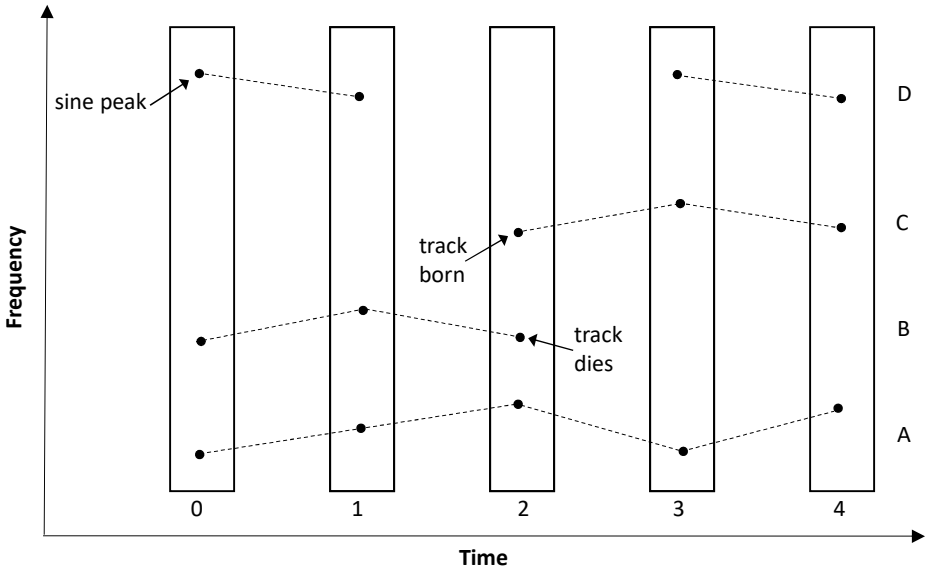
Bhimsen Joshi, Marwa, Tintal
Bandish: Guru Bina Gyan Na Pave

Source separation with sinusoidal modeling



For the smooth evolution of the signal, sine components are detected in each frame and linked to tracks from the previous frame based on frequency proximity.

Peak tracking



Shortcomings of Harmonic Sinusoidal Modeling

- Need for accurate pitch tracking for correct reconstruction of highly time-varying segments, and the subsequent total suppression in the background music.
- All the energy at a harmonic location is assigned to the target source => spectral distortion of target source with "bleed in" of accompanying instruments. Arises from the binary masking. A soft mask would work better if we had prior estimate of one of the components (vocal or background).

43

MIR tasks for Singing

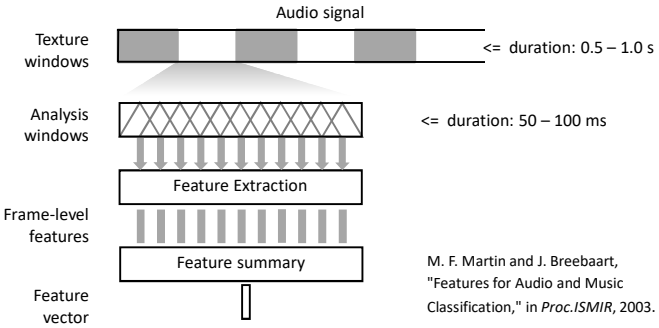
- Singing voice detection
- Lyric alignment
- Key/mode/raga identification
- Singing style/genre recognition
- Singing skill evaluation
- Structural segmentation

- Vocal timbre analysis
- Singing synthesis

44

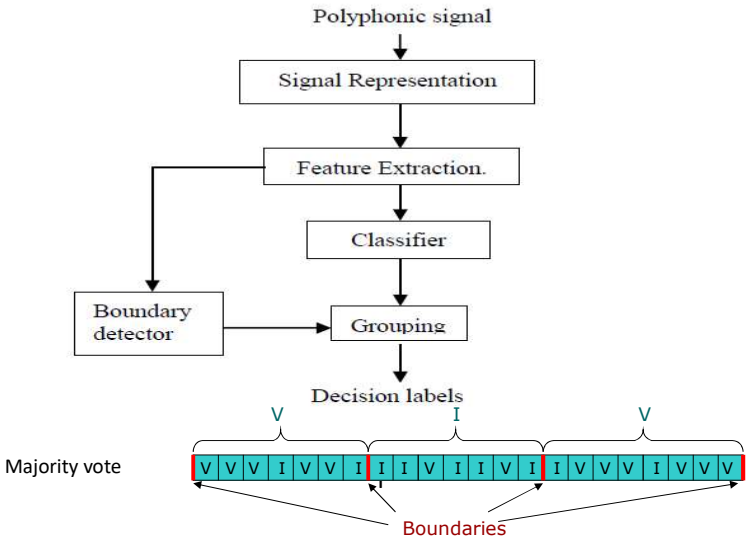
Feature extraction

- The temporal pattern of frame-level features can offer important cues to signal identity



45

Singing voice detection using Timbre characteristics



EXPERIMENTS

Cross-cultural Datasets*

Polyphonic music with lead vocal and pitched instrumental background

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

*From: V. Rao, C. Gupta and P. Rao, LNCS 7836, Springer 2011

EXPERIMENTS

Cross-cultural Datasets

Genre	*Singing*	*Dominant Instrument*
I Western	Syllabic. No large pitch modulations. Voice often softer than instrument.	Mainly flat-note (piano, guitar). Pitch range overlapping with voice.
II Greek	Syllabic. Replete with fast, pitch modulations.	Equal occurrence of flat-note plucked-string /accordion and of pitch-modulated violin.
III Bollywood	Syllabic. More pitch modulations than western but lesser than other Indian genres.	Mainly pitch-modulated wood-wind & bowed instruments. Pitches often much higher than voice.
IV Hindustani	Syllabic and melismatic. Varies from long, pitch-flat, vowel-only notes to large & rapid modulations.	Mainly flat-note harmonium (woodwind). Pitch range overlapping with voice.
V Carnatic	Syllabic and melismatic. Replete with fast pitch modulations.	Mainly pitch-modulated violin. F0 range generally higher than voice but has some overlap in pitch range.

EXPERIMENTS

Proposed features

- Distinct static and dynamic feature sets
- Features extracted using a harmonic sinusoidal model representation

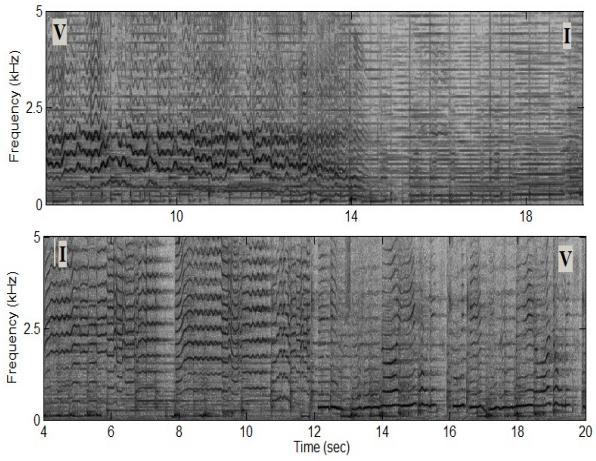
C1 *Static timbral*	*C2* *Dynamic timbral*	*C3* *Dynamic F0-Harmonic*
F0	Δ 10 Harmonic powers	Mean & **median** of **ΔF0**
10 Harmonic powers	Δ SC & Δ SE	Mean, median & Std.Dev. of ΔHarmonic ε [0 2 kHz]
Spectral centroid (SE)	**Std. Dev.** of **SC** for 0.5, **1** & 2 sec	Mean, median & Std.Dev. of ΔHarmonic ε [2 5 kHz]
Sub-band energy (SE)	**MER** of **SC** for 0.5, 1 & **2** sec	Mean, median & Std.Dev. of ΔHarmonics 1 to 5
	Std. Dev. of **SE** for **0.5**, 1 & 2 sec	Mean, **median**, Std.Dev. of **ΔHarmonics 6:10**
	MER of SE for 0.5, 1 & 2 sec	Mean, **median**, Std.Dev. of **ΔHarmonics 1:10**
		Ratio of mean, median & Std.dev. of ΔHarmonics 1 to 5 : ΔHarmonics 6 to 10

FEATURE EXTRACTION

Timbral statics (C1), dynamics (C2) and **F0 dynamics** (C3)

North Indian Classical Music (Hindustani) C1+C3

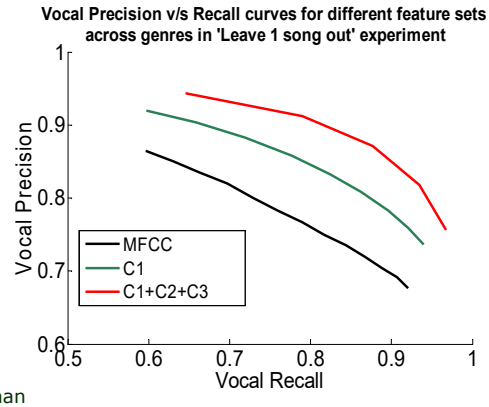
Indian film music (Bollywood) C1+C2



EXPERIMENTS

Results*

- Cross-validation experiment
  - Leave 1 song out
- Feature combination
  - Concatenation
  - Classifier combination
- Baseline features
  - 13 MFCCs [Roc07]
- Evaluation
  Recall (%) and precision (%)
- Overall Results
  - C1 better than baseline
  - C1+C2+C3 much better
  - Classifier combination better than feature concatenation



*From: V. Rao, C. Gupta and P. Rao, LNCS 7836, Springer 2011

Lyrics Alignment

- Achieved via singing voice separation followed by forced alignment with acoustic phone models adapted to singing voice.
- Challenges
  - Residual interference and artifacts from separation
  - Singing style influence on phone acoustics
  - Lack of matched training data

52

Melodic similarity: Applications

- Detecting similarity in tune, key or raga from the vocal melody. Types of cues:
  - distributional (hierarchy of notes)
  - structural (phrases, motifs)
- Genre or style recognition: singing style influences the melodic contour for the same tune or tonal material
- Singing skill assessment
- "Mood" for playlist creation

53

Key profiles for Western music

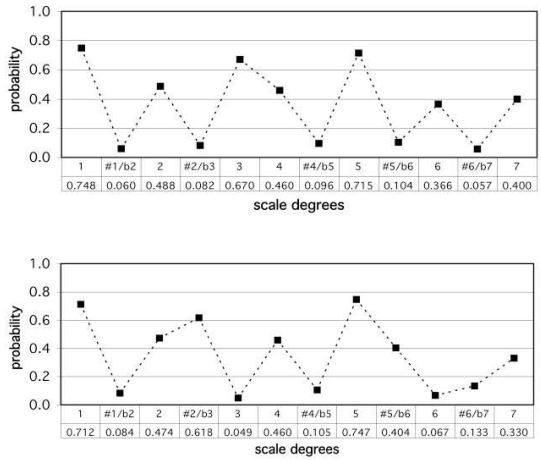
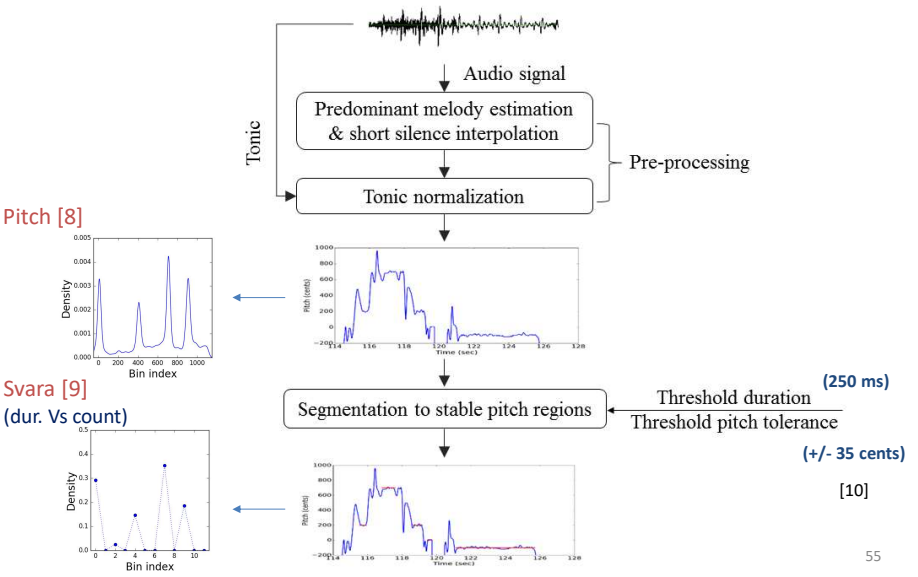


Figure 1. Key-profiles for major keys (above) and minor keys (below).

From Temperly, A Bayesian key-finding model, in Music and AI, Springer, 2002.

54

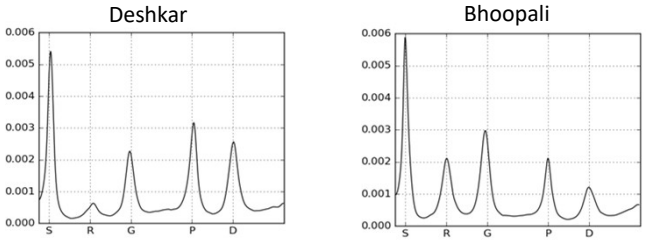
Raga (or key) identification from pitch distribution

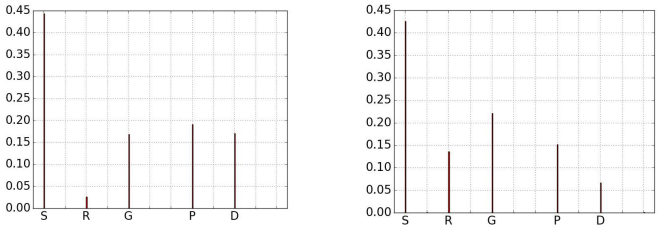


55

Discriminating "allied" ragas*

Tonal material: S R G P D

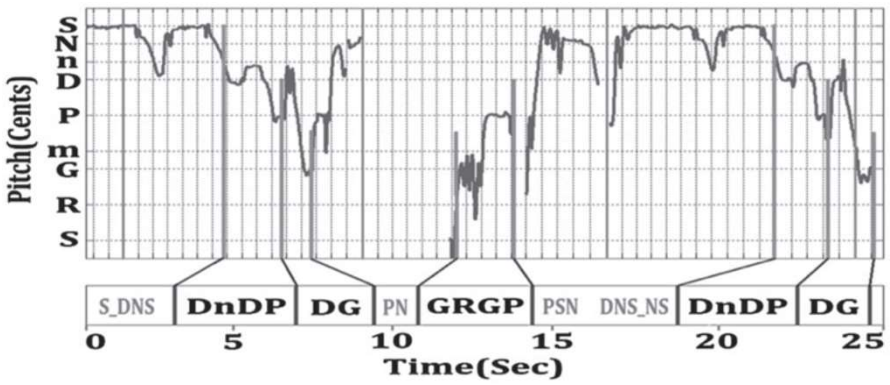




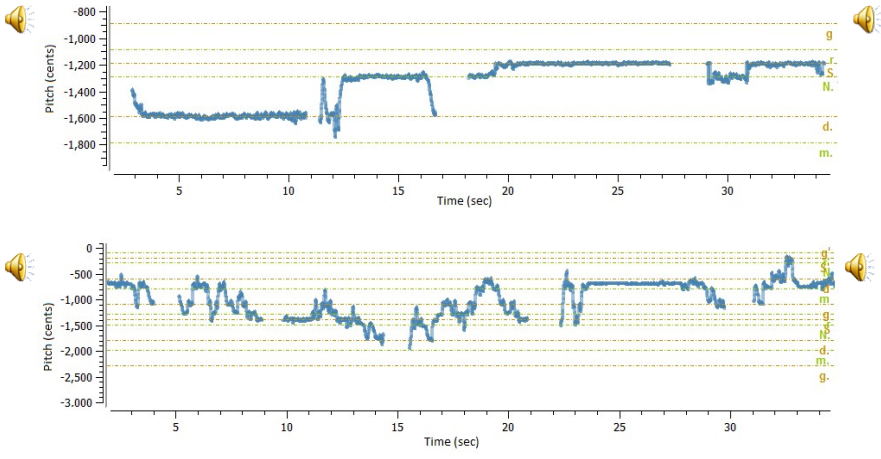
56

*From: Ganguli, K.K. and Rao, P., Proc. ISMIR 2017

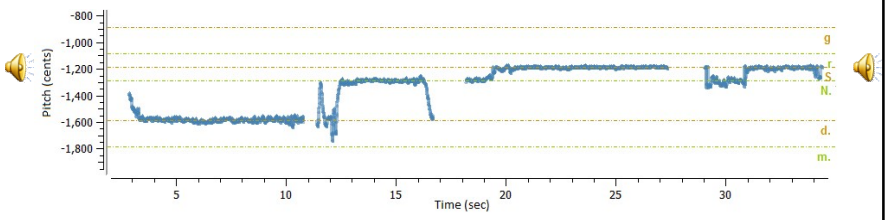
Raga phrases: note labelling?



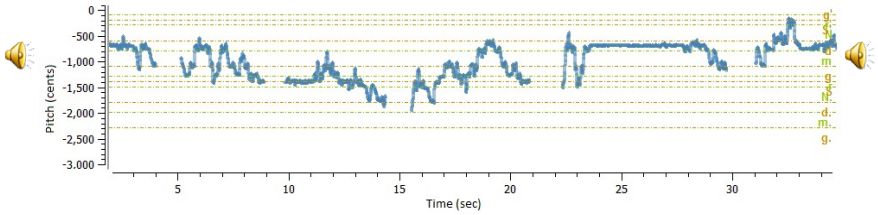
Style variations



Style variations*



Raga Todi: Malini Rajurkar



Raga Subhapanthuvarali: M S Subhalaxmi

*From: A. Vidwans, K. Ganguli and P. Rao, Proc. Compmusic Workshop, 2012

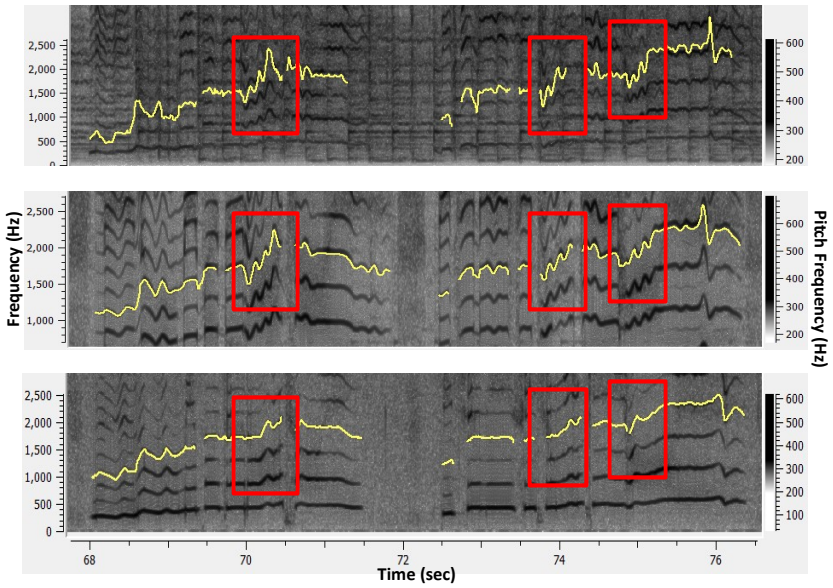
Assessing Vocal Performance

Various performance parameters:

- mean pitch for each note and interval sizes in cents; and vibrato rate and depth.
- inter-onset intervals between notes; tempo information;
- relative dynamic level between notes.

60

Model versus Trained, Amateur Singers



61 of 50

Observed differences*

- The “notes” rendered by both, trained and amateur singers, are similar to the reference
- Observe the difference in the rendition of **ornaments**

Model Singer Trained Singer Amateur Singer

*From: C. Gupta and P. Rao, CMMR/FRSM 2011, Springer LNCS 7172

62 of 50

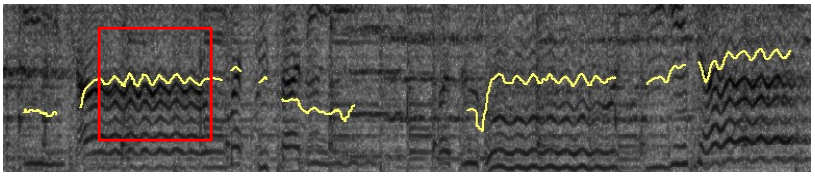
Ornamentation

- Style, emotions, *gharana (school/style) characteristics, raga characteristics and even the personal characteristics influence the ornaments.*
- **Ornaments** enhance the basic melodic contour and contribute to musicality/ expressiveness.

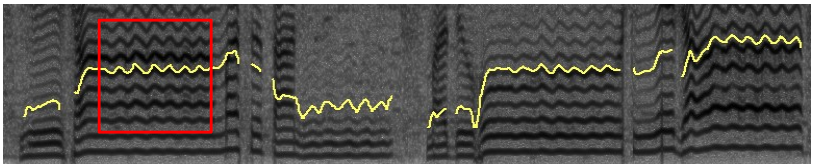
  e.g. grace note, a glide between two notes, multiple oscillations of a single note, oscillation between notes….
- Of the ornaments, those that are often transcribed in notation are *Meend, Andolan, Khatka, Murki, Gamak, Zamzama*

  [**ITC SRA site:** http://www.itcsra.org/alankar/alankar.html]

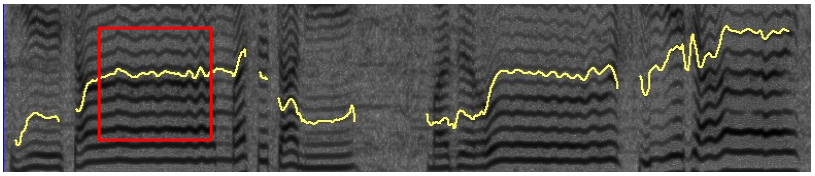
Ornament : Vibrato



Model Singer



Good Singer



Amateur Singer

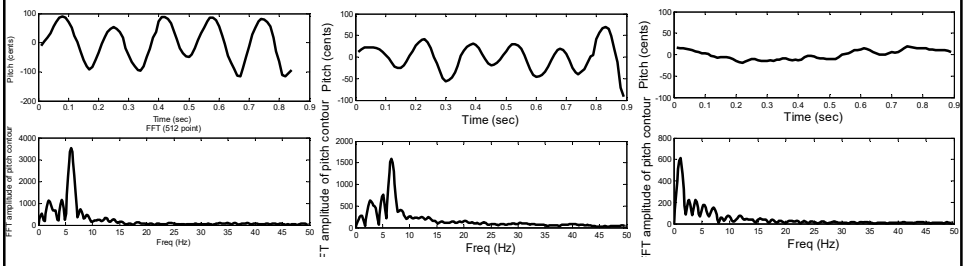
Vibrato : Modelling [Nak2006],[Ami2009]

- **Vibrato :** a deliberate, periodic fluctuation of pitch that can be parameterized by its **rate** (the number of vibrations per second) and **extent** (the amplitude of vibration from an average pitch on the vibrato section)
- Empirically, vibrato commonly has a periodicity of **5 – 8 Hz** and extent range is **30 – 150 cents**

Vibrato : Objective Measure

- **Relative Energy Ratio (RER)** [Ami2009]

$$\text{Energy Ratio, } ER = \frac{\sum_{k=k_{4.5Hz}}^{k_{7.5Hz}} |Z(k)|^2}{\sum_{k=k_{2Hz}}^{k_{10Hz}} |Z(k)|^2} \qquad RER = \frac{ER_{Test}}{ER_{Ref}}$$



Reference Singer

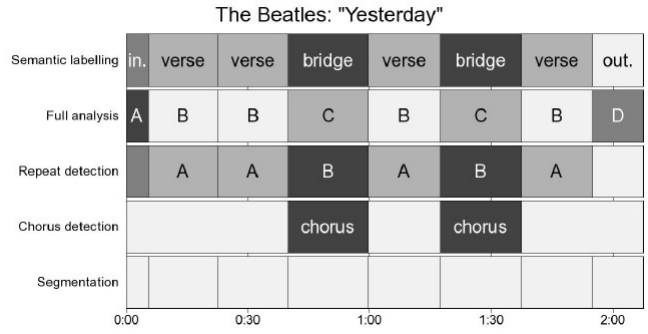
Good Singer
RER = 0.97

Bad Singer
RER = 0.4

Structural Segmentation

- At large time scales, music is highly structured.
- The segmentation may be based on
  - Melody/harmony,
  - Timbre
  - Rhythm
- Homogeneity and Repetition
- Applications: Browsing, Audio summary generation

67



Yesterday, all my troubles seemed so far away
Now it looks as though they're here to stay
Oh, I believe in yesterday

Suddenly, I'm not half the man I used to be
There's a shadow hanging over me.
Oh, yesterday came suddenly

Why she had to go I don't know she wouldn't say
I said something wrong, now I long for yesterday

Yesterday, love was such an easy game to play
Now I need a place to hide away
Oh, I believe in yesterday

Why she had to go I don't know she wouldn't say
I said something wrong, now I long for yesterday

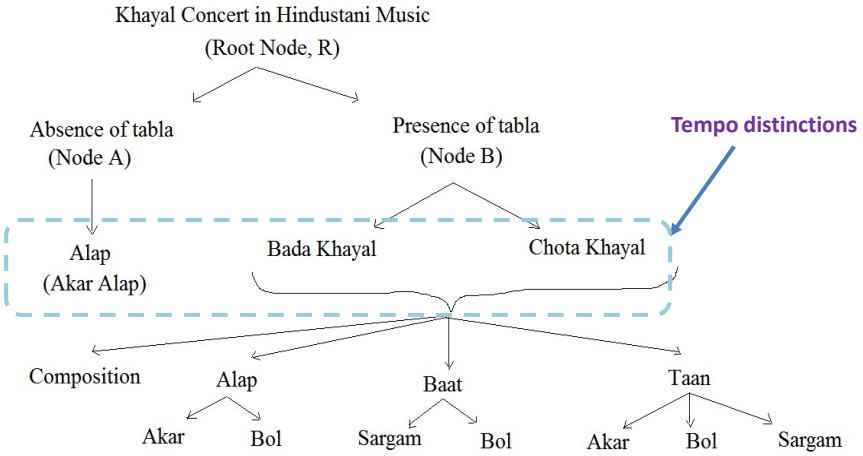
Yesterday, love was such an easy game to play
Now I need a place to hide away
Oh, I believe in yesterday

Mm mm mm mm mm mm mm

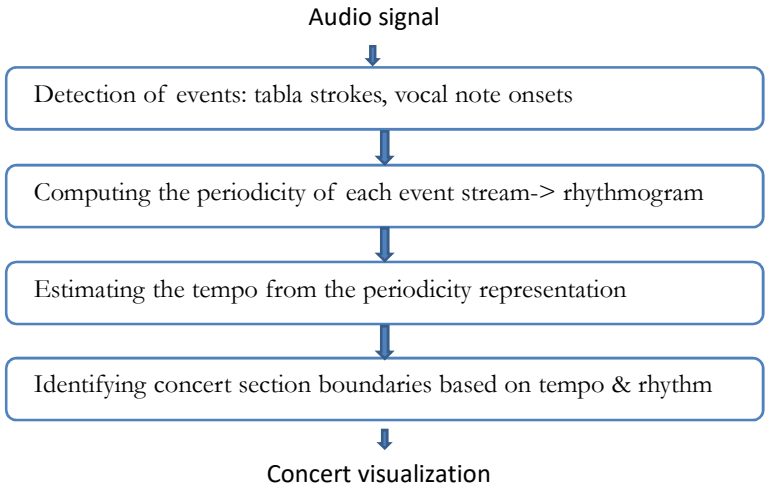
68

Hindustani Concert Segmentation

Khayal Vocal Concert Structure



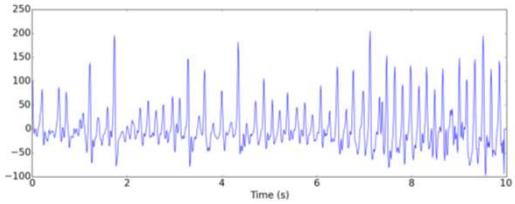
Concert segmentation with rhythm detection*



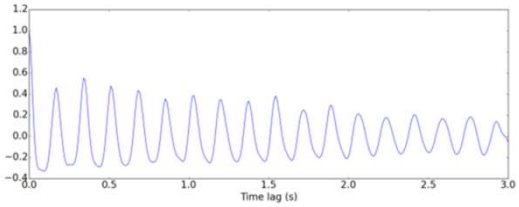
*From: T.P.Vinutha, S. Suryanarayana, K. K. Ganguli and P. Rao, Proc. ISMIR 2016

Tempo Detection

Onset detection function

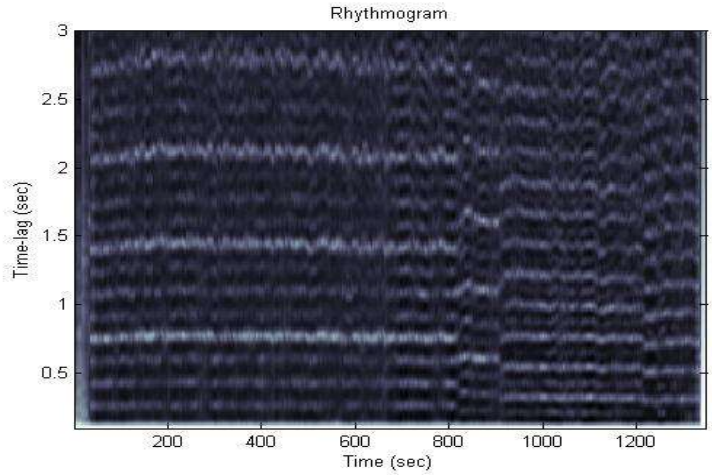


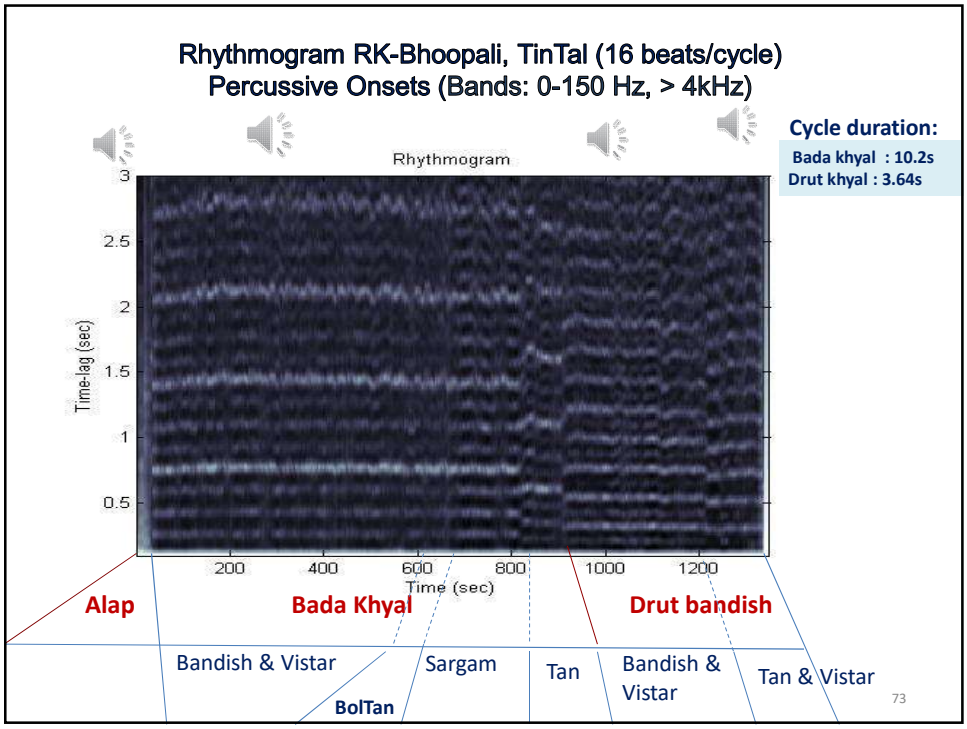
Autocorrelation function



71

Rhythmogram RK-Bhoopali, TinTal (16 beats/cycle)
Percussive Onsets (Bands: 0-150 Hz, > 4kHz)





Conclusion

- Understanding the acoustic and structural aspects of music is important to developing signal processing algorithms for music applications.
- Time-frequency representations play a key role in the detection and similarity modeling of music attributes.
- Human listeners remain superior to machines in nearly all tasks. This drives continuing research on modeling perception.
- Modeling a-priori knowledge effectively is crucial to success on the more challenging problems where complex decisions are involved and expert knowledge can help, or where large datasets are not available.

The future… Perceptual AI*

- MIR systems on audio signals are seeing a plateauing of performance in recent years.
- Under-performing MIR systems are likely a result of deficiencies in feature representation rather than in the classifier.
- Traditional "hand-crafted" signal-level attributes to be replaced by feature learning with deep processing architectures.
- Integrating features at various time scales needs to be dealt with.
- Potential of unsupervised learning to deal with shortage of labeled training data.

*Humphrey, Bello and LeCun, J. Intell. Inf. Sys, 2013

75

MIREX

Music **I**nformation **R**etrieval **E**valuation e**X**change

- Homepage:http://www.musicir.org/mirex/wiki/MIREX_HOME
- Annual evaluation campaign for MIR algorithms, coupled to the ISMIR conference
  - A set of community-defined formal evaluations through which a wide variety of state-of-the-art systems, algorithms, and techniques are evaluated under controlled conditions.
- Reference: J. Stephen Downie. "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," Acoustical Science and Technology Japan, 29(4), pp. 247-255, 2008.

76

Mailing lists

- MIREX: evalfest@lists.lis.illinois.edu
- ISMIR: community@ismir.net
- WiMIR: https://groups.google.com/forum/#!forum/wimir
- CompMusic friends: compmusic-friends@llista.upf.edu
- Auditory: auditory@lists.mcgill.ca
- IRCAM: music-ir@listes.ircam.fr
- Google Magenta: https://groups.google.com/a/tensorflow.org/forum/#!forum/magenta-discuss

A good summary: http://www.justinsalamon.com/music-technology.html

77

MIR Toolkits

- Essentia library: http://essentia.upf.edu/documentation/
- LibROSA library: https://github.com/librosa/librosa
- MIR toolbox: https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox
- SMS tools: https://www.upf.edu/web/mtg/sms-tools
- Melodia: https://www.upf.edu/web/mtg/melodia
- SM toolbox: https://www.audiolabs-erlangen.de/resources/MIR/SMtoolbox/
- Summary: https://www.ismir.net/software-tools.php

78

Datasets

- RWC music database: https://staff.aist.go.jp/m.goto/RWC-MDB/
- CompMusic datasets: http://compmusic.upf.edu/datasets
- Million song dataset: https://labrosa.ee.columbia.edu/millionsong/
- Dunya research corpora: http://compmusic.upf.edu/corpora
- MusicBrainz metadata repository: https://musicbrainz.org/

79

Thank you

Acknowledgements

My students at I.I.T. Bombay

Funding from CompMusic ERC project from UPF Barcelona (2011-2016)

81