ISMIR Lecture Notes

# Signal Processing Methods for Sound Recognition

Dr. E. Benetos

Conferences – ISMIR 2005, ICA 2007, MPEG 2009, CMMR 2012, WAC 2017, Audio Mostly 2017, SysMus 2017

Software – Sonic Visualiser (+Vamp plugins), SoundBite, Tony

**Machine Listening for Sound Scenes**

* Sound scene recognition
* Sound event detection
* Source separation
* Noise monitoring/reduction

**Sound Event Detection ( SED )**

Main goal: identify labels and start/end times for sound events

**Matrix Factorization**

Spectrogram ( frequency-time representation ) – treat as a matrix and factorize into two matrices

NMF ( Nonnegative Matrix Factorization)

VFxT = WFxK x HKxT

K refers to the rank of the matrices. Can be thought of as the number of con-current sound events you want to detect.

Nonnegative matrices – W (dictionary of templates containing signatures of events) , H ( activation matrix)

PLCA – Probabilistic Latent Component Analysis

* Decompose a bivariate distribution over frequency and time
* Probabilities of frequencies over particular component ( W )
* Probabilities of component over time ( H )

**Recognizing multiple overlapping sounds**

Mel and ERB spectrograms – they are essentially different filterbanks

Mel- frequency cepstrum : representation of the short term power spectrum of a sound.

* P(t) : spectrogram energy ( known quantity )
* First term : spectral template for event class s , exemplar c, and sound state q ( fixed, pre-extracted )
* Second term : event activation over time
* Third term : exemplar contribution for each event class, over time
* Fourth term : sound state contribution for each event class, over time. Controlled by an event HMM

Application source code : <https://code.soundsoftware.ac.uk./projects/sound-event-detection-plca>

Tracking sequences over time using Hidden Markov Models ( discrete ) or a State Space Model ( SSM )

**Linear Dynamical System** is a state space model where all the conditional probability distributions are linear-Gaussian.

* Given some output you can infer information about some hidden variables/features.
* The hidden variables are linked by a transition matrix

Approach : Use LDS to track multiple concurrent sound events

Motivation : Assume the posteriogram is the observation in an LDS, with the latent states corresponding to the ‘clean’ posteriogram. Essentially clean up the model which was unsatisfying and noisy.

**Future perspectives**

* Use of context to better inform models
* Sound event taxonomy
* Language/ library modeling

# Automatic Music Transcription by Dr. Emmanouil Benetos

Dr. E. Benetos

**Introduction**

Mid-level and parametric representation:

* Pitch onset, offset, stream, loudness
* Audio time(s)

Notation:

* Note name, key, rhythm, instrument
* Uses score time (beat)

Prescriptive notation – an interpretation of the performance

Descriptive notation – provides the specifics of the performance

**Subtasks**

* Pitch detection
* Onset/offset detection
* Instrument identification
* Rhythm parsing
* Identification of dynamics/expression
* Typesetting

**Challenges**

* Music sources are not independent. The events are strongly dependent
* Inferring musical attributes from a mixture is an extremely underdetermined problem
* Data annotation is time consuming, leading research to be constrained towards specific tasks. Eg. Pianos with sensors in the keys (disk klavier)
* State of the art methods really struggle with symphonic pieces.
* The large variety of music

**Concurrent sound sources**

Concurrent sound sources interfere with each other.

* C4 (46.7%), E4 (33.64%), G4(60%)

The fundamental frequency identifies the notes but there are also frequencies present at integer multiples of the fundamental frequency.

**Consonant intervals and the problems they pose**

Octave - A2 = 110 Hz, A3 = 220 Hz

Perfect Fifth – A2 = 110Hz, E3 = 165Hz

Overtone series …



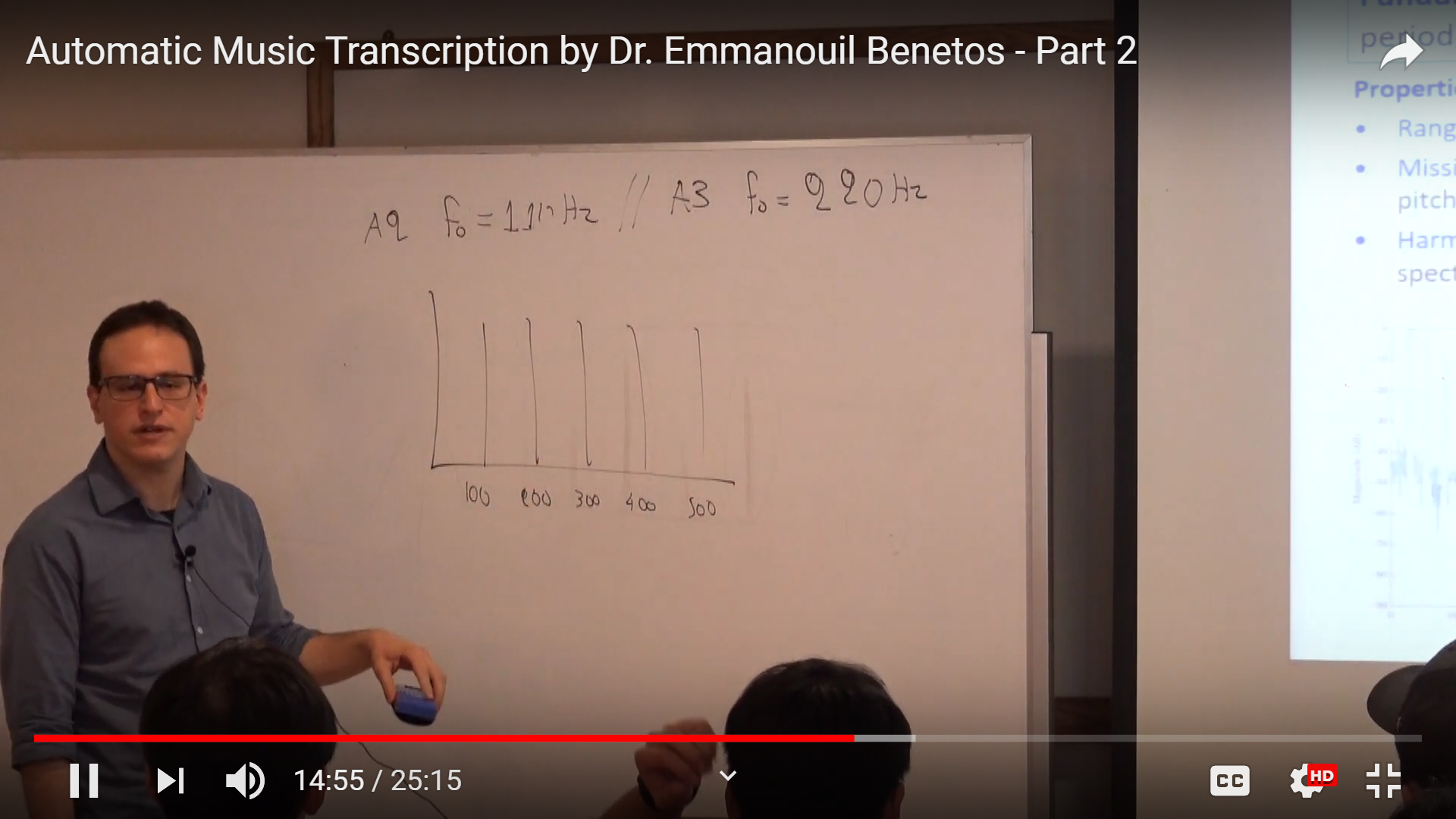
Source: E. Benetos, ‘Automatic Music Transcription’, Tutorial at the Department of Computer Science, National University of Singapore, available: <http://c4dm.eecs.qmul.ac.uk/nus-amt-tutorial/> [Accessed 19/8/2019]

And so on ……

How do we determine what note the energy is coming from?

**Pitch Perception**

* A sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at 40dB SPL.
* **People hear pitch on a logarithmic scale !**
* Pitch is salient in the range from 30Hz-5kHz
* The fundamental frequency does not need to be present for a pitch to be perceived
* For a sound with harmonic partials to be heard as a tone, its spectrum must include at least 3 successive harmonics of a common frequency



Source: E. Benetos, ‘Automatic Music Transcription’, Tutorial at the Department of Computer Science, National University of Singapore, available: <http://c4dm.eecs.qmul.ac.uk/nus-amt-tutorial/> [Accessed 19/8/2019]

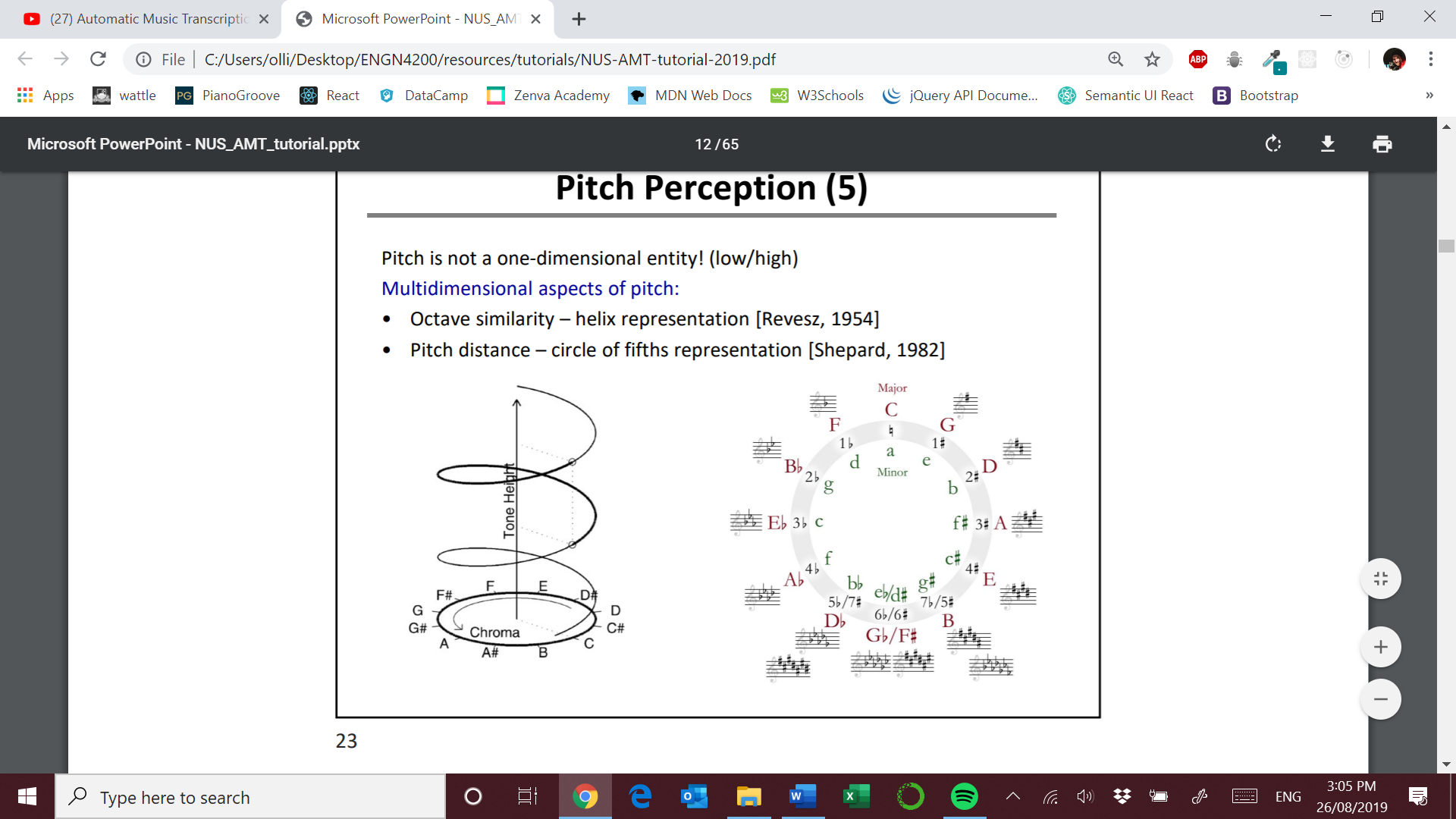
Interestingly the lower notes on the piano are missing the lower harmonics!

Some instruments are quasi-harmonic ( i.e. The partials are shifted slightly )

Pitch perception theories inform the creation of AMT systems: pattern matching and autocorrelation models. Slide 22 of the presentation provides some references to some pitch perception models.

**Multi-dimensional aspects of pitches**

* Octave similarity – helix representation
* Circle of fifths - pitch distance



Source: E. Benetos, ‘Automatic Music Transcription’, Tutorial at the Department of Computer Science, National University of Singapore, available: <http://c4dm.eecs.qmul.ac.uk/nus-amt-tutorial/> [Accessed 19/8/2019]

**Human Transcription**

Implicitly humans are very good at:

* Style detection
* Instrument identification
* Beat tracking