CS4347

Sound and Music Computing

L3: Introduction to Music Representation, Analysis and Transcription

Wang Ye

wangye@comp.nus.edu.sg

Office: AS6-04-08

Topics to Cover (*selective approach*)

Part A: The Core

- Introduction
- Review of DFT, Audio Representation, and Machine Learning
- Music Representation, Analysis and Transcription
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)
- Generative Models for Text-to-Speech (TTS) & Singing Voice Synthesis (SVS)

Midterm break

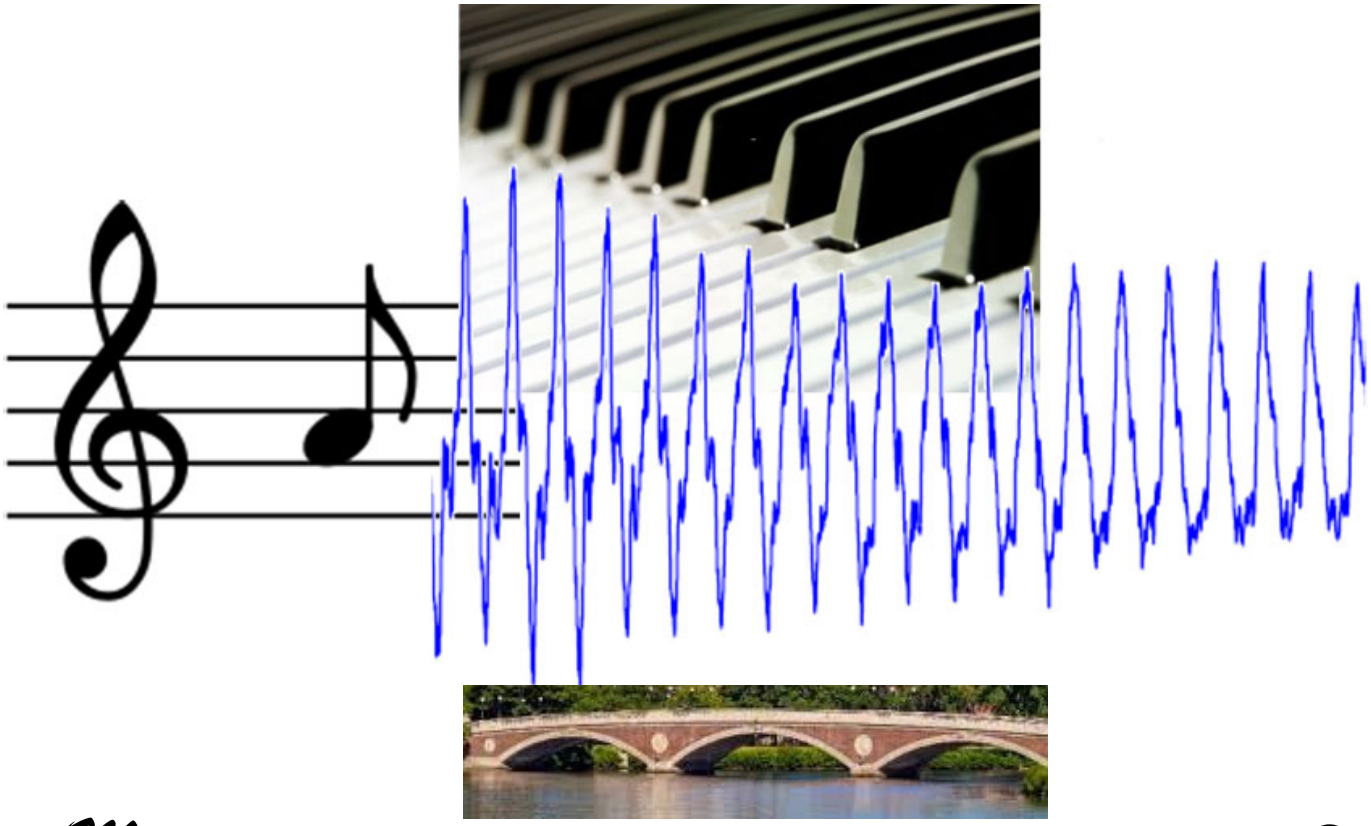
Part B: The Breadth

- Singing voice processing
- Music production audio effects
- Automatic Music Generation
- Synthesis of sound & music – a DSP approach
- Project presentations/demo

Today's topics

1) Recap what we have learnt last week
2) Music representations – music notation demystified
3) Music analysis (e.g. transcription)

Music Representations



Music — ***CS4347/CS5647*** — Computing

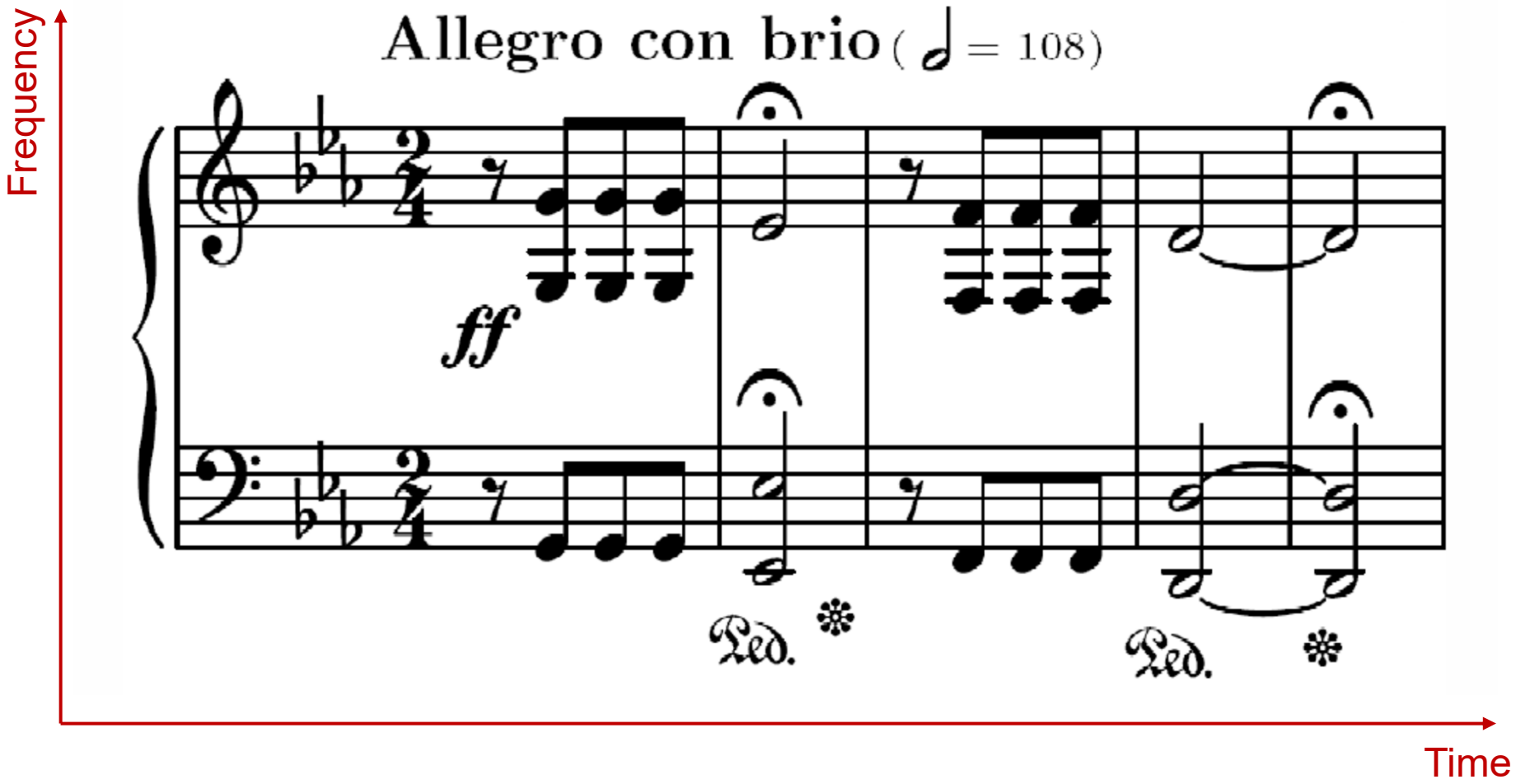
ASR (CS5241) AMT (CS4347)

What is music?

The Concise Oxford Dictionary defines music as "the art of combining vocal or instrumental sounds (or both) to produce beauty of form, harmony, and expression of emotion"

Why do we listen to music?

Sheet Music Representations

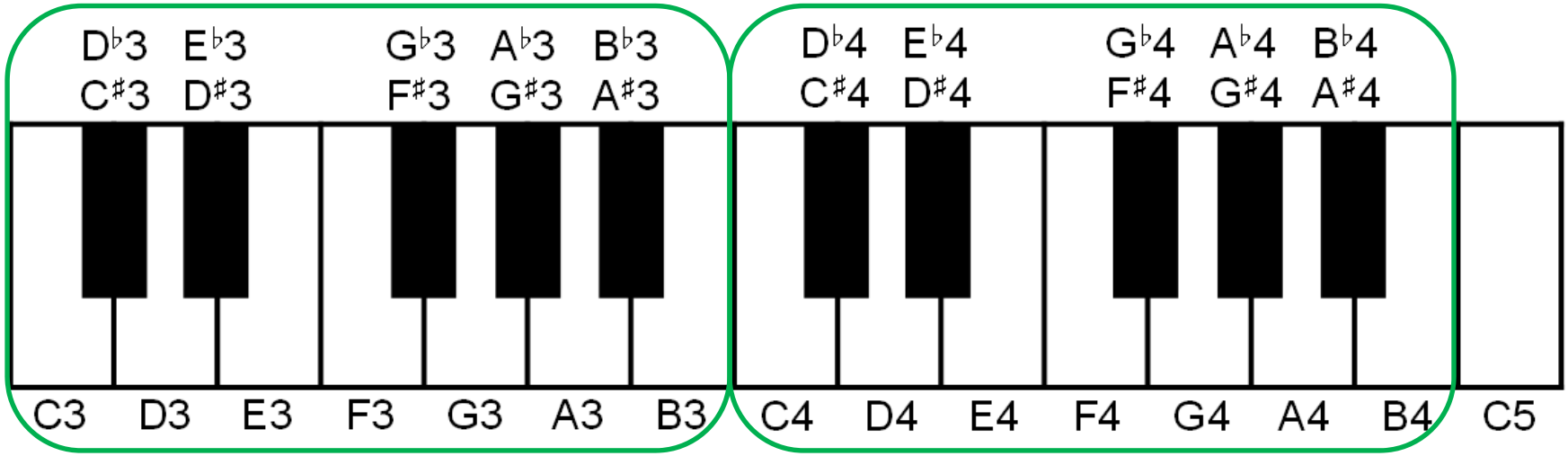


Textual description of music – the language of musicians!

Let's try to demystify it with the spectrogram representation!

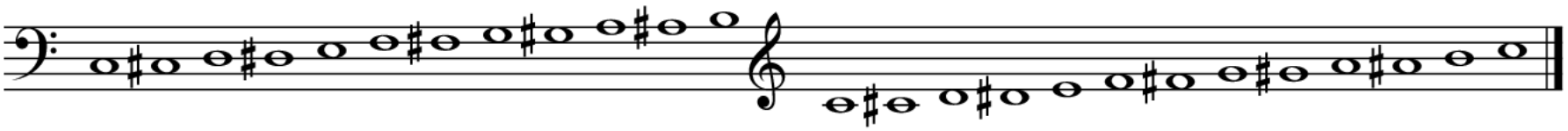
Sheet Music Representations

Twelve-tone equal-tempered scale in western music



1 octave

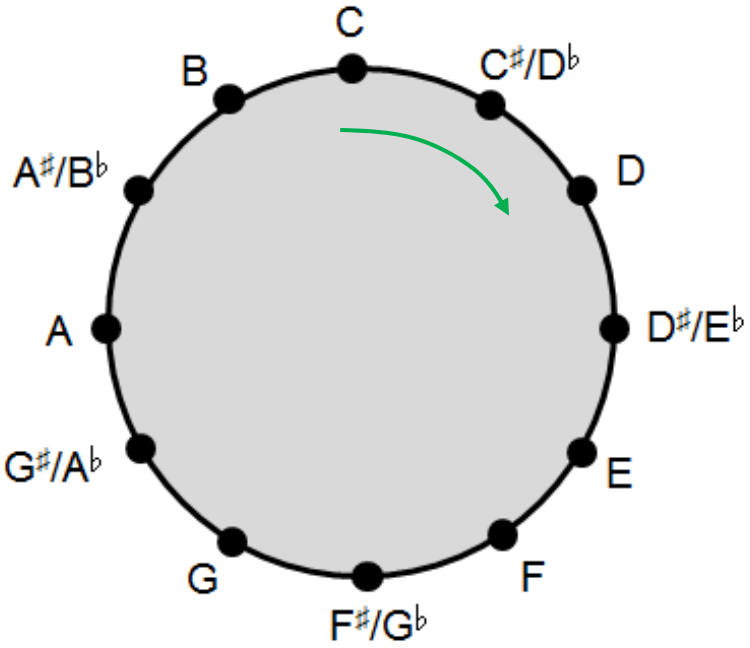
1 octave

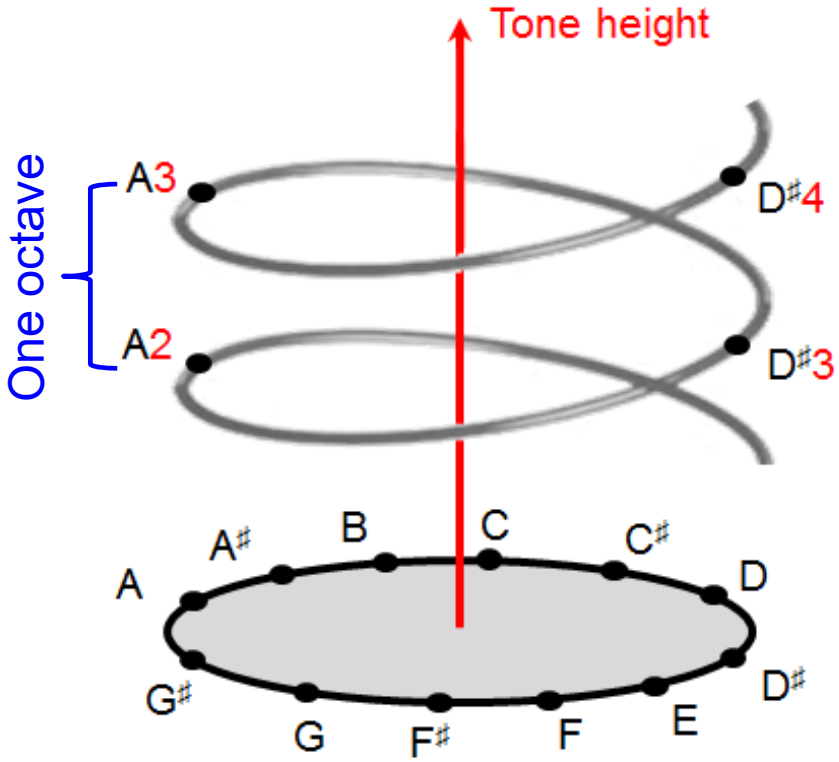


Compare symbols in the sheet music to words and phonemes in speech!

Sheet Music Representations

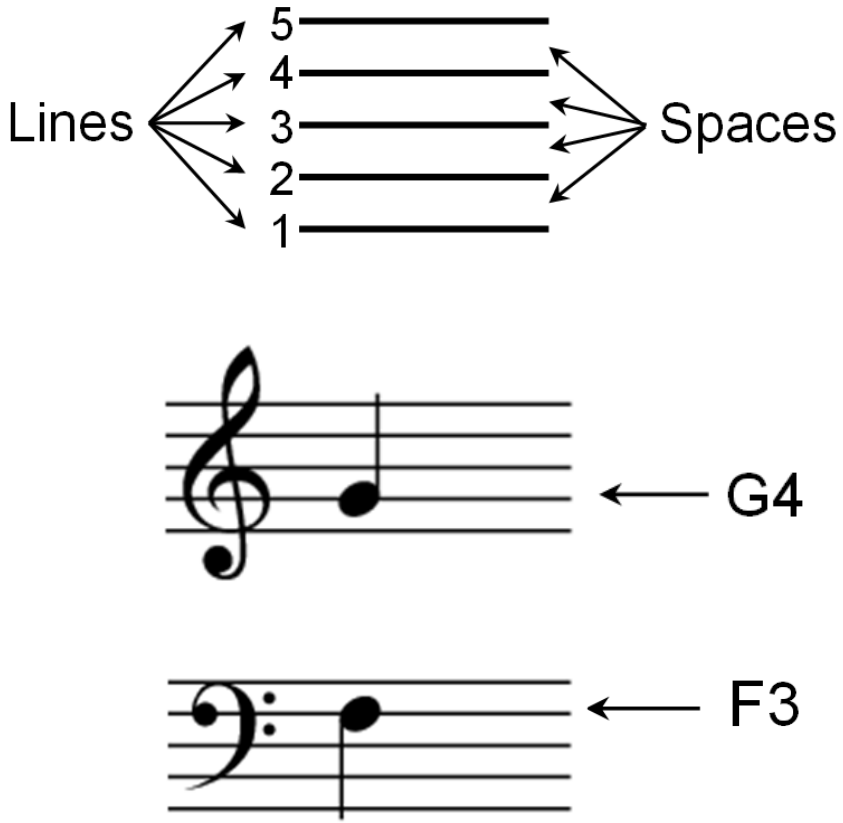
Twelve-tone equal-tempered scale in western music





Do not confuse this spiral representation with the unit circle for DFT!

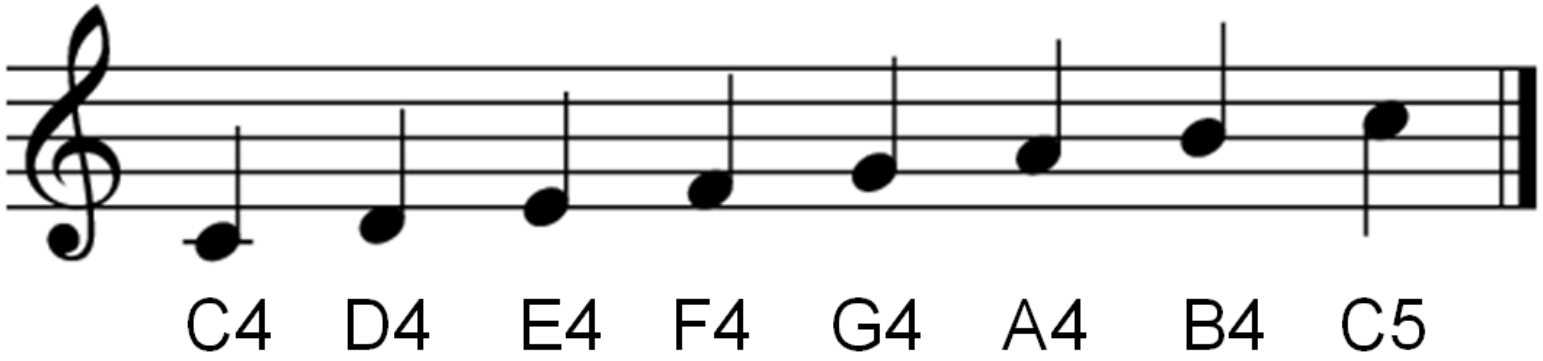
Sheet Music Representations



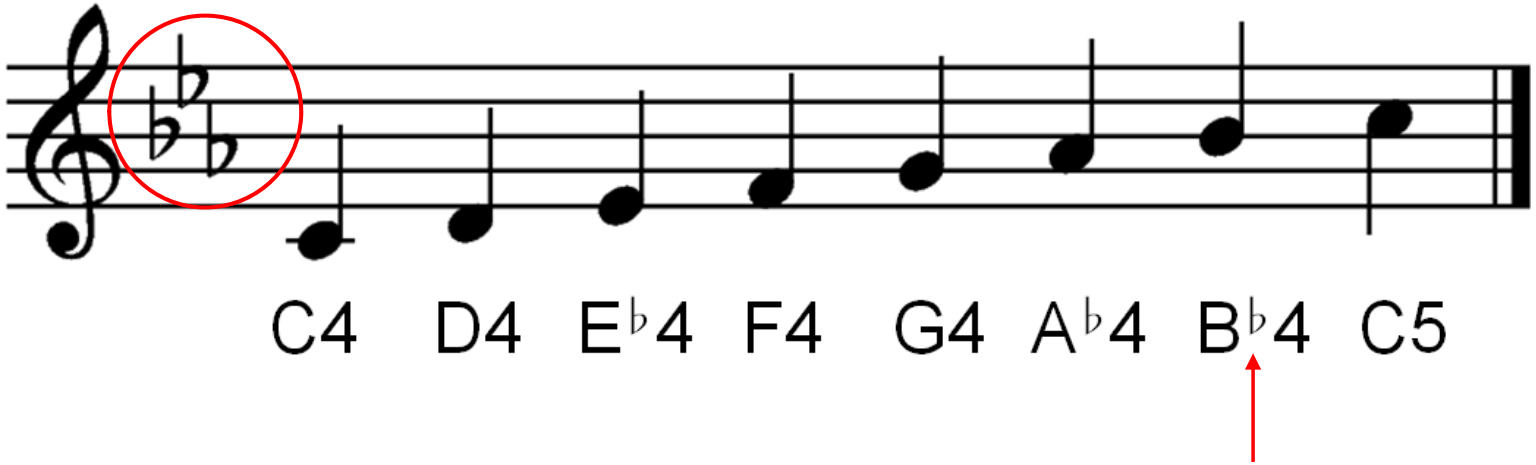
Compare symbols in the sheet music to words and phonemes in speech!

Sheet Music Representations (*pitch*)

C-major scale starting with C4 and ending with C5

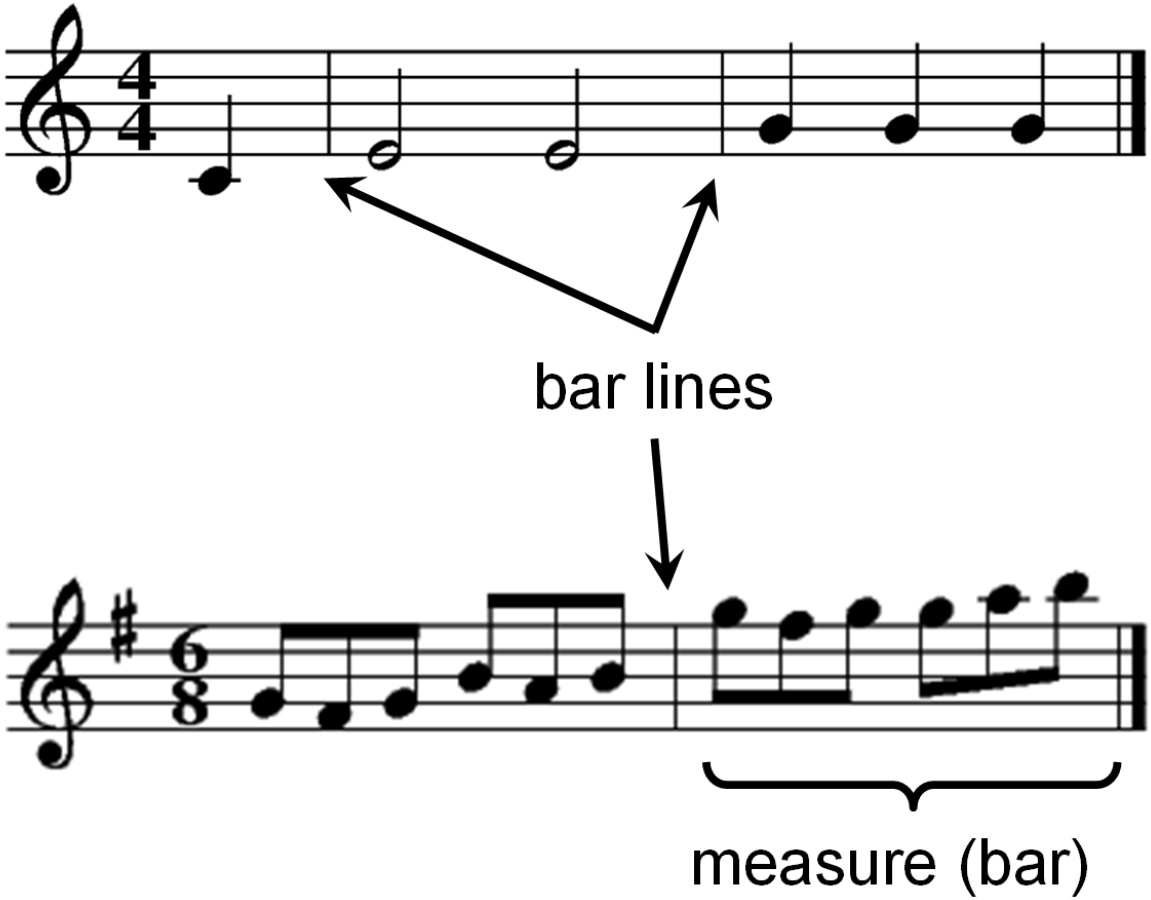


Key signature consisting of three flats converting the notes into a C-Minor scale



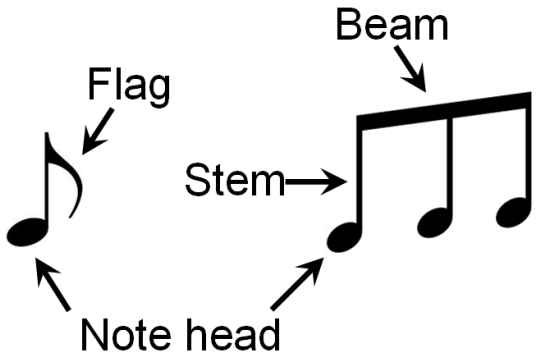
A flat (♭) lowers a note by a semitone.

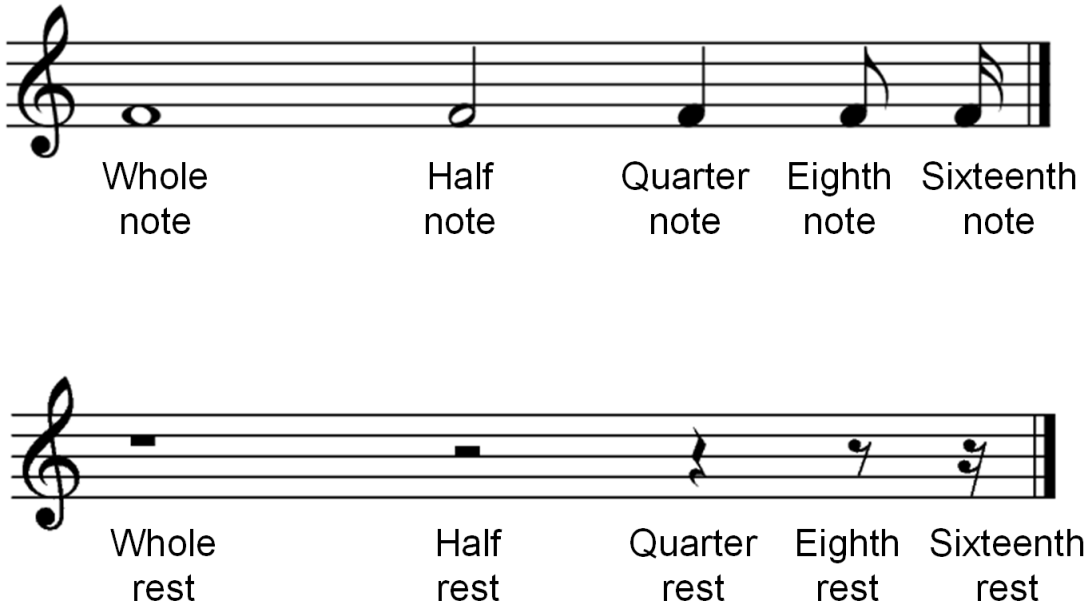
Sheet Music Representations (*time*)



Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

Sheet Music Representations

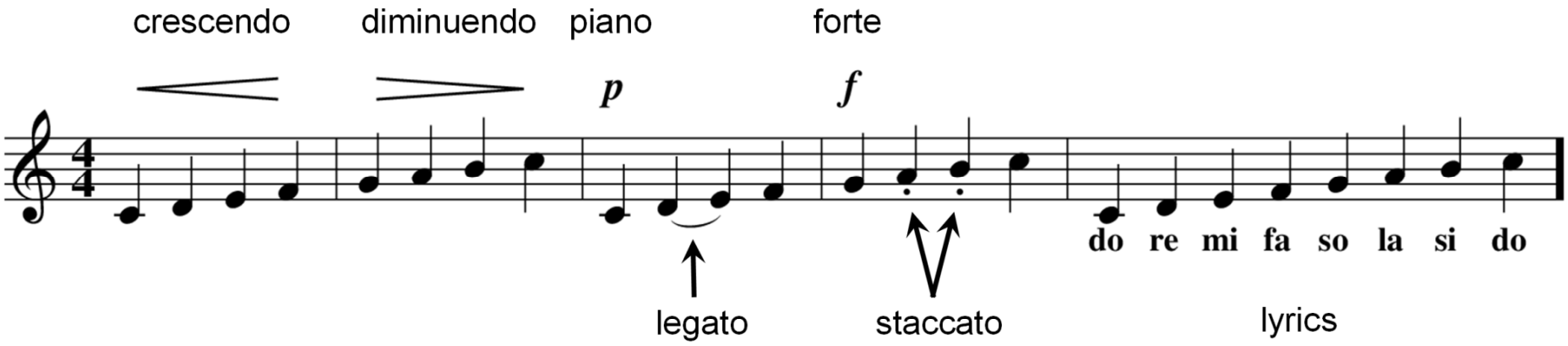




Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

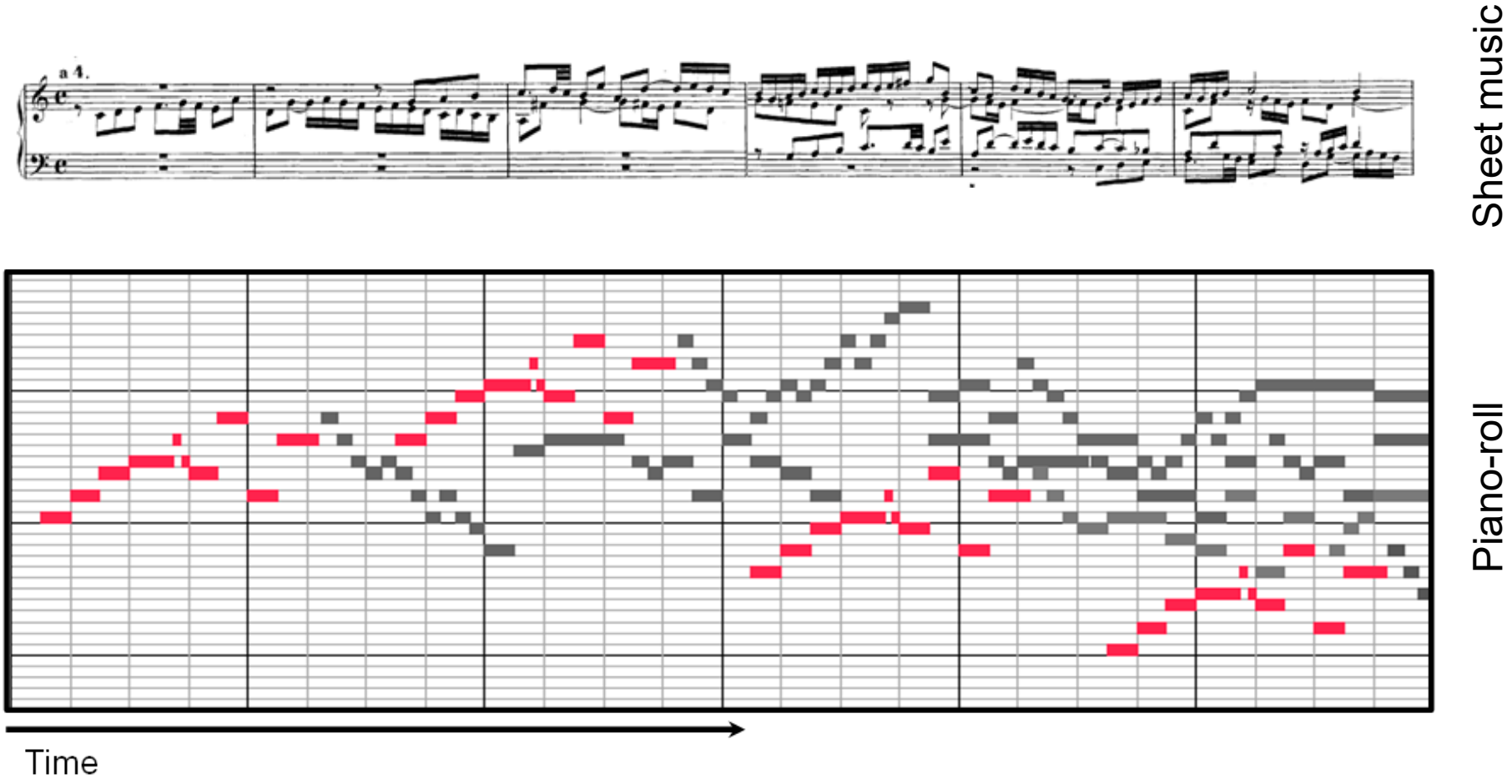
Sheet Music Representations

A very intuitive language – a demo session by a pianist would be helpful!



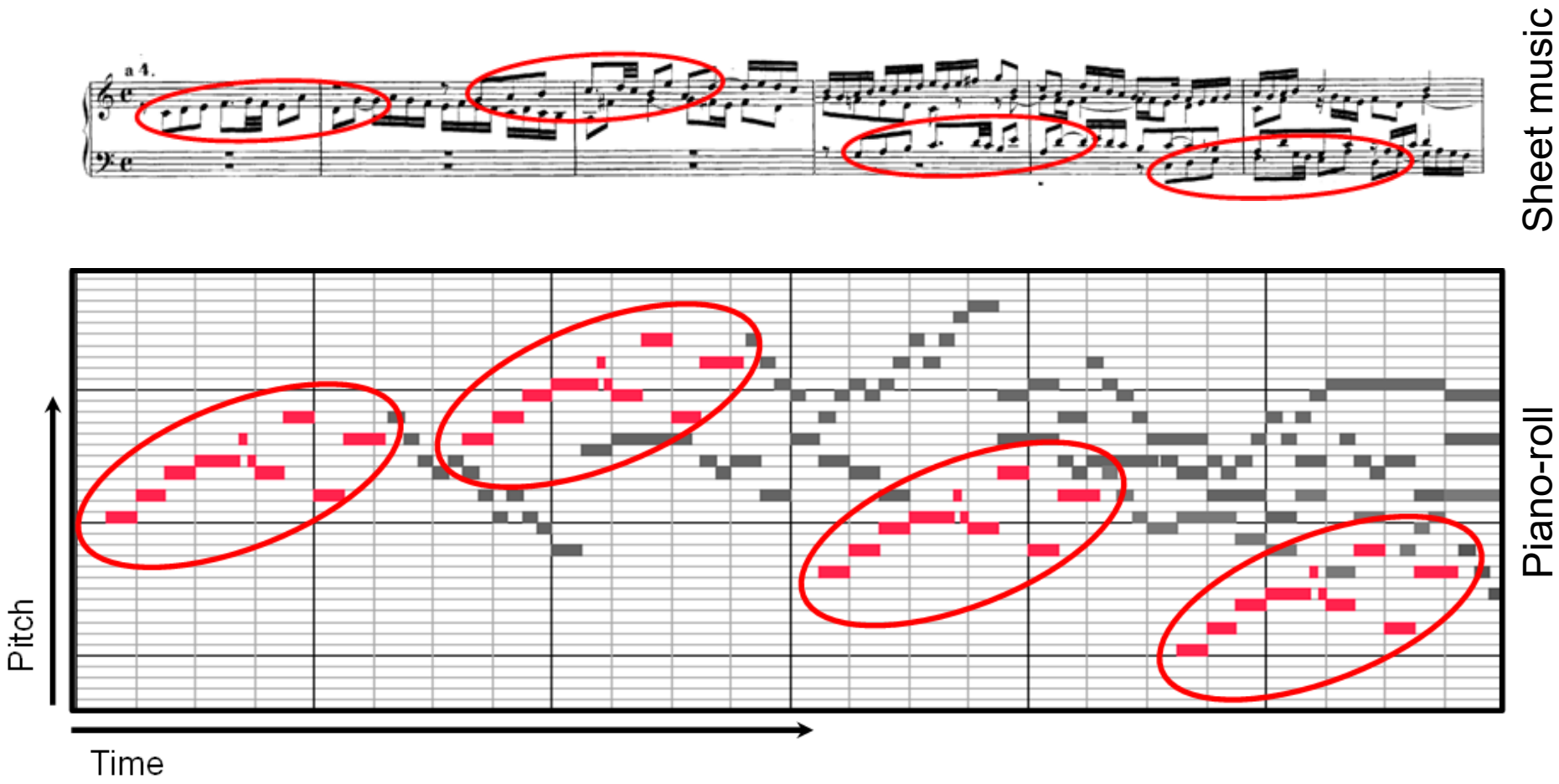
Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

Symbolic Representations

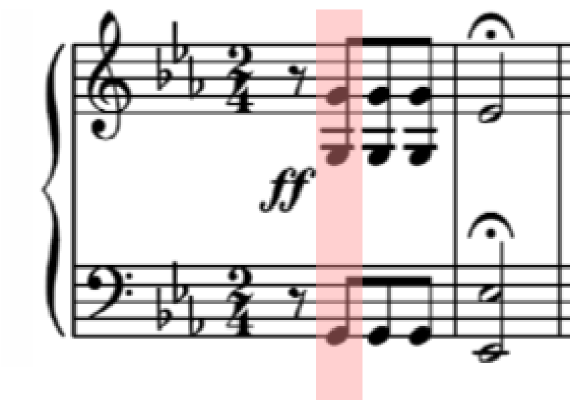


Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

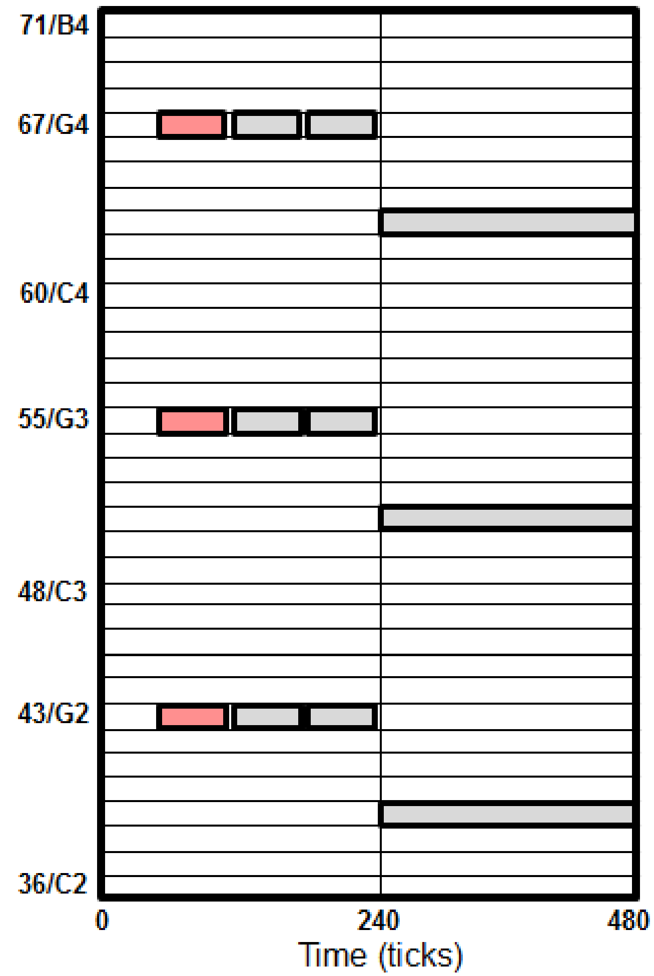
Symbolic Representations



Symbolic Representations



Time (Ticks)	Message	Channel	Note Number	Velocity
60	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	63	100
0	NOTE ON	2	51	100
0	NOTE ON	2	39	100
240	NOTE OFF	1	63	0
0	NOTE OFF	2	51	0
0	NOTE OFF	2	39	0



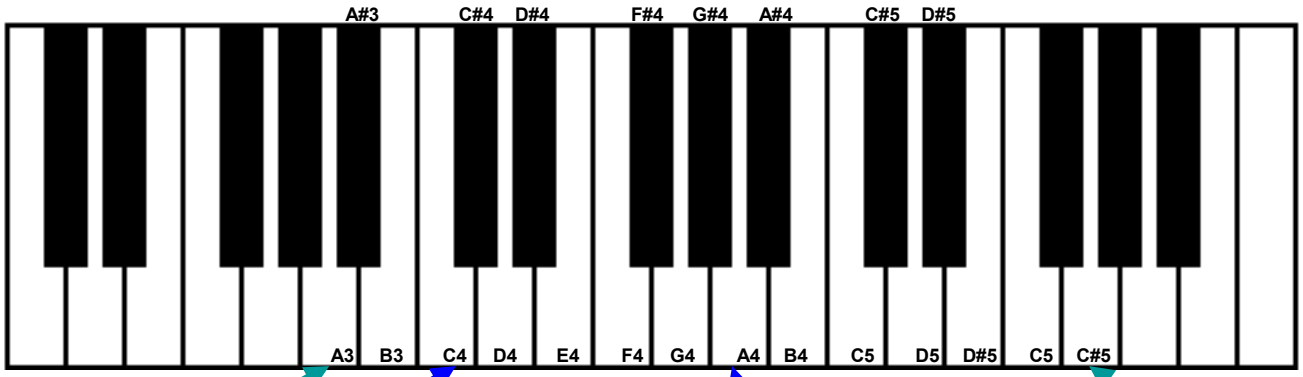
Sheet music

MIDI

Piano-roll

Mapping Pitch to Notes, to Keyboard Position

- Keyboard players usually take bearings from “Middle C” (C4)
- Orchestral Players tune to the A above “Middle C” (A4)



NOTE	A3	A#3	B3	C4	C#4	D4	D#4	E4	F4	F#4	G4	G#4	A4	A#4	B4	C5	C#5	D5	D#5	C5	C#5	…	…
PITCH	220	233	247	262	277	294	311	330	349	370	392	415	440	466	493	523	554	587	622	659	698	…	…
MIDI No	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	…	…

Note Name (With ANSI index), Perceived Pitch (rounded to nearest Hertz) & MIDI No.

NOTE	A	A#	B	C	C#	D	D#	E	F	F#	G	G#	A	A#	B	C	C#	D	D#	C	C#	…	…

Note Name

Scales, Frequency Ratio & Cents

Given a frequency ratio r, the number of cents c is:

$c = (1200/\log_{10}(2))*\log_{10}(r))$

Given c the number of cents, to convert back to ratio r:

$r = 10^{\{c(\log10(2))/1200\}}$

C_0	16.352	C_3	130.81	C_6	1046.5
	17.324		138.59		1108.7
D_0	18.354	D_3	146.83	D_6	1174.7
	19.445		155.56		1244.5
E_0	20.602	E_3	164.81	E_6	1318.5
F_0	21.827	F_3	174.61	F_6	1396.9
	23.125		185.00		1480.0
G_0	24.500	G_3	196.00	G_6	1568.0
	25.957		207.65		1661.2
A_0	27.500	A_3	220.00	A_6	1760.0
	29.135		233.08		1864.7
B_0	30.868	B_3	246.94	B_6	1975.5
C_1	32.703	C_4	261.63	C_7	2093.0
	34.648		277.18		2217.5
D_1	36.708	D_4	293.66	D_7	2349.3
	38.891		311.13		2489.0
E_1	41.203	E_4	329.63	E_7	2637.0
F_1	43.654	F_4	349.23	F_7	2793.8
	46.249		369.99		2960.0
G_1	48.999	G_4	392.00	G_7	3136.0
	51.913		415.30		3322.4
A_1	55.000	A_4	440.00	A_7	3520.0
	58.270		466.16		3729.3
B_1	61.735	B_4	493.88	B_7	3951.1
C_2	65.406	C_5	523.25	C_8	4186.0
	69.296		554.37		4434.9
D_2	73.416	D_5	587.33	D_8	4698.6
	77.782		622.25		4978.0
E_2	82.407	E_5	659.26	E_8	5274.0
F_2	87.307	F_5	698.46	F_8	5587.7
	92.499		739.99		5919.9
G_2	97.999	G_5	783.99	G_8	6271.9
	103.83		830.61		6644.9
A_2	110.00	A_5	880.00	A_8	7040.0
	116.54		932.33		7458.6
B_2	123.47	B_5	987.77	B_8	7902.1

Frequencies of Notes in the tempered scale

Notes naming: From Music to MIDI

- Notes of a western scale are named in ascending order from the first seven letters of the alphabet, A-G and they repeat to represent the same notes (notes of the same chroma) at higher or lower octaves.
  - e.g. 440Hz is A, 220Hz is A one octave lower

- American Standards Association extension: Number (register) appended to the note's letter name so as to associate a note to its fundamental frequency.
  - e.g 440Hz is A4, 220Hz Is A3, 880Hz is A5

- MIDI: Numbers are assigned to note pitches for ease of representation.
  - e.g. A3 is assigned 69, A4 assigned 81 (69+12 reflecting A4 as 12 semitones, 1 octave higher than A3)

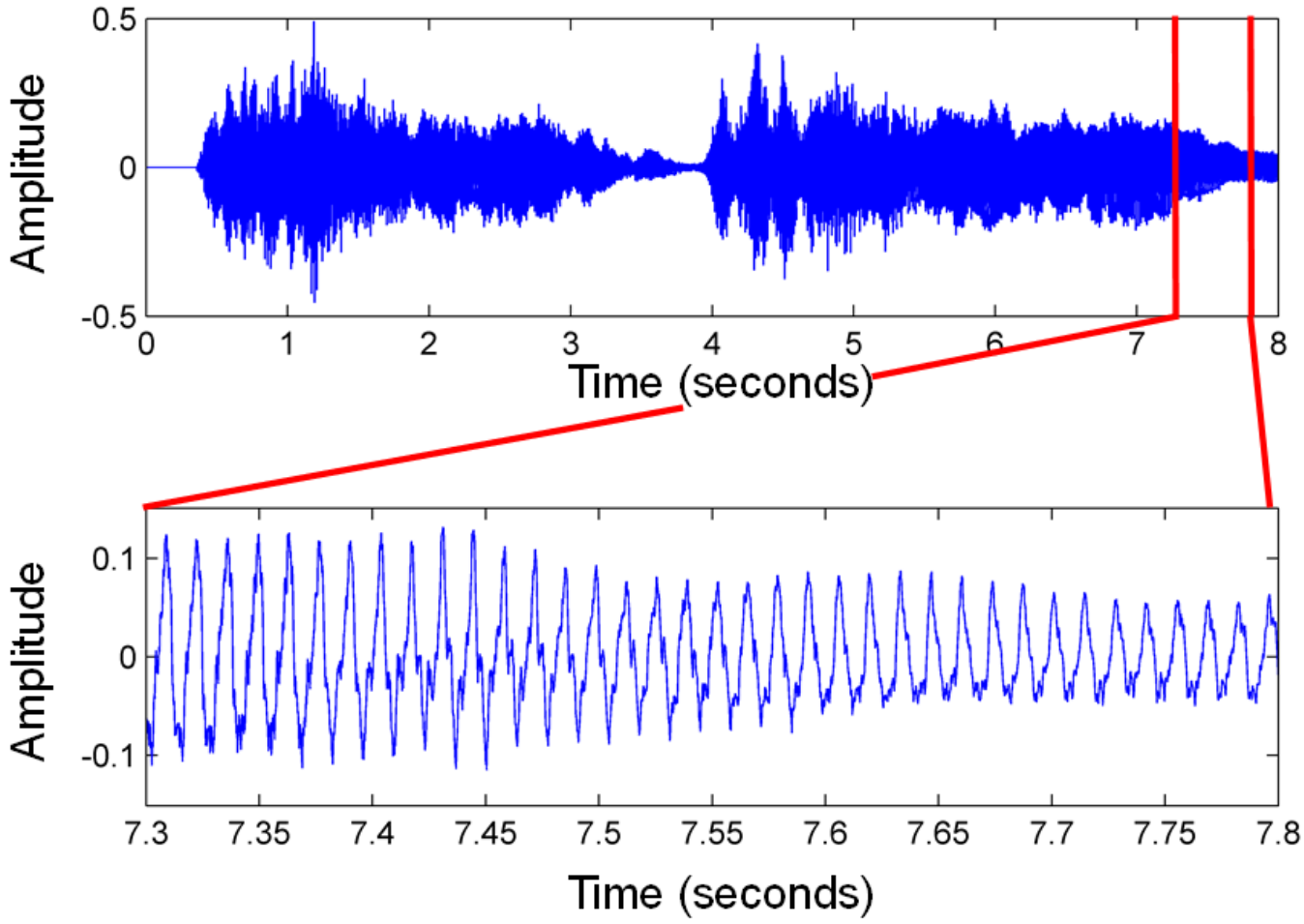
- Converting fundamental frequency (f) to a MIDI note number (n) and vice versa*:

$$n = 12 \times \log_2 \left[\frac{f}{440} \right] + 69$$

$$f = 440 \times 2^{\left(\frac{n-69}{12}\right)}.$$

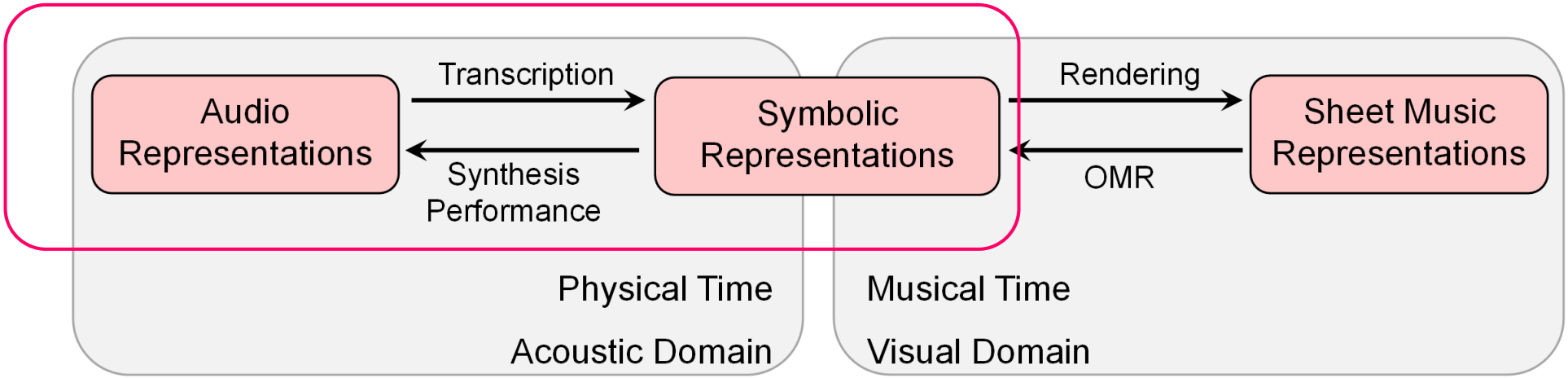
*Assuming equal temperament scale

Audio Representations



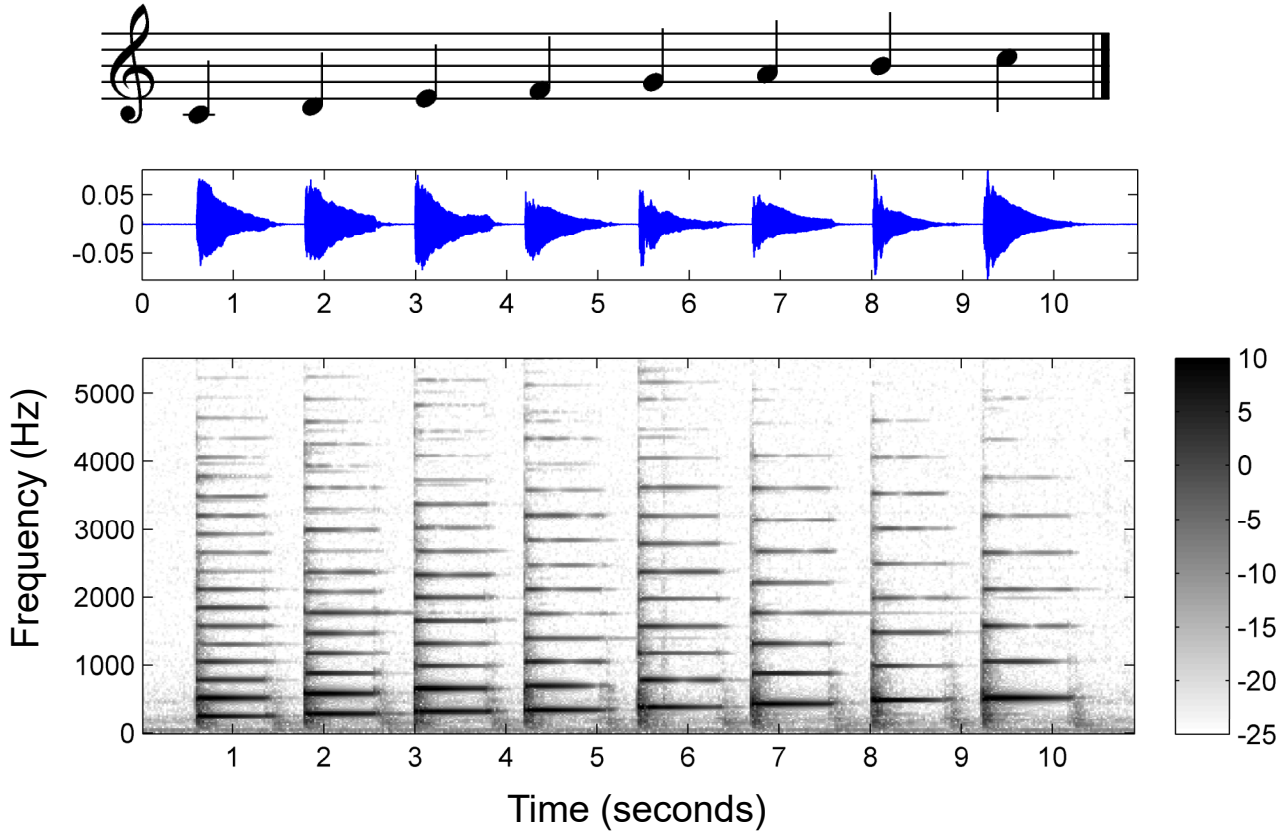
Connecting the Dots

MIDI and Piano-roll



Optical Music Recognition (OMR)

Music Representations



Sheet music

Waveform

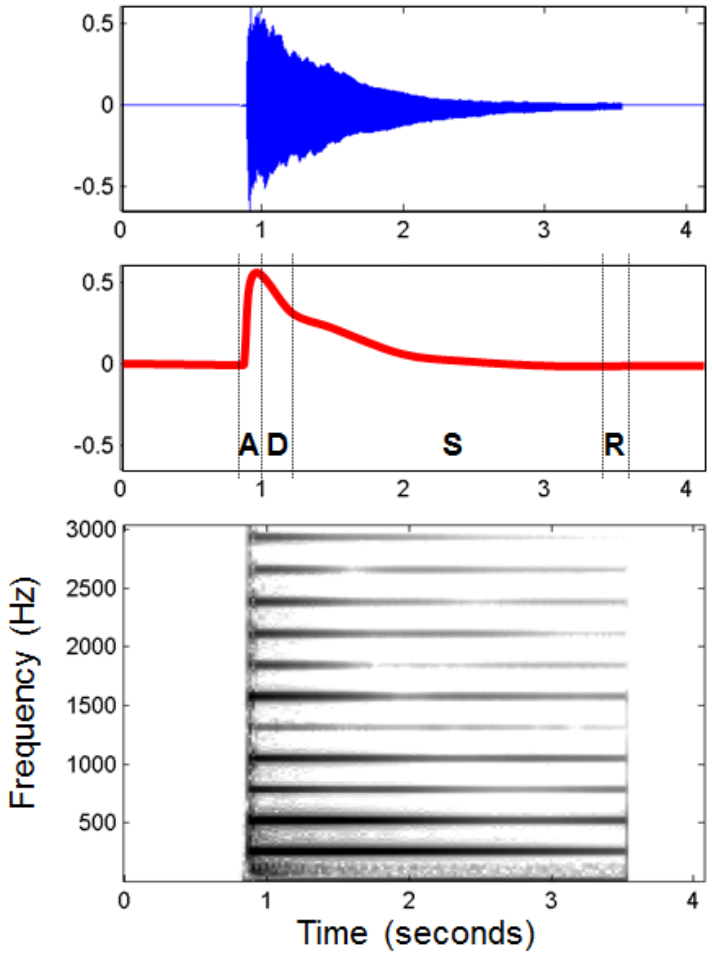
Spectrogram

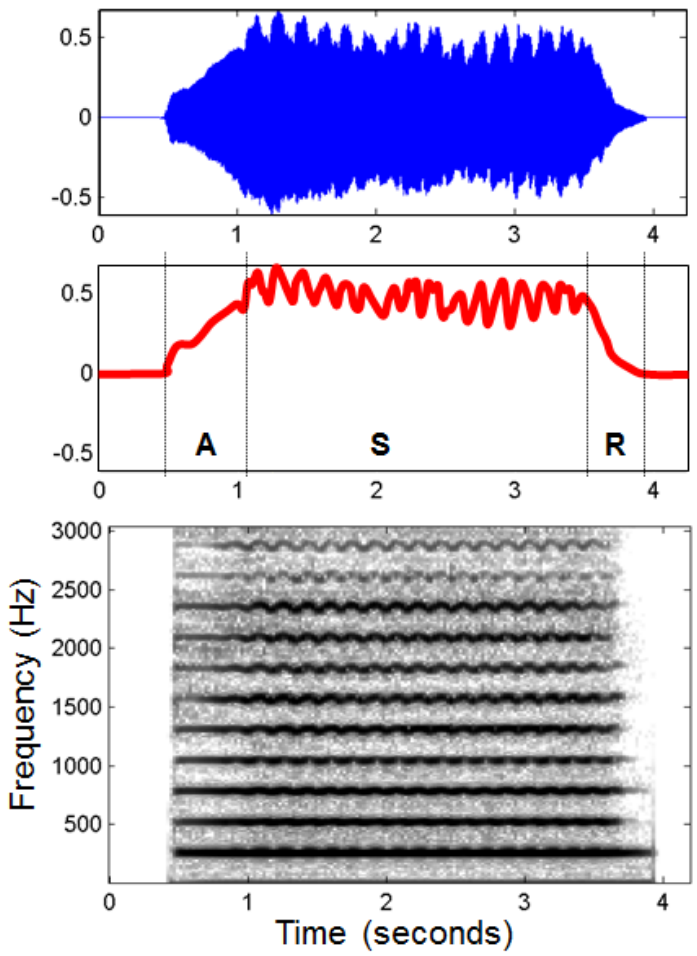
How do you convert the TF representation to a piano roll?

Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

Music Representations

3 basic parameters for an AMT: ***pitch, onset and offset***





Waveform

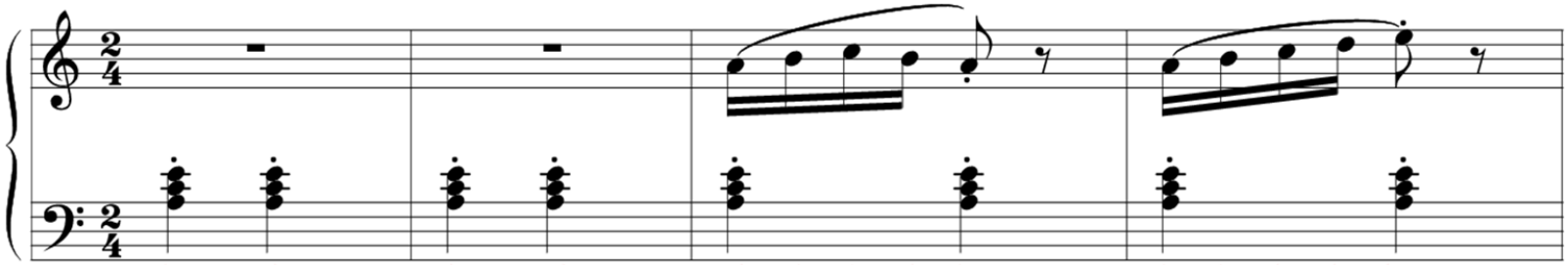
Contour

Spectrogram

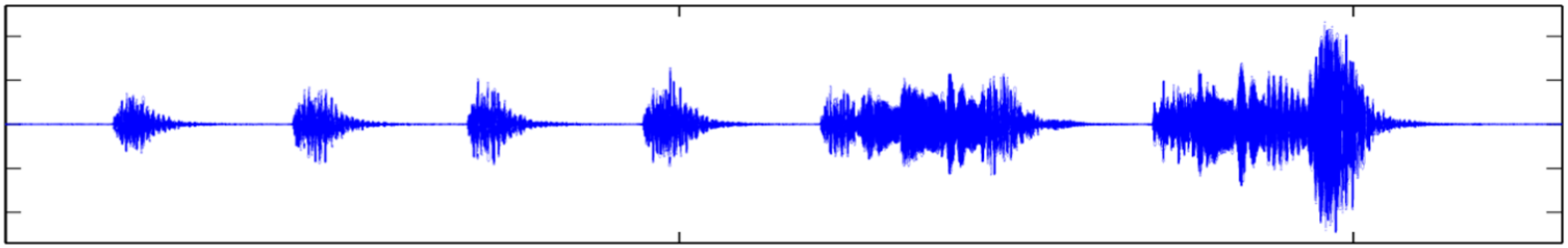
Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

$$V_{F \cdot T} \approx W_{F \cdot k} \cdot H_{k \cdot T}$$

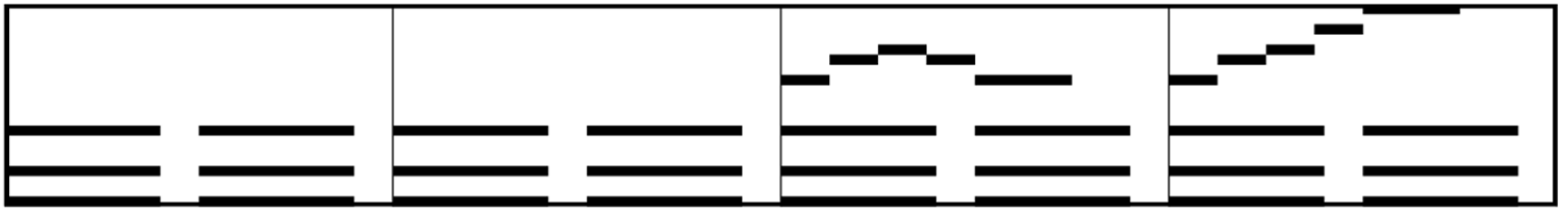
Music Representations



Sheet music



Waveform



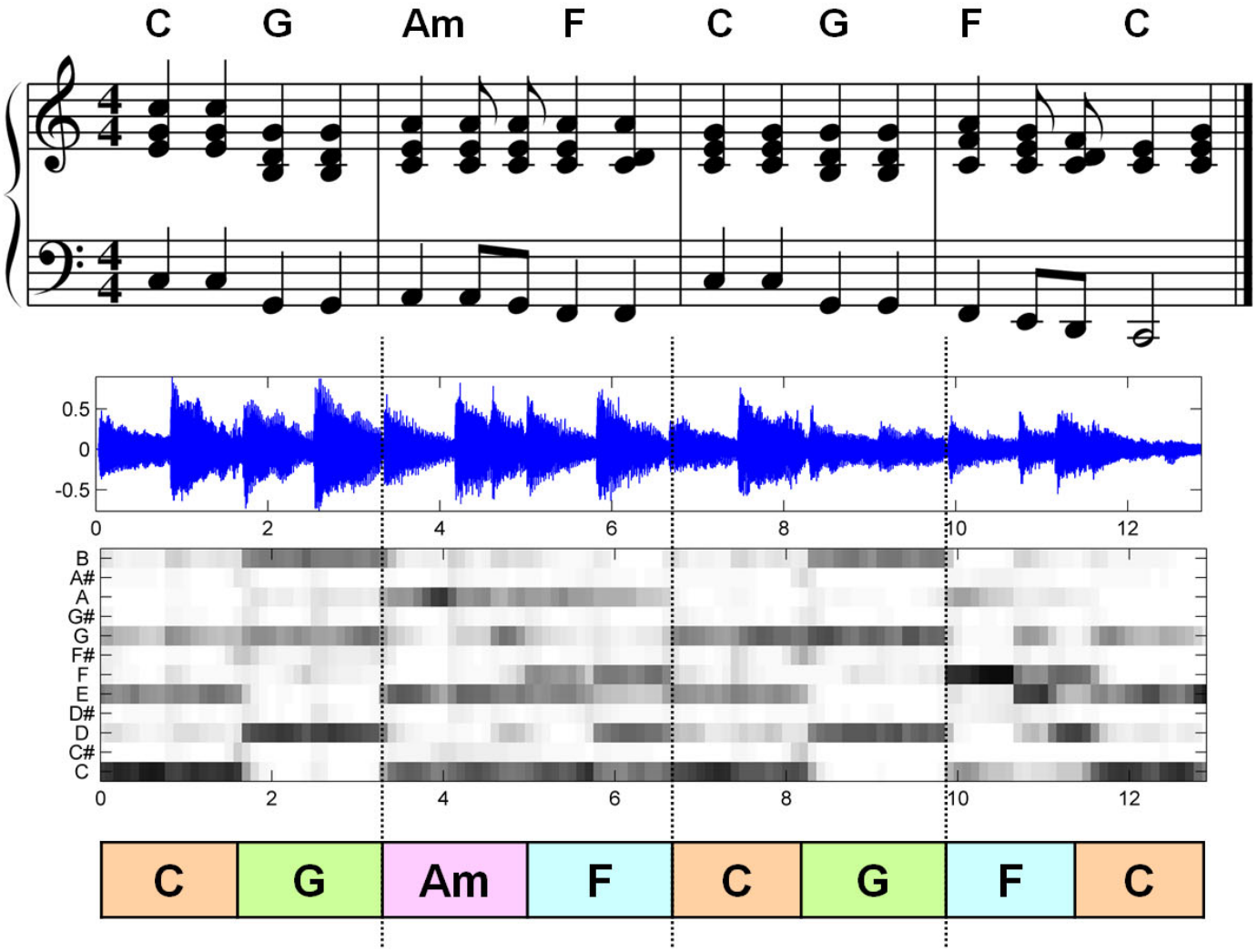
Piano-roll

Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

Non-negative matrix factorization (NMF)

$$V_{F \cdot T} \approx W_{F \cdot k} \cdot H_{k \cdot T}$$

Music Representations



Sheet music

Waveform

Chroma

Chord recognition

Meinard Müller, Fundamentals of Music Processing
ISBN: 978-3-319-21944-8, Springer, 2015

Music Representations (MIDI format)

MIDI (**M**usical **I**nstrument **D**igital **I**nterface) is a technical **standard** that describes 3 components:

- Protocol
- Digital Interface
- Connectors

What is the purpose of MIDI?

- In 1970 each digital synthesizer manufacturer was creating its own system to link the input device (usually keyboard) to the sounds
- In 1983 the full MIDI 1.0 detailed specification was released

MIDI is a standardized way of communication between synthesizers, computers, samplers and other musical equipment.

What is MIDI data?

MIDI does **NOT** carry sound, it carries **event messages**.

What type of events ?

- The start of a note
- The end of a note
- Changing the parameters of the synthesizer
- Other controls and configuration messages

What is a MIDI file?

A **MIDI** recording, or **MIDI** file can be thought of as being an enhanced score:

- Exact dynamics (quantified between 0-127)
- Exact timings
- Other information (eg: sustain pedal)

A **MIDI** file does **NOT** contain the music. It contains just “instructions” to play the music, thus the file size of a MIDI recording is greatly reduced.

How can we listen to a MIDI file ?

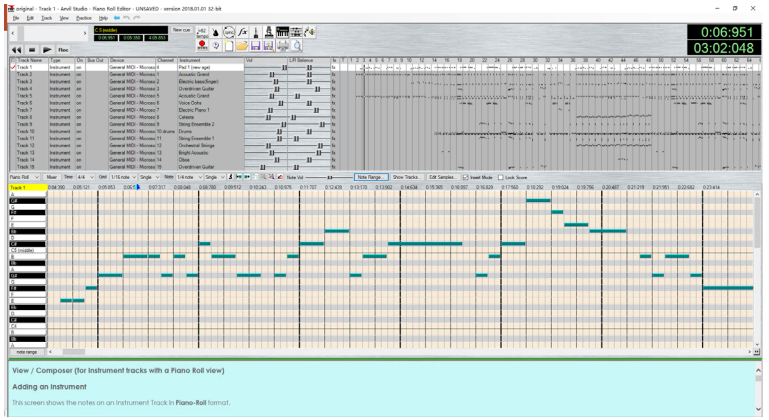
To listen to a MIDI file we need one or more pieces of software or hardware that are able to play the MIDI “instructions”:

- Synthesizers
- Samplers
- Virtual Instruments

Can you hear the difference?

Pros and Cons of MIDI

- Easy to edit
  - Change notes
  - Change tempo
  - Change instruments
  - Add tracks
  - …



- Free software: GarageBand, Anvil Studio, …
- Python libraries: mido, pretty-midi, …

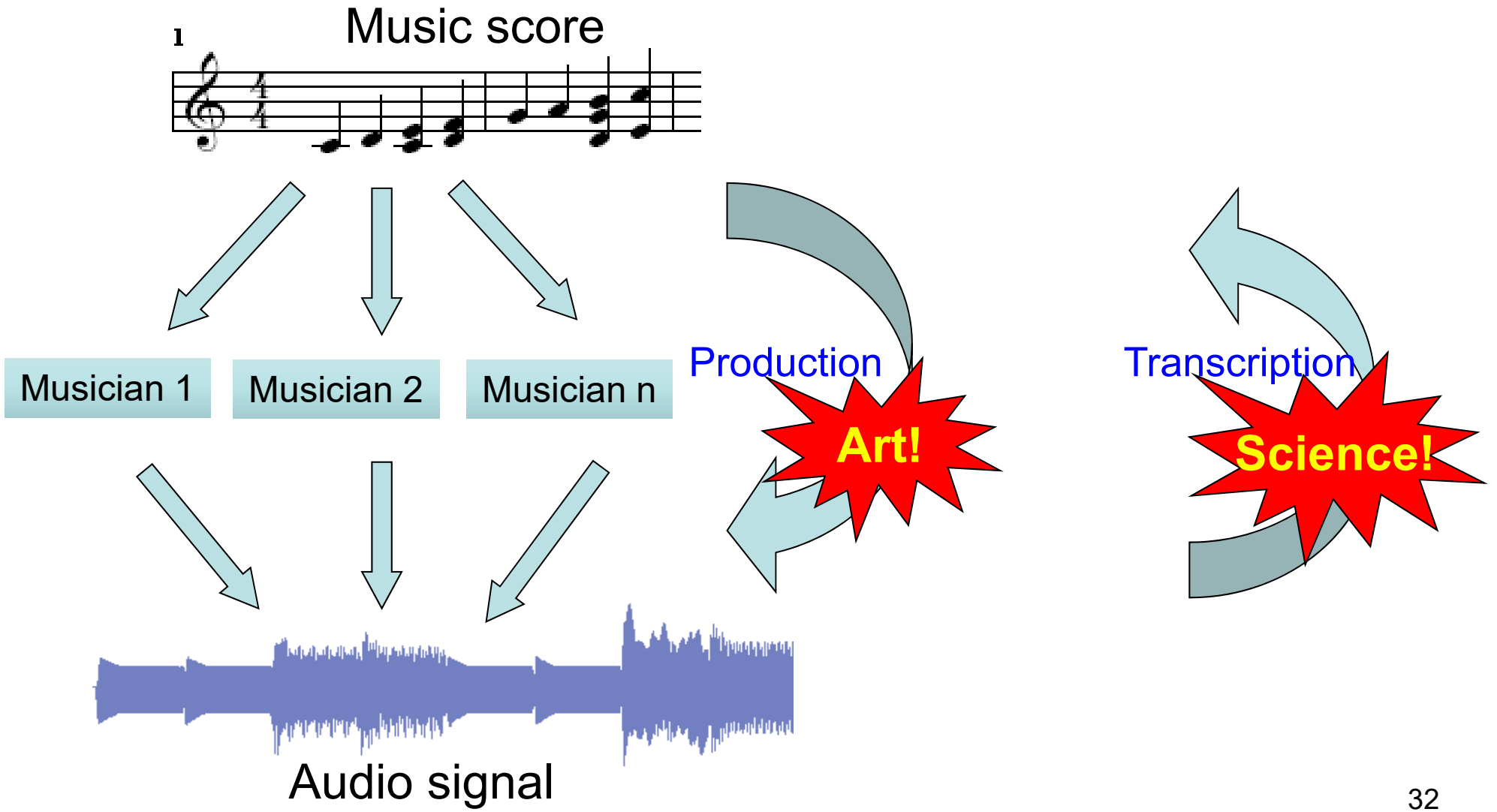
- Inconsistent sound effects
- No vocals
- Less realistic



Today's topics

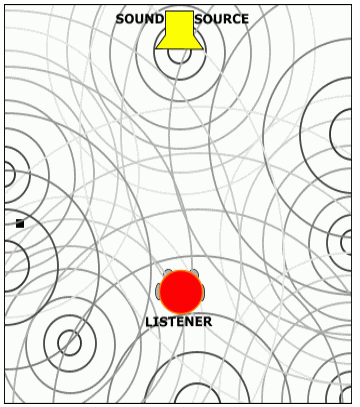
1) Recall what we have learnt last week
2) Music representations – music notation demystified
3) Music analysis (e.g. transcription)

Music Production and Transcription



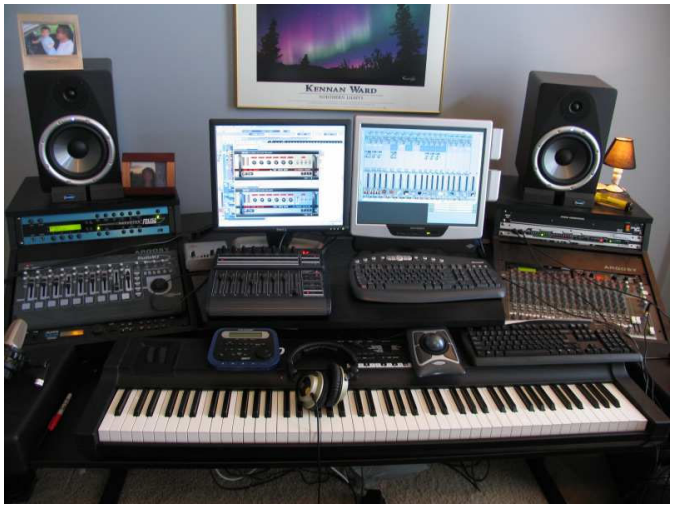
Acoustics

- Reverberation (echo) is caused by reflected waves
  - exaggerated in enclosed rooms
  - amount depends on material, room size

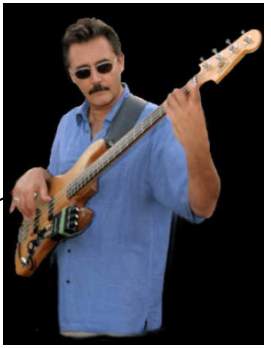


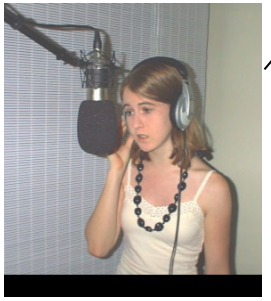
- Recording studios typically have little reverb
  - Acoustically treated walls absorb / diffuse sounds.

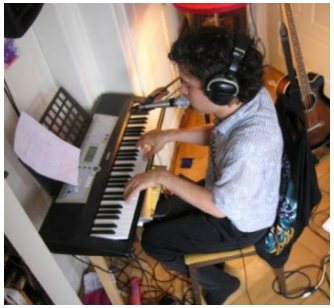
Recording



- Studio Engineer





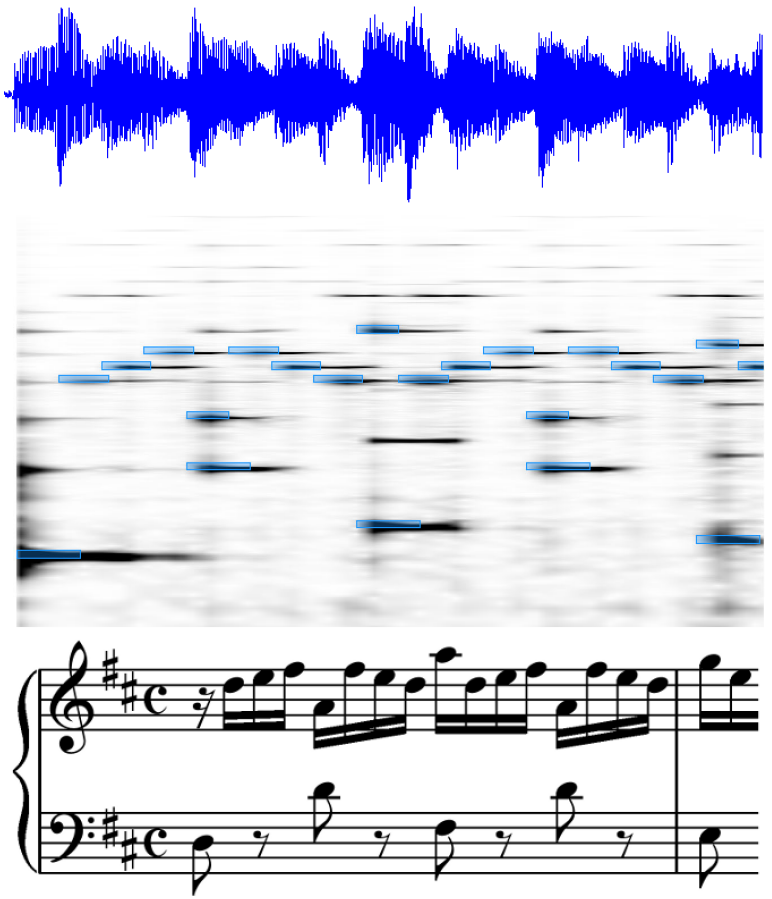


Mixing

- Balance the relative volume, frequency, and dynamical content of a number of sound sources
  - Drums 
  - Bass 
  - Guitar 1 
  - Guitar 2 
  - Vocals 
  - End product 
- Usually, digital WAVE files in 24-bit, 44.1kHz or higher

Automatic Music Transcription (AMT)

Automatic music transcription (AMT): the process of converting an acoustic musical signal into some form of music notation (e.g. staff notation, MIDI file, piano-roll,…)



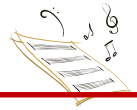
Fundamental (and open) problem in music information research

Applications:

- Search/annotation of musical information
- Interactive music systems
- Music education
- Music production
- Digital/computational musicology

In comparison to ASR, the progress of AMT is quite limited. Why?

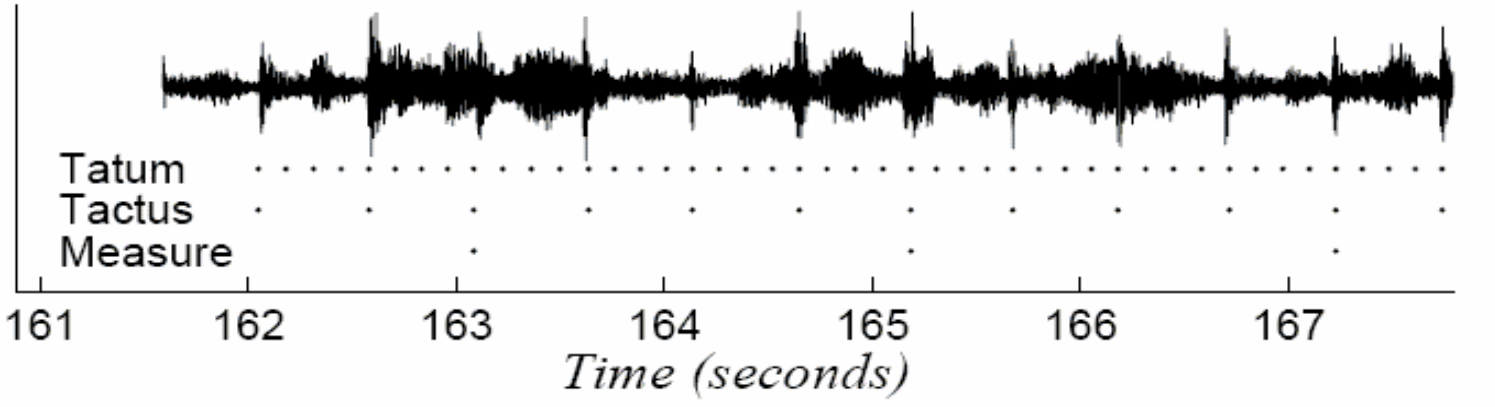
Music Transcription: What to transcribe?

Musical Terminology	Matching Physical property
 Duration/Onsets	Time measured in samples
 Pitch	Fundamental Frequency F0
 Loudness	Logarithmic scale (simplifies)
 Timbre/Tone Color	Difficult, no single physical attribute describes it completely

- Vibrato and other expressive features
- Lyrics

Temporal Structure

- Goal is to **automatically find musically relevant time information** (e.g. start of single notes) in the recorded signal



Temporal Structure

Examples:

Percussive sources:

Drums, Cymbals; rather simple to find by looking at signal energy

Harmonic sources:

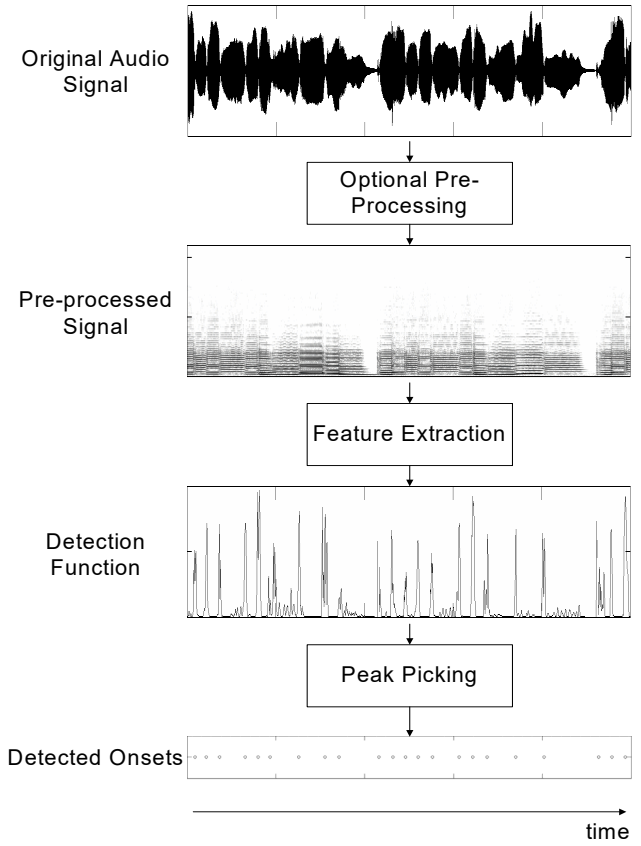
Singing voice, violin, trumpet; much harder to derive temporal structure from harmonic changes in signal

Conclusion:
We need more sophisticated methods for harmonic signals

Temporal Structure

A typical signal processing scenario of an onset detection system

Key is to find numerical methods which allow for reliable peak picking



Temporal Structure

Examples of automatically derived time information:

- **High-frequency tick** (claves) indicates tatum pulse (**note level**)
- **"Side stick"** (snare) indicates the tactus pulse (**beat level** or foot tap)
- **Bass drum** indicates the measure pulse (**musical measure**)

Source: http://www.cs.tut.fi/~klap/iiro/meter/index.html

Pitch and harmonic structure

Pitch:

- perceptual attribute, allowing ordering of sounds on a frequency-related scale from low to high
- Defined as the frequency of a sine wave that is matched to the perceived sound by a human listener

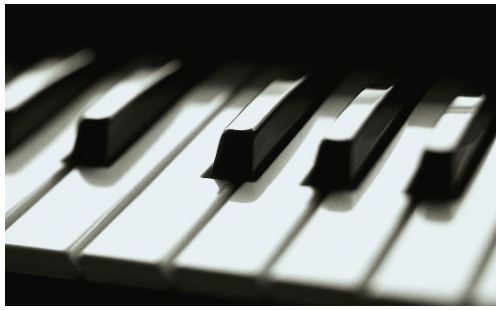
Fundamental frequency:

- Is the corresponding physical term
- Only defined for periodic or nearly periodic musical signals

Commonly used pitch detection algorithms include YIN and pYIN.

Pitch and harmonic structure

- Musical Scale
  - A3=220 Hz
  - Exponentially Stepped $A3 * 2^{1/12} = 233$
  - Semitone Step= $\sqrt[12]{2}$
  - Octave Step= 2 (A3 -> A4; 220 Hz -> 440 Hz)



Semitone

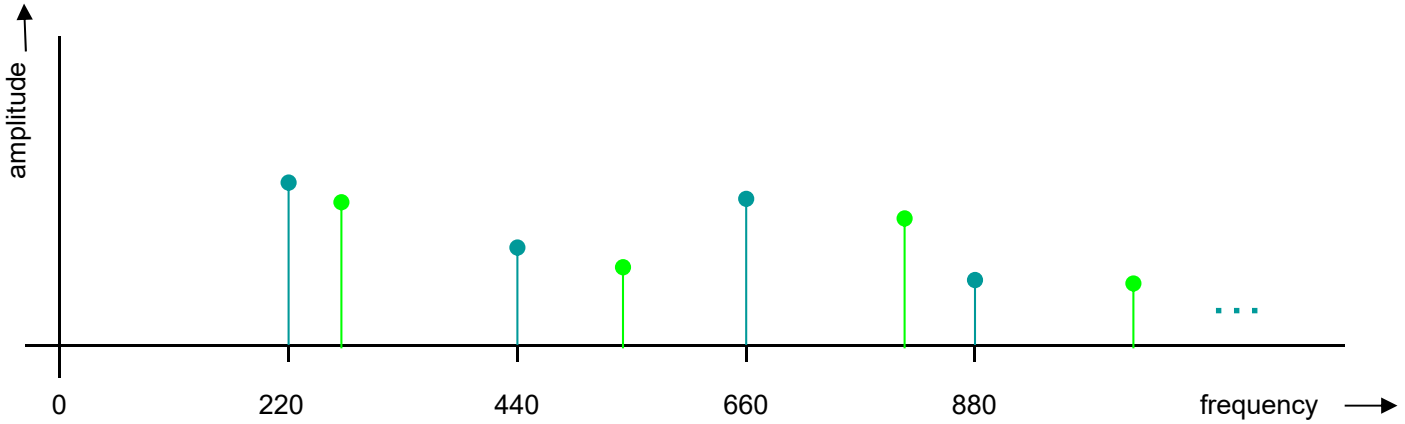
Note	Freq (Hz)	Note	Freq (Hz)
A3	$A3 * 2^{0/12} = 220$	C#4	$A3 * 2^{4/12} = 277$
A#3	$A3 * 2^{1/12} = 233$	D4	$A3 * 2^{5/12} = 294$
B3	$A3 * 2^{2/12} = 247$	D#4	$A3 * 2^{12/12} = 440$
C4	$A3 * 2^{3/12} = 262$	E4	$A3 * 2^{24/12} = 880$

The musical scale approximates the critical bandwidth in our ear!

Pitch and harmonic structure

A challenging problem: multi-pitch detection (polyphonic sound)

Example: A3 (220 Hz) and C4 (262 Hz)



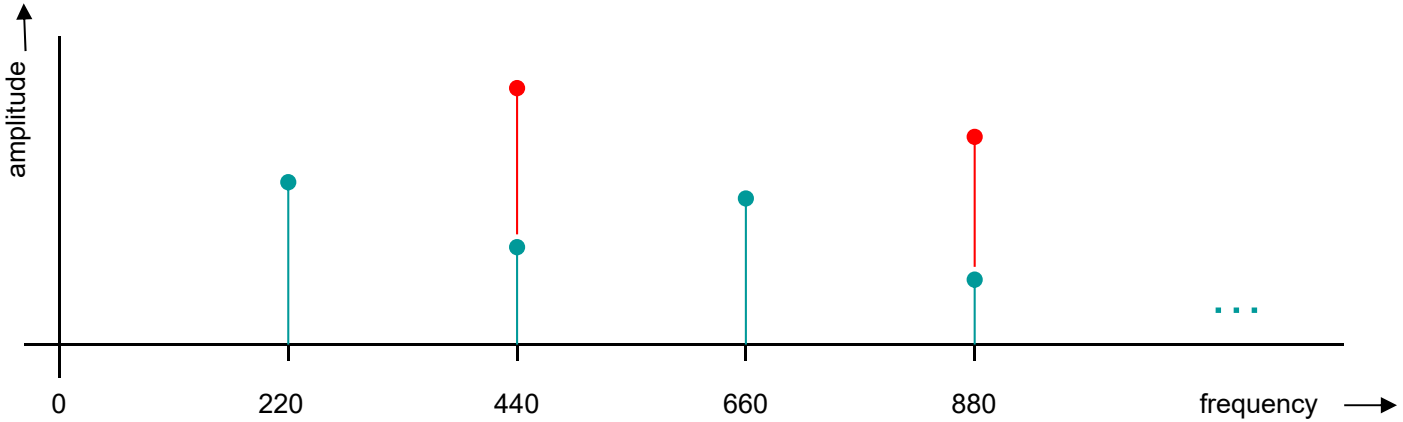
Relation to audio source separation task!

Demystify the cocktail party effect!

Pitch and harmonic structure

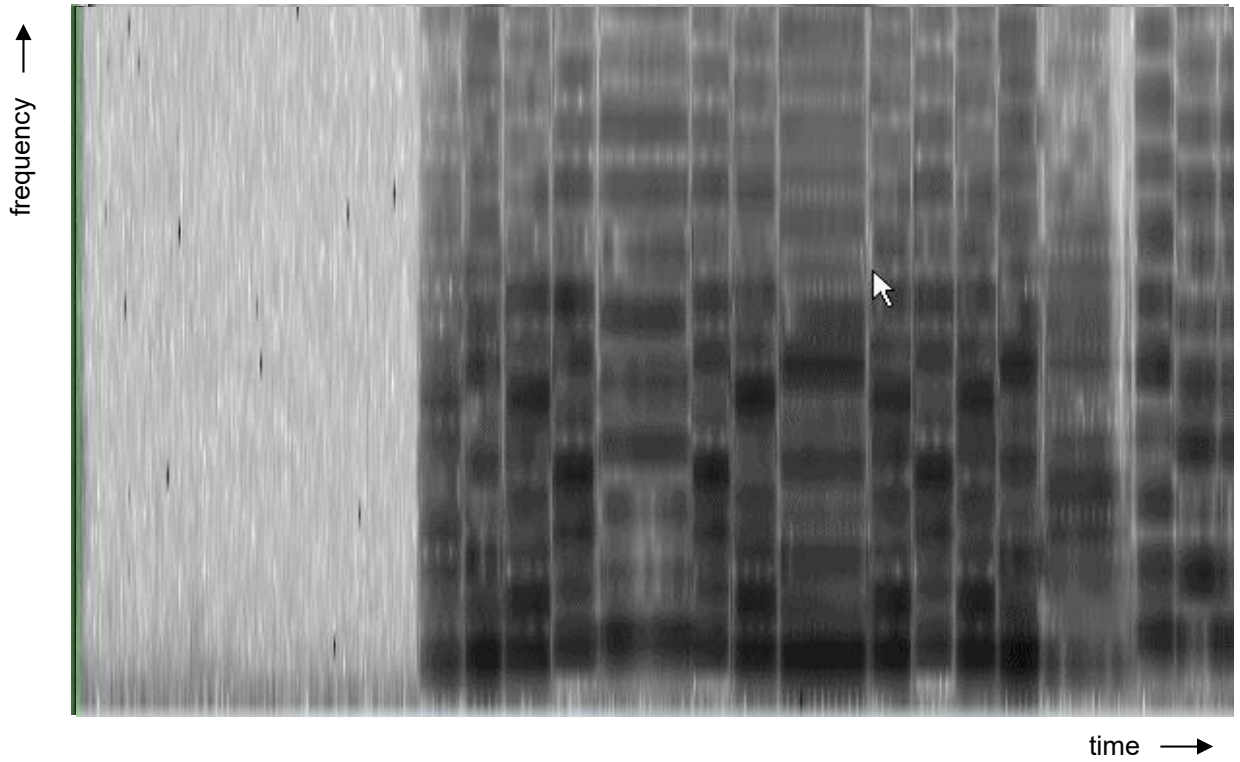
A challenging problem: multi-pitch detection (polyphonic sound)

Problem: A3 (220 Hz) and A4 (440 Hz)
The harmonics overlap and such it is very difficult to separate the two tones



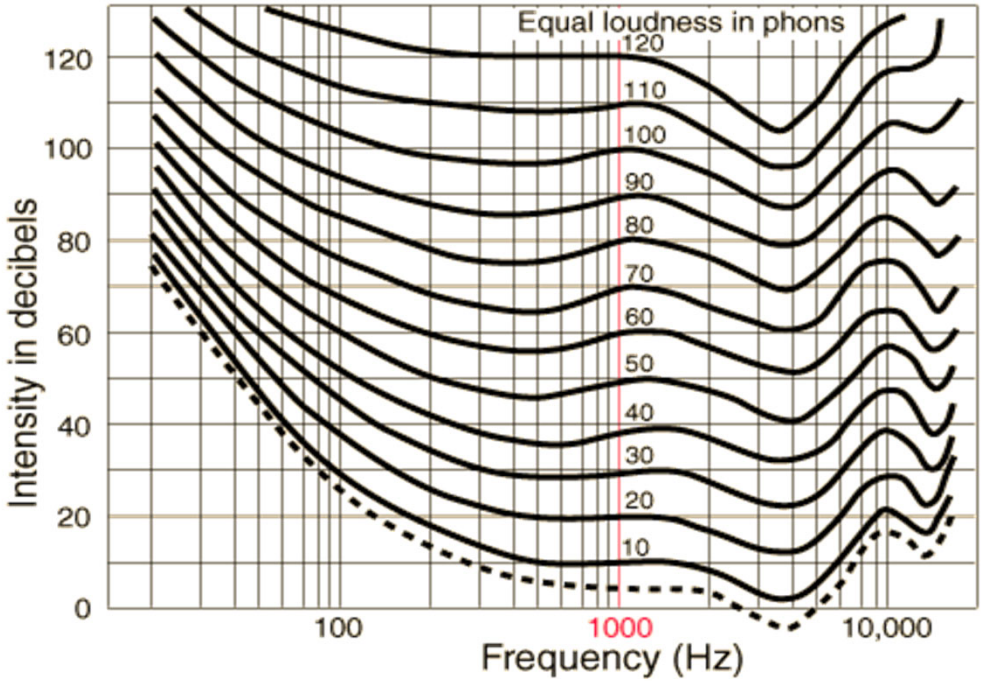
Pitch and harmonic structure

One major problem of the FFT is the Heisenberg-uncertainty



Loudness

- Perceived loudness has no-trivial connection to its physical properties
- Models exist, to computationally simulate the perceived loudness, based on psychoacoustics
- In music processing things are **simplified** by using a logarithmic scale



Timbre

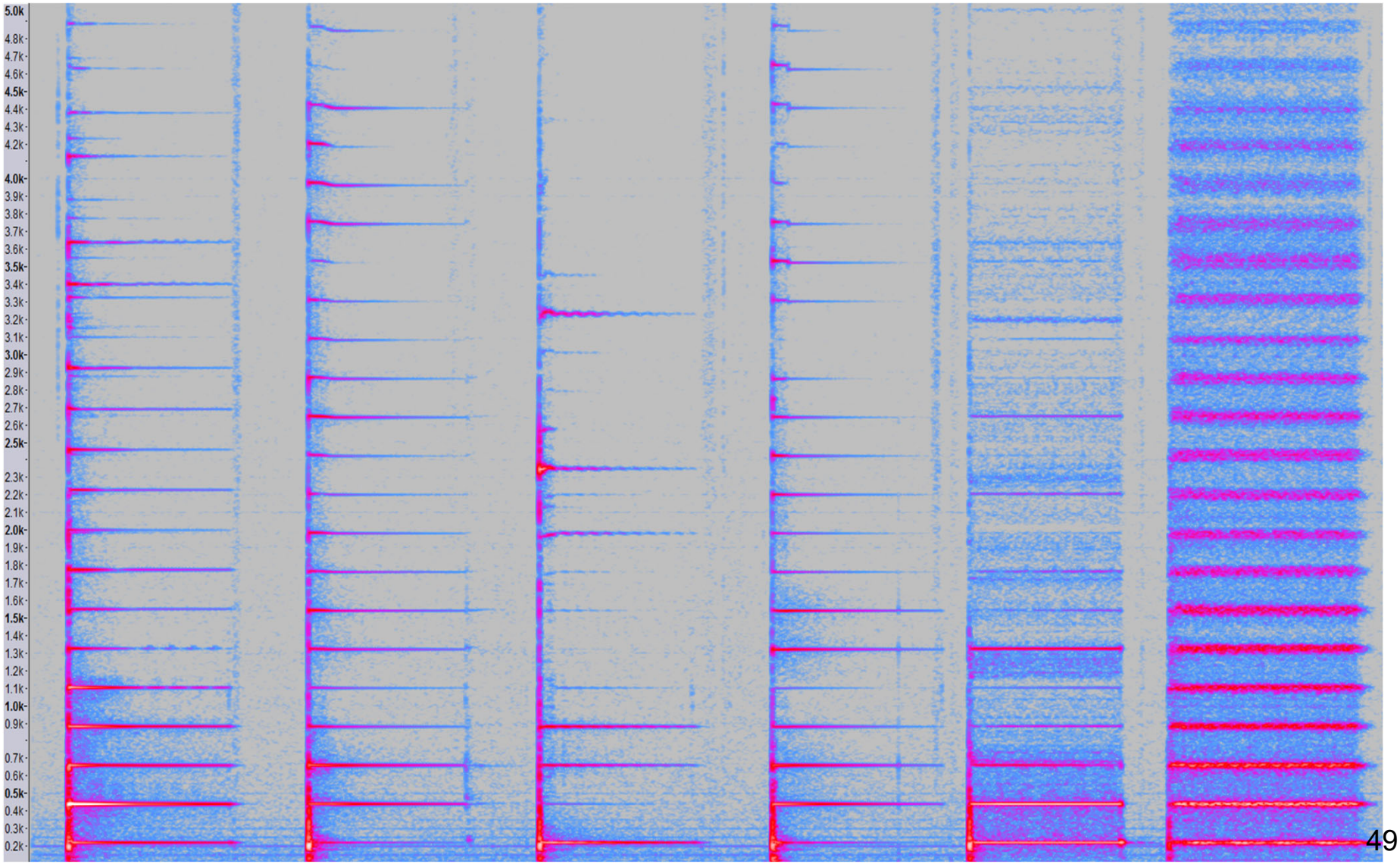
Timbre is difficult to describe and analyze, but is an active research problem

- Timbre is what distinguishes the sounds of different instruments playing the same note

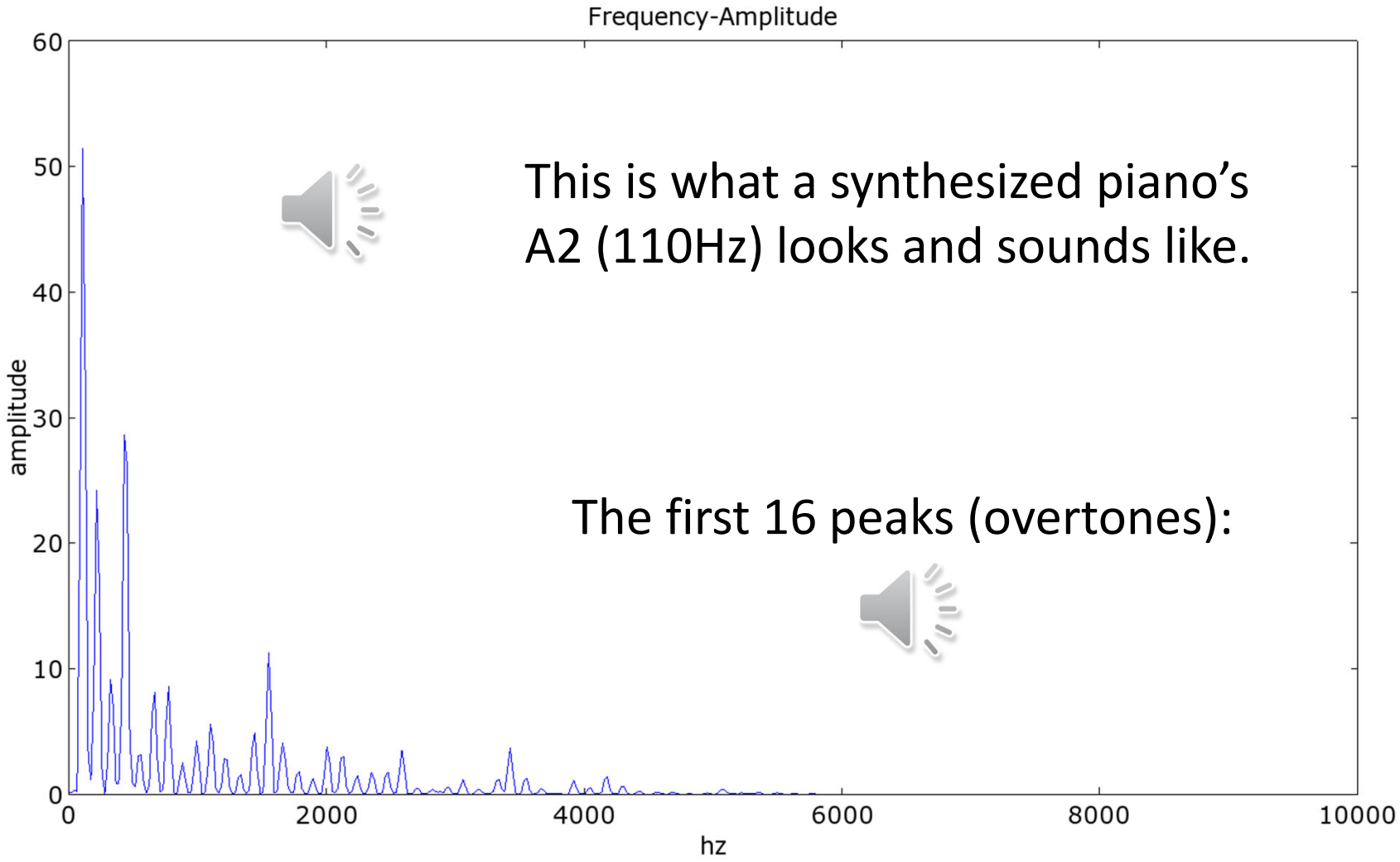
- 3 things to keep in mind about timbre:
  - The same instrument has many timbres
  - It's not just a snapshot of the spectral envelope but depends on temporal variables such as decay
  - It is intricately linked to speech analysis (*e.g., my voice quality changes during the lecture*)

Timbre

piano harpsichord vibraphone guitar organ string orch.



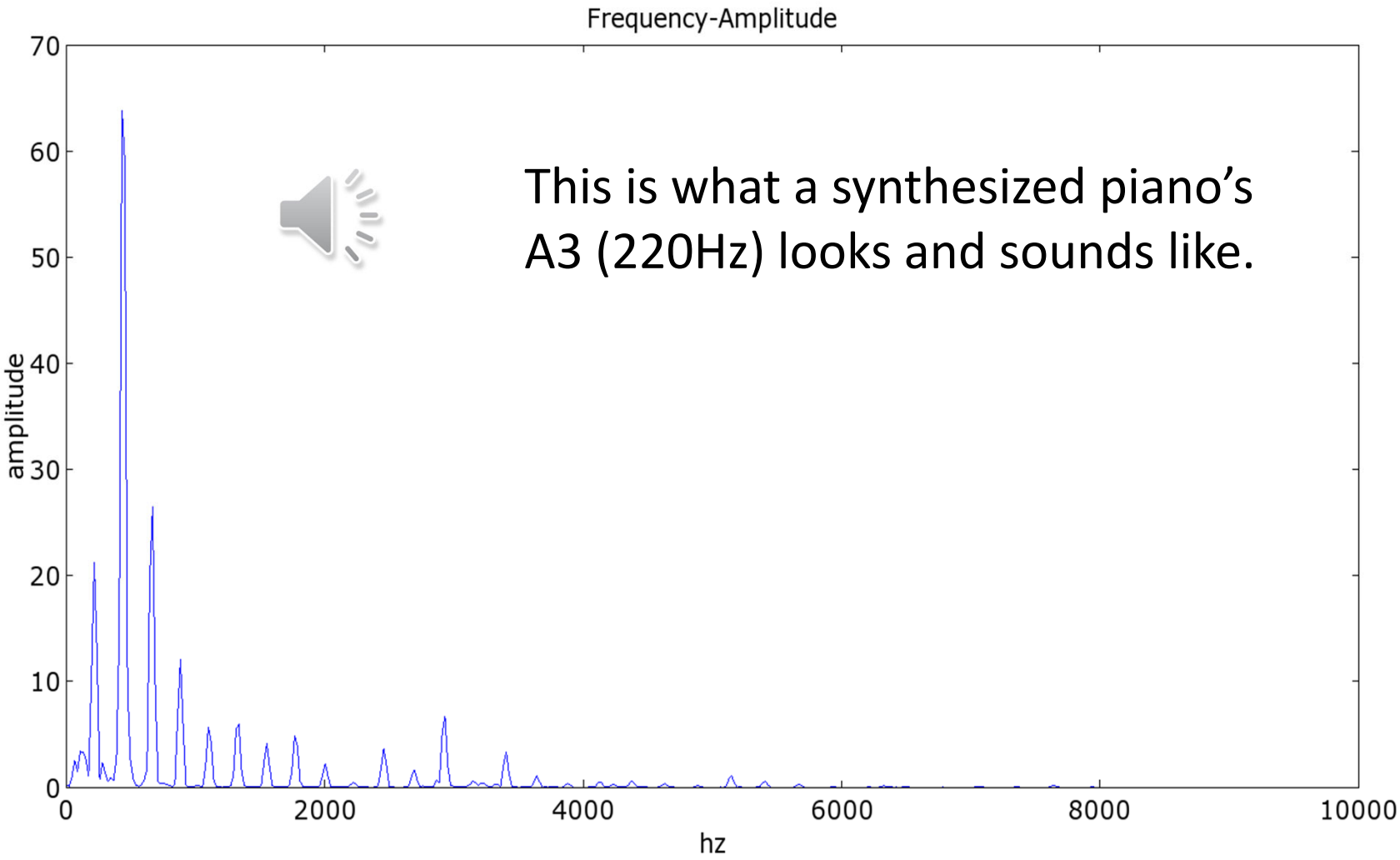
One Piano, Many Timbres



This is what a synthesized piano's A2 (110Hz) looks and sounds like.

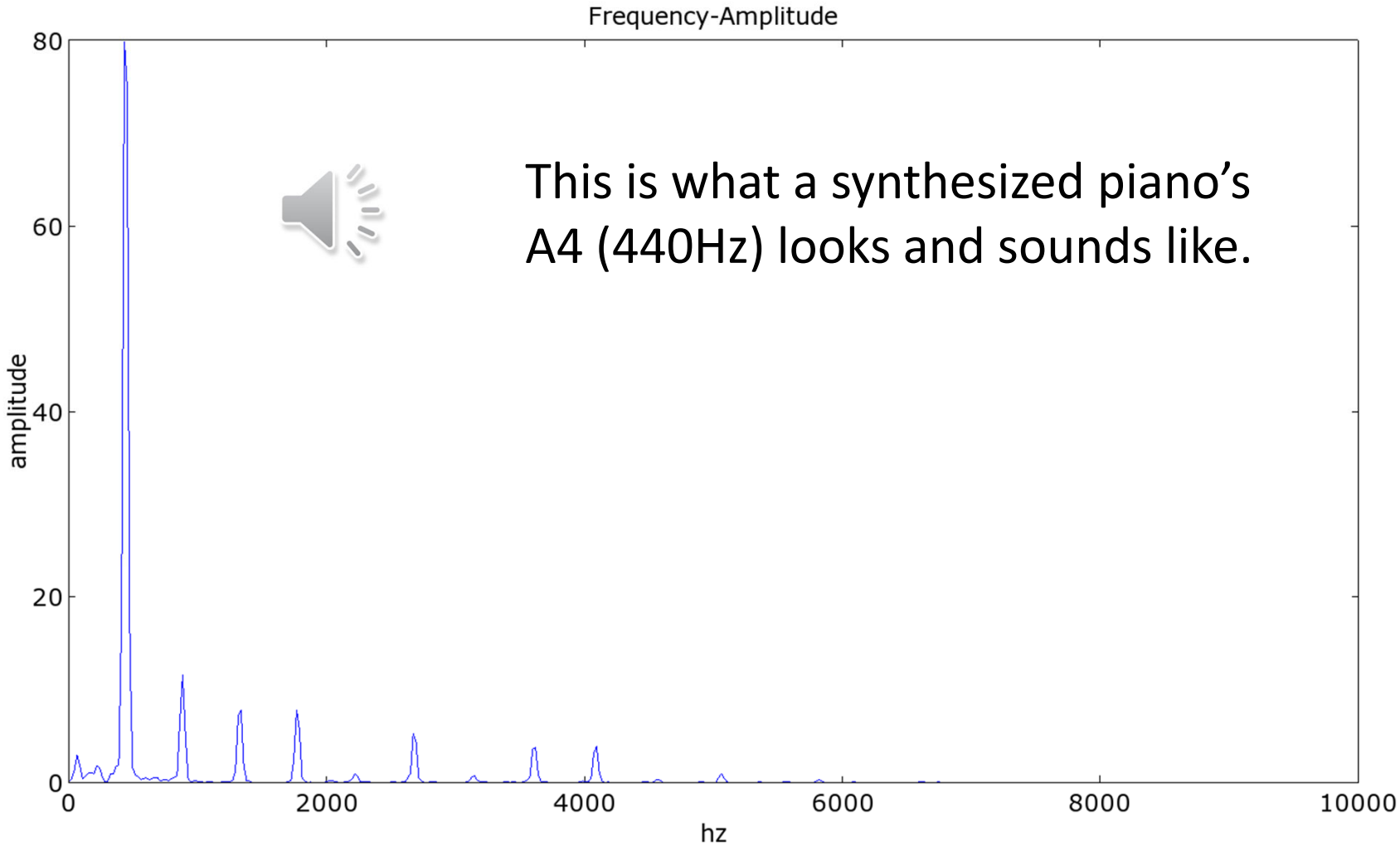
The first 16 peaks (overtones):

One Piano, Many Timbres



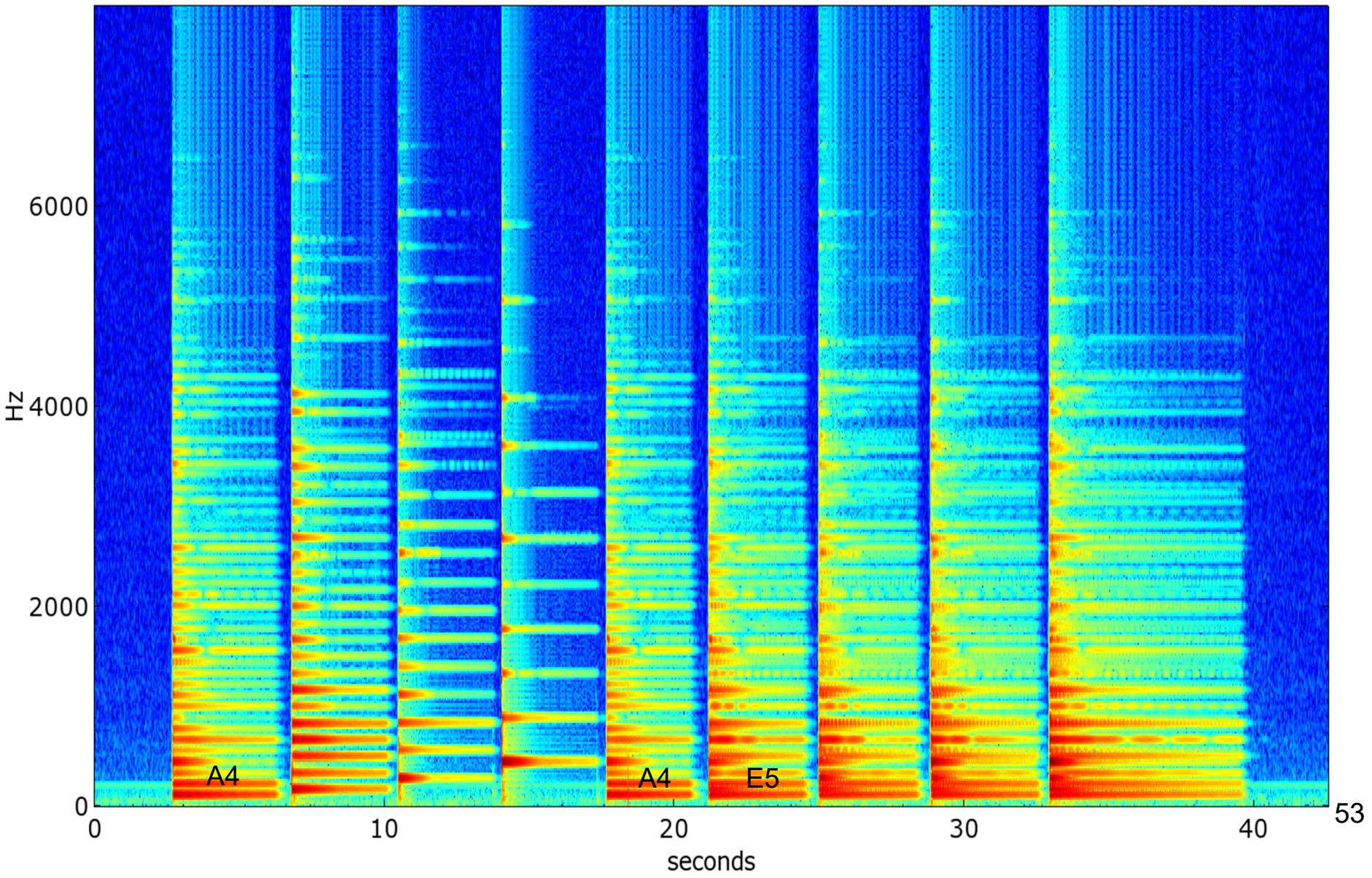
This is what a synthesized piano's A3 (220Hz) looks and sounds like.

One Piano, Many Timbres

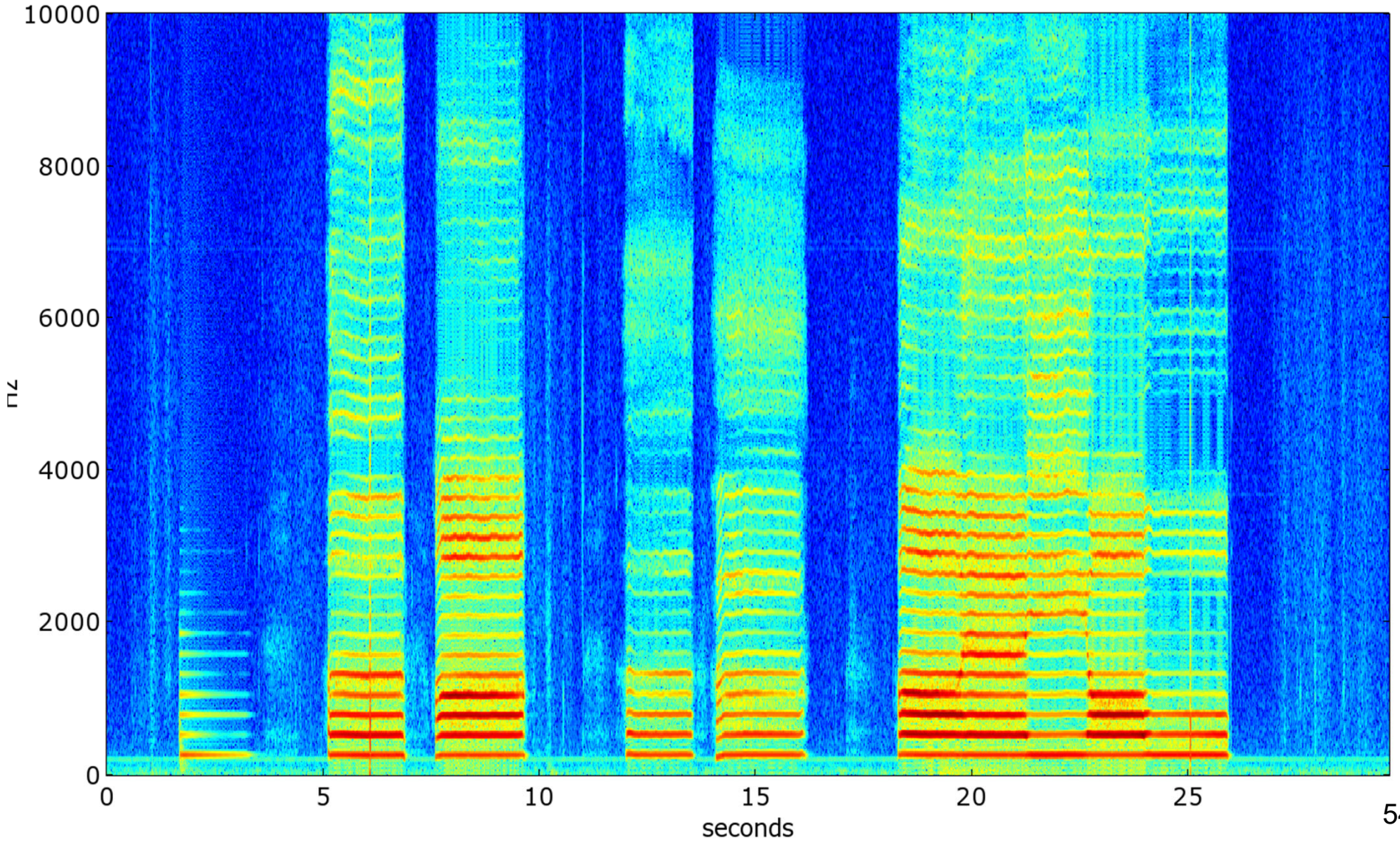


This is what a synthesized piano's A4 (440Hz) looks and sounds like.

Decays



Speech: Voices & Vowels



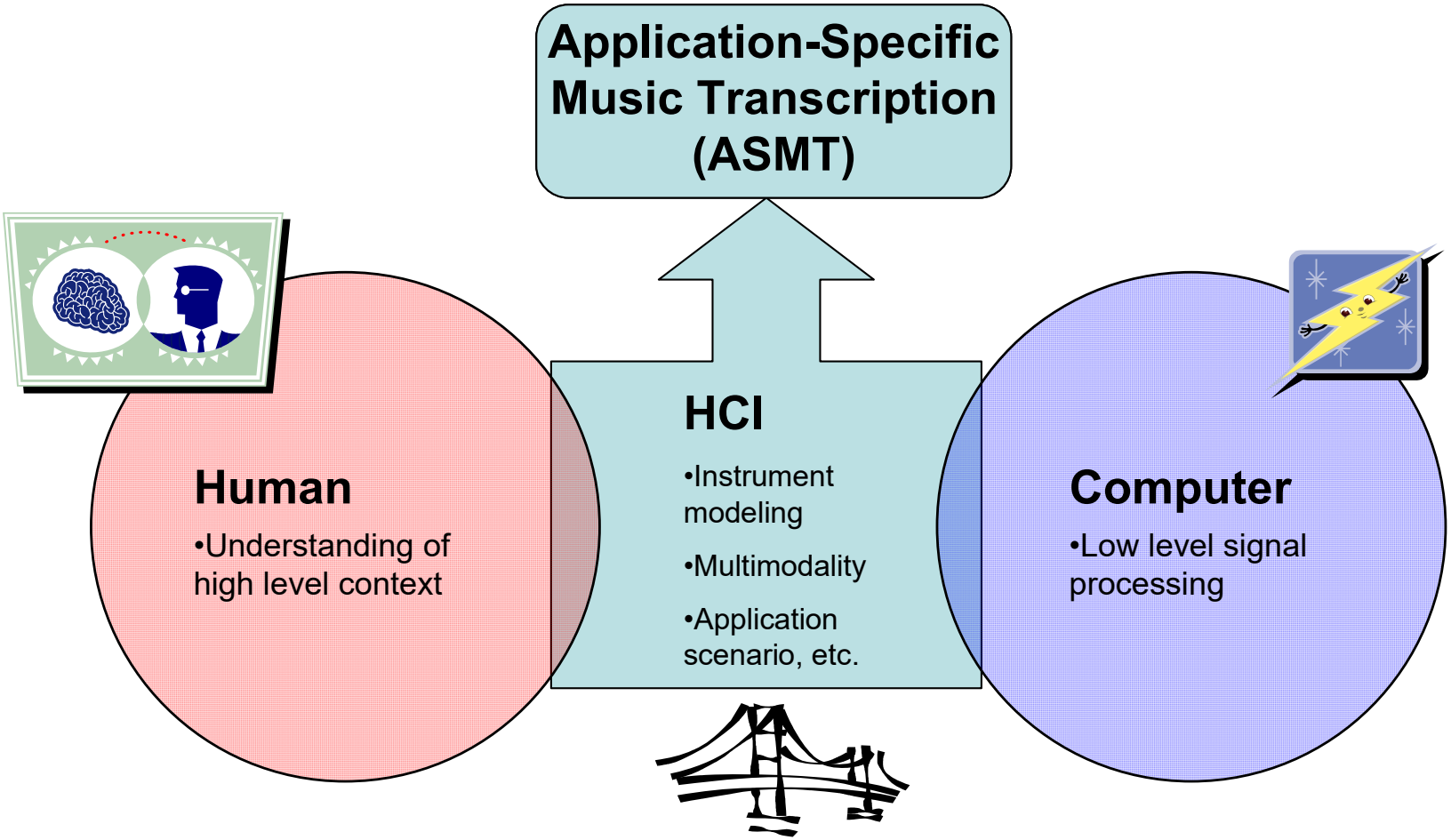
How do we process timbre?

- Each genre has its own “sounds”
- The “sound” is in large part created by the instrumentation of the piece of music
- Despite the complexity of the audio signal and the variability within an instrument, the brain can parse out each instrument, each voice.
- Machine has quite a long way to go

State-of-the-Art and Future Directions

- HCI to leverage strengths from both human and computer
- Multimedia fusion to improve transcription performance
- Instrument model to improve transcription performance
- State-of-the-Art results

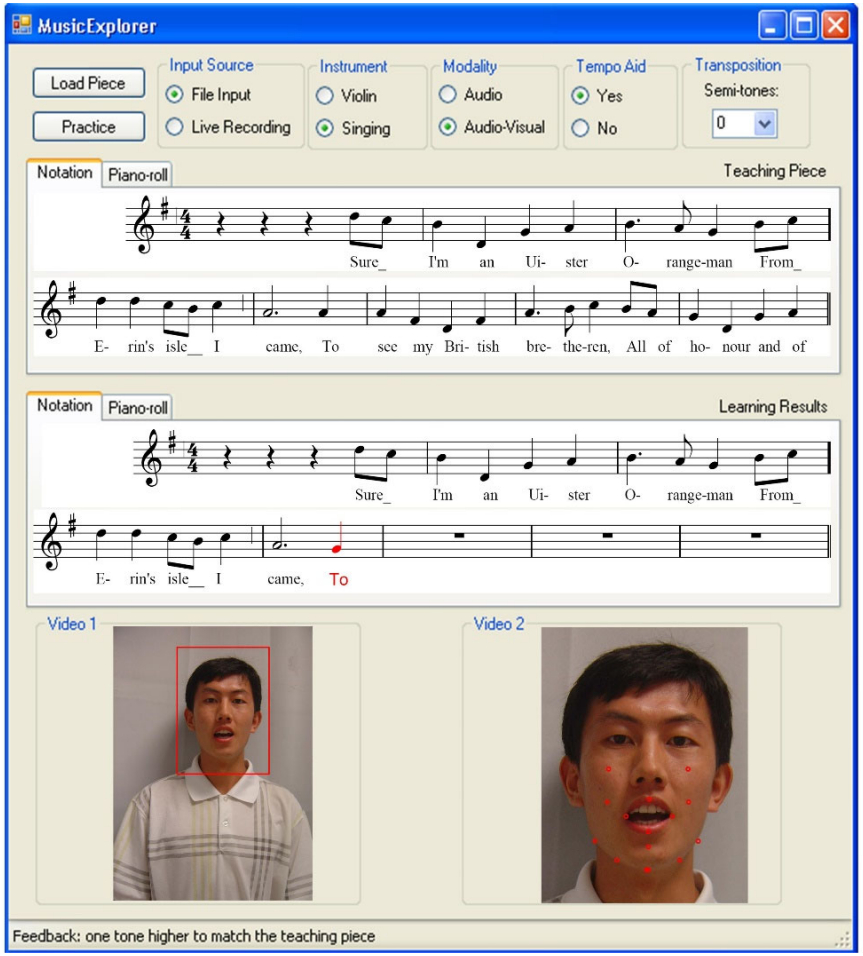
How could we address the AMT challenges?

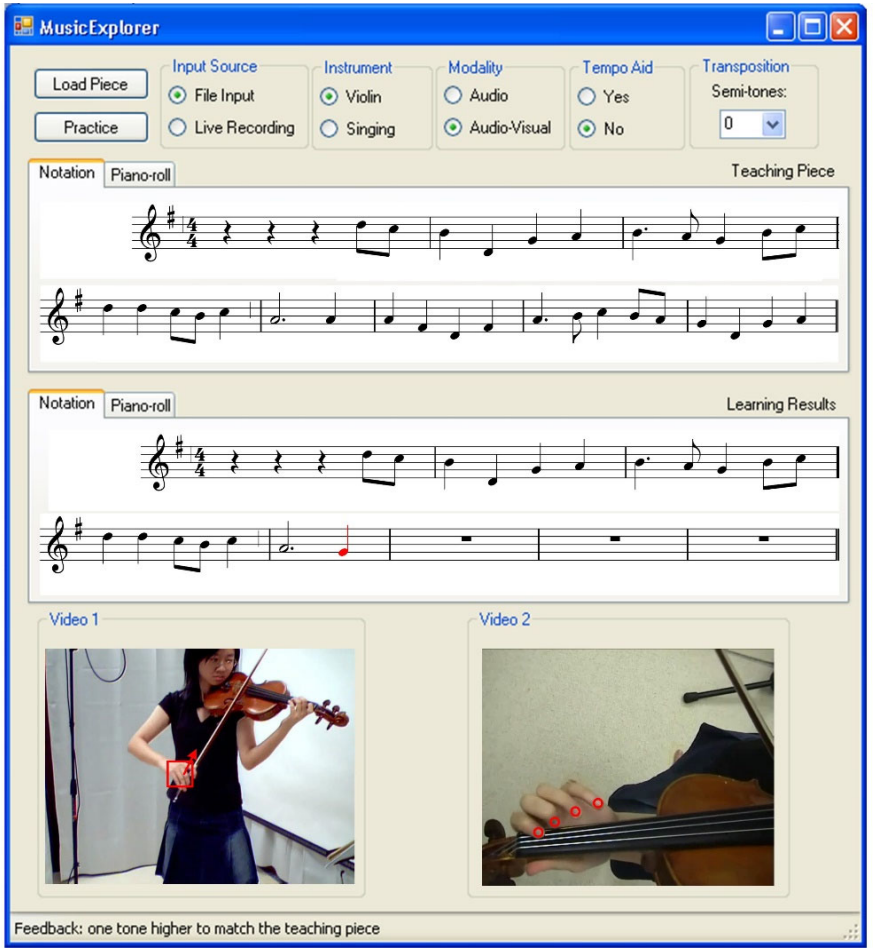


Ye Wang & Bingjun Zhang, “Application-Specific Music Transcription for Tutoring”, IEEE Multimedia, Vo.15, No.3, July-September 2008

https://smcnus.comp.nus.edu.sg/wp-content/uploads/2018/08/wang2008application.pdf

A multimodal approach

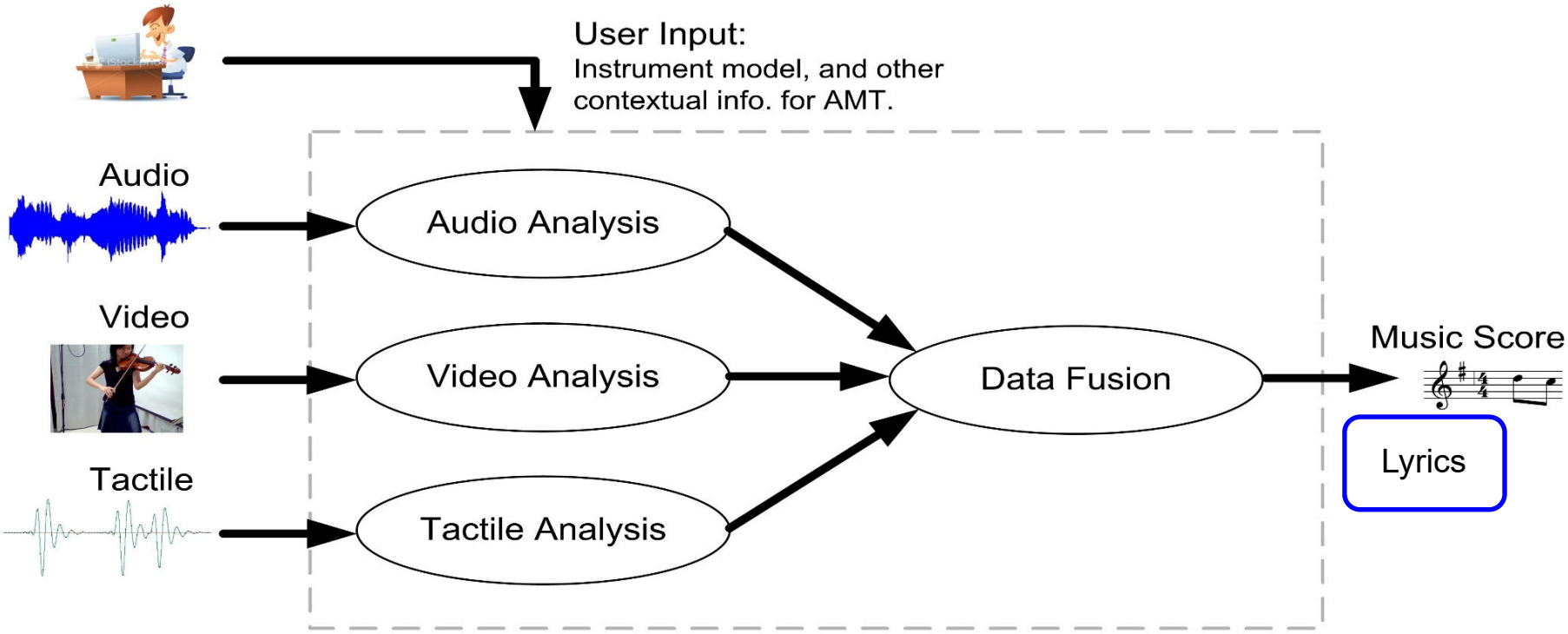




Ye Wang & Bingjun Zhang, “Application-Specific Music Transcription for Tutoring”, IEEE Multimedia, Vo.15, No.3, July-September 2008

Multimedia fusion to improve transcription

This is the theme of CS4347 group project!



Wang & Zhang, "Application-Specific Music Transcription for Tutoring", IEEE Multimedia, Vo.15, No.3, July-September 2008

Gu, Ou, Ong, & Wang, "MM-ALT: A Multimodal Automatic Lyric Transcription System", ACM Multimedia 2022

https://smcnus.comp.nus.edu.sg/archive/pdf/2022_ACM_MM_MM-ALT.pdf

Polyphonic piano music transcription (examples)

Input audio 

Synthesized audio using transcribed MIDI 

Mozart Sonata K. 331, 3rd movement

Input audio 

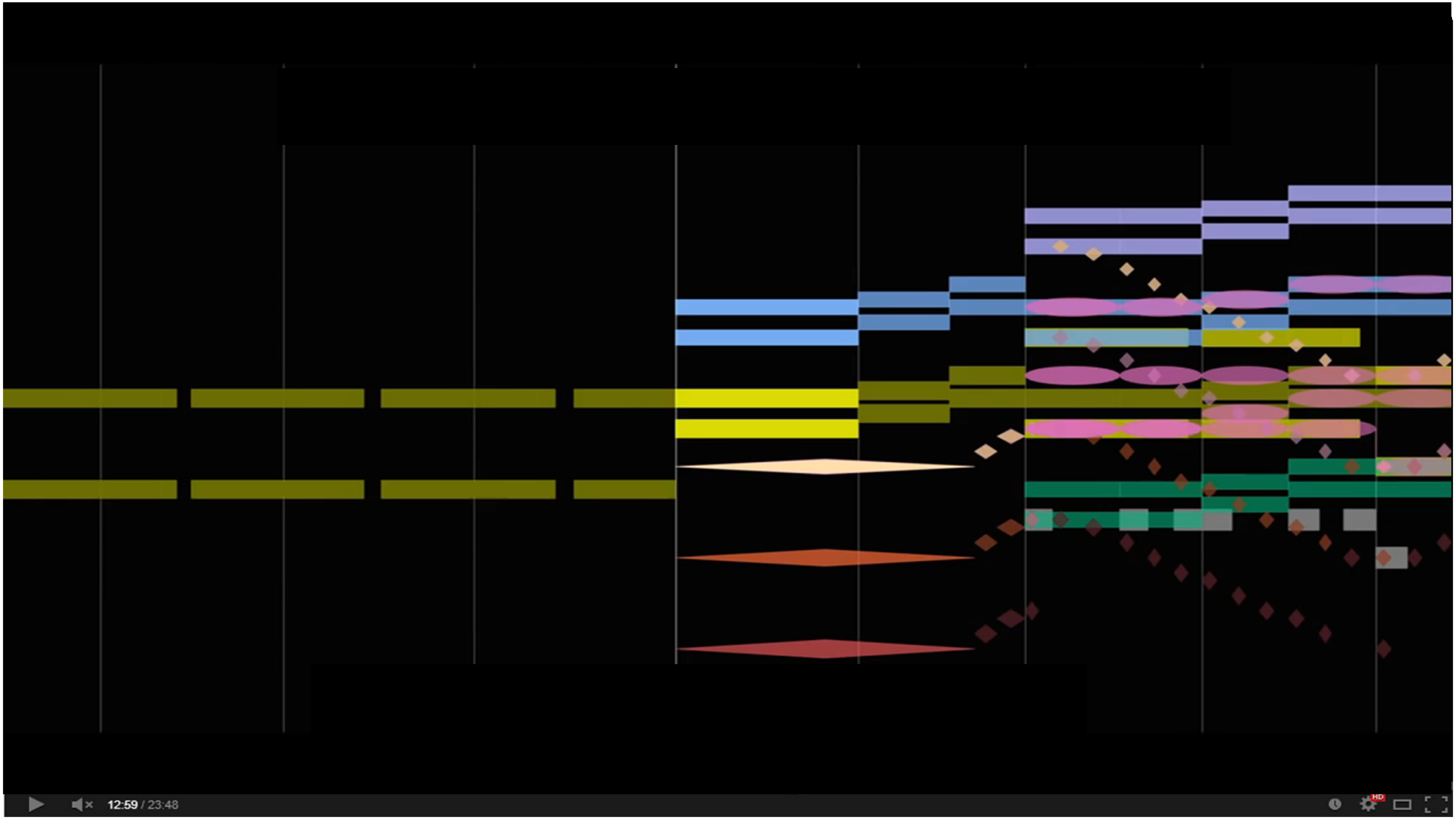
Synthesized audio using transcribed MIDI 

Chopin Etude Op. 25 No. 3

Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, Douglas Eck, Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 2018

What's in the music?

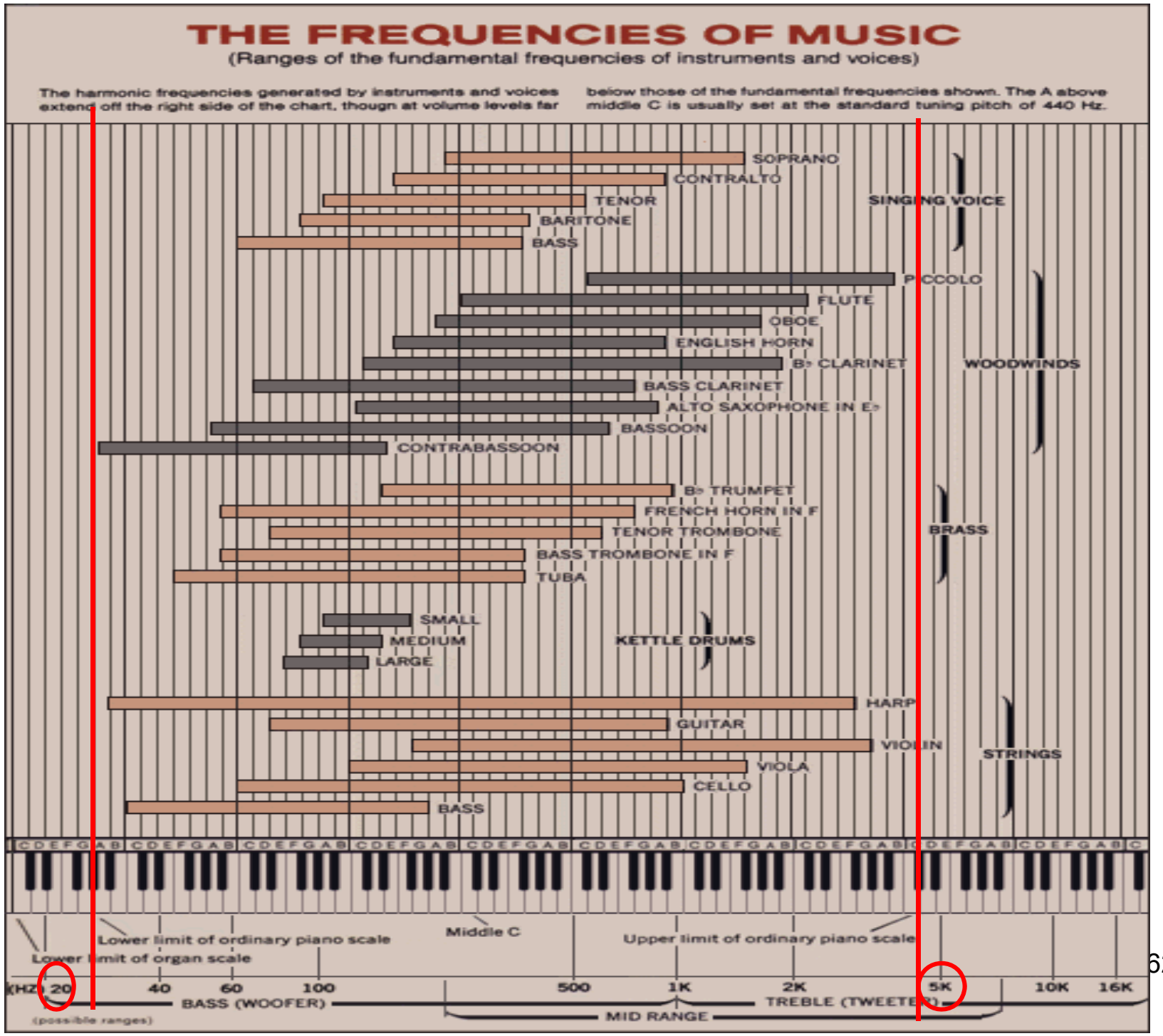
Ode to Joy from Beethoven's 9th Symphony, view 2



http://www.youtube.com/watch?v=ljGMhDSSGFU&t=12m55s

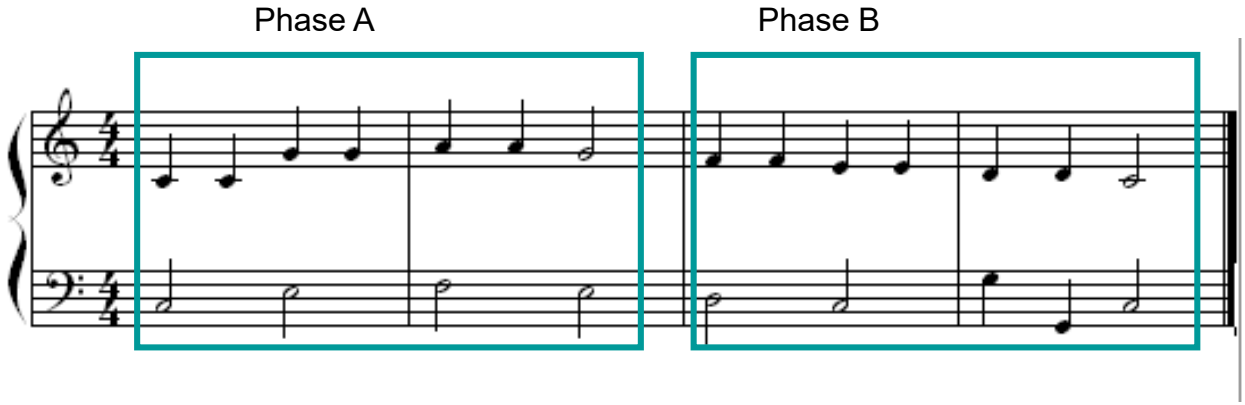
Performance was recorded and stored in MIDI format, which easily visualizes.

Can we one day do so directly from wav/mp3 format?


THE FREQUENCIES OF MUSIC
(Ranges of the fundamental frequencies of instruments and voices)
The harmonic frequencies generated by instruments and voices extend off the right side of the chart, though at volume levels far below those of the fundamental frequencies shown. The A above middle C is usually set at the standard tuning pitch of 440 Hz.
SOPRANO
CONTRALTO
TENOR
BARITONE
BASS
SINGING VOICE
PICCOLO
FLUTE
OBOE
ENGLISH HORN
B♭ CLARINET
BASS CLARINET
ALTO SAXOPHONE IN E♭
BASSOON
CONTRABASSOON
WOODWINDS
B♭ TRUMPET
FRENCH HORN IN F
TENOR TROMBONE
BASS TROMBONE IN F
TUBA
BRASS
SMALL
MEDIUM
LARGE
KETTLE DRUMS
HARP
GUITAR
VIOLIN
VIOLA
CELLO
BASS
STRINGS
C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C
Lower limit of organ scale
Lower limit of ordinary piano scale
Middle C
Upper limit of ordinary piano scale
(HZ) 20
40
60
100
500
1K
2K
5K
10K
16K
BASS (WOOFER)
TREBLE (TWEETER)
MID RANGE
(possible ranges)

What else did we miss out?

- Phrases, Sentences, Paragraphs
  - The way words are grouped together to make phrases, sentences and paragraphs is paralleled in music
  - Phrases are grouped together to form musical sentences, sentences grouped together to form musical paragraphs.
  - Eventually leading to musical Form



Twinkle Twinkle Little Star is of the form A-B-C-C-A-B

Challenge: Global structure of music beyond the low level structures of notes, rhythms
Simple case: Verses, Chorus of a pop music

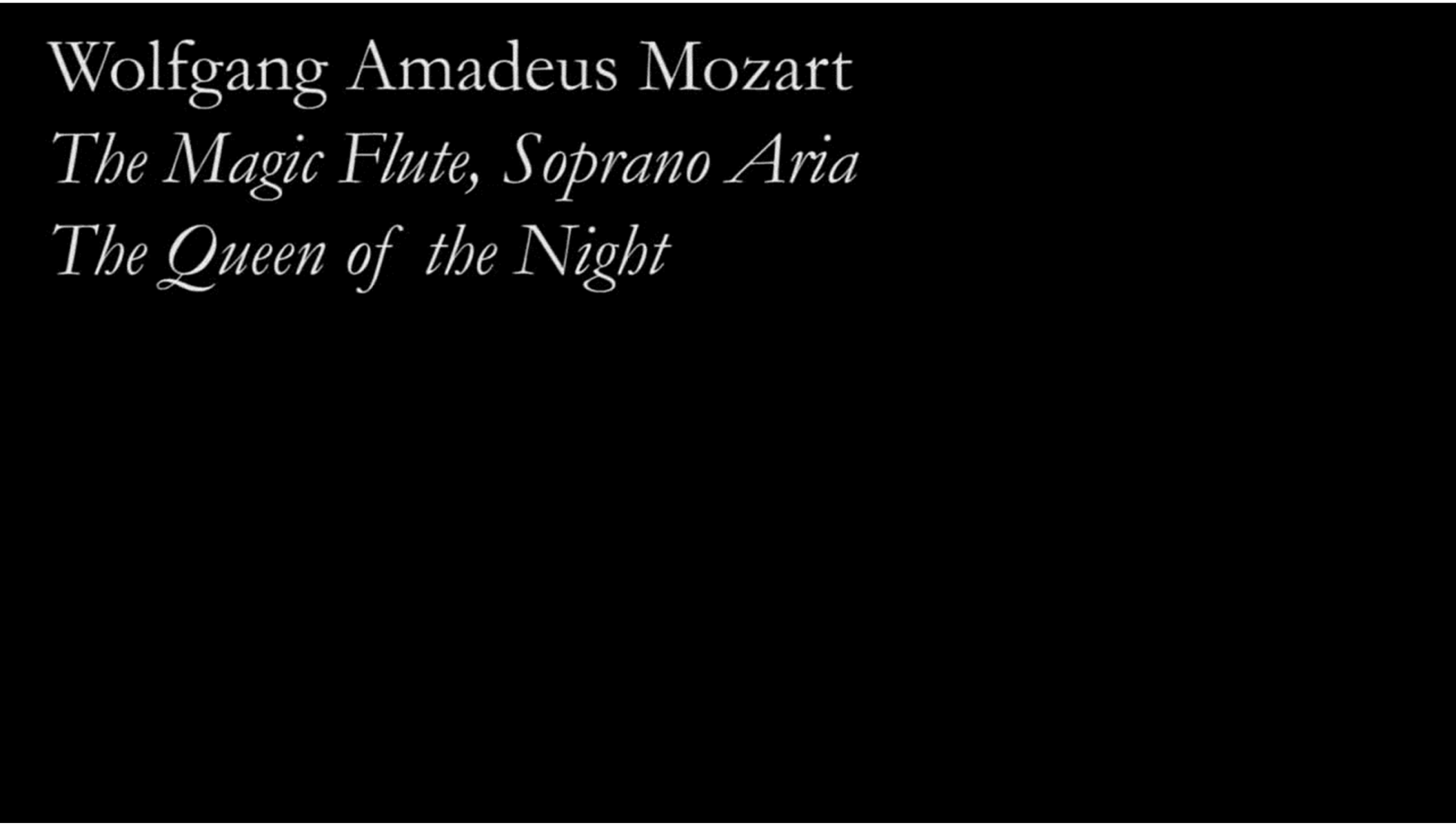
Summary of Today's Lecture

1) Recap what we have learnt last week
2) Music representations – music notation demystified
3) Music analysis (e.g. transcription)

A few words about the group project

Multimodal presentation of music

Expressive visualization of music



Multimodal presentation of music

On our class concert + technical demos