Introduction to Music Information Retrieval

Zhiyao Duan
Associate Professor of ECE and CS
University of Rochester

Presentation at WiSSAP 2023, IIT Kanpur, December 18-21, 2023


Calgary
Vancouver
Regina
Winnipeg
Victoria
Seattle
WASHINGTON
Portland
MONTANA
NORTH DAKOTA
MINNESOTA
Minneapolis
WISCONSIN
MICHIGAN
Toronto
Rochester
Ottawa
Montreal
Quebec City
NEW BRUNSWICK
PRINCE EDWARD ISLAND
NOVA SCOTIA
MAINE
VERMONT
NEW HAMPSHIRE
MASSACHUSETTS
CT
RI
New York
OREGON
IDAHO
SOUTH DAKOTA
WYOMING
Detroit
Chicago
PENNSYLVANIA
NJ
IOWA
NEBRASKA
OHIO
MARYLAND
DE
Washington
INDIANA
ILLINOIS
Indianapolis
WEST VIRGINIA
VIRGINIA
NEVADA
Sacramento
San Francisco
San Jose
UTAH
Denver
United States
Kansas City
COLORADO
KANSAS
MISSOURI
KENTUCKY
NORTH CAROLINA
Bermud
CALIFORNIA
Las Vegas
Nashville
TENNESSEE
Charlotte
SOUTH CAROLINA
OKLAHOMA
Los Angeles
ARKANSAS
Atlanta
ARIZONA
NEW MEXICO
MISSISSIPPI
ALABAMA
GEORGIA
San Diego
Phoenix
Dallas
Tucson
El Paso
Jacksonville
BAJA CALIFORNIA
TEXAS
LOUISIANA
SONORA
Austin
Houston
New Orleans
Orlando
FLORIDA
CHIHUAHUA
San Antonio
Miami
The Bahamas
Gulf of California
COAHUILA
Turks and Caicos Islands
NUEVO LEON
Gulf of Mexico
Havana
BAJA CALIFORNIA SUR
SINALOA
Monterrey
DURANGO
Cuba
Dominican Republic
Haiti
TAMAULIPAS

Audio Information Research (AIR) Lab

Machine Understanding of Sounds

MUSIC INFORMATION RETRIEVAL
- Music transcription, alignment
- Source separation
- Generation
- Interactive performance

SPEECH PROCESSING
- Separation and enhancement
- Verification and anti-spoofing
- Emotion analysis
- Diarization
- Text-to-speech
- Voice conversion

ENVIRONMENTAL SOUND UNDERSTANDING
- Sound search by vocal imitation
- Sound event detection
- Source localization
- HRTF modeling
- Smart acoustics

AUDIO-VISUAL PROCESSING
- Talking face generation
- Music performance analysis and generation
- Audio-visual source separation

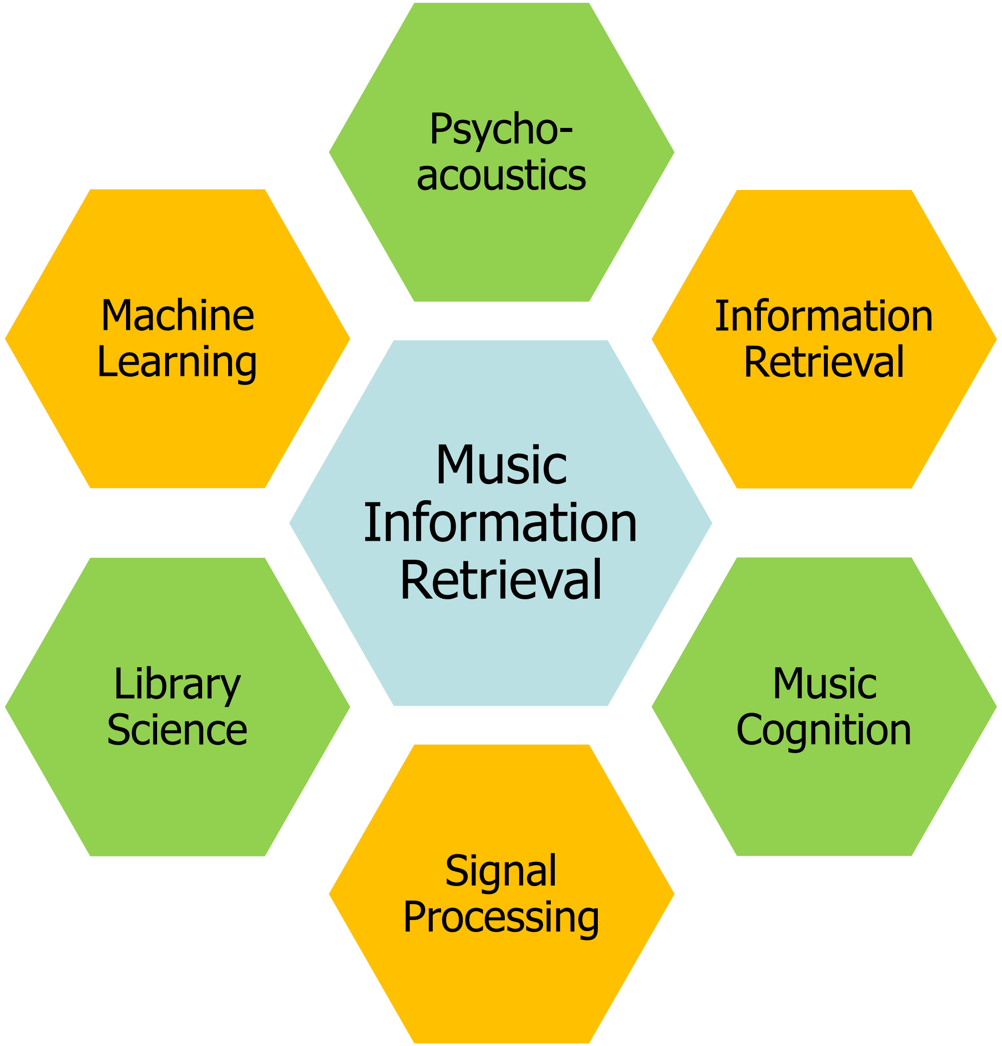
Outline

- MIR overview
- Auditory sensation
- Psychoacoustic inspirations
- Music audio features

What is Music Information Retrieval?

- Definition from https://ismir.net/about/
- MIR focuses on the research and development of computational systems to help humans better make sense of music data
- MIR draws from a diverse set of disciplines, including music theory, psychology, neuroscience, library science, computer science, electrical engineering, and machine learning

Related to Many Fields



MIR Products

Music search: Shazam, Spotify®, SoundHound Instant Music Search and Discovery, pandora®, QQ Music, gracenote A NIELSEN COMPANY

Music education: YOUSICIAN, flowkey, Violy, Tonara

Music generation: amper music, magenta, 字节跳动 ByteDance, AIVA

Music production: iZOTOPE, melodyne, CUBASE

Music Data and MIR Tasks

	Analysis	Synthesis / Generation	Relevant disciplines
User data (e.g., listening activity, EEG)	Behavior analysis, neural signal analysis	?	Psychology, neuroscience, human-computer interaction
Metadata (e.g., genre, artist, year)	Clustering, recommendation	Tagging, captioning	Library science
Video recordings	Gesture analysis, audiovisual association	Cross-modal generation	Computer vision, multimedia
Audio recordings	Transcription (melody, rhythm, chord), separation, tagging	Sound synthesis, acoustic music generation	Acoustics, signal processing
Symbolic (e.g., sheet music, MIDI)	Harmonic analysis, computational musicology	Symbolic music generation	Music theory, musicology, natural language processing

… and of course, machine learning!

Demos – Symbolic Analysis

- Python software framework for counterpoint analysis
  - Vertical Interval Successions (VIS)

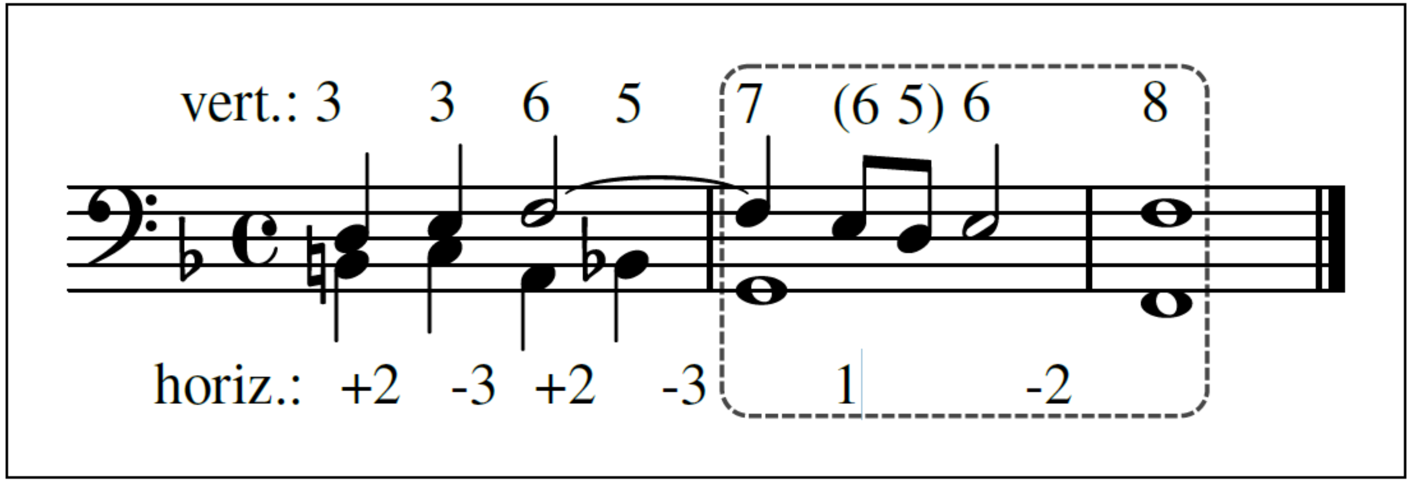


Figure 1. Symbolic score annotated with vertical and horizontal intervals. A common contrapuntal module appears in the box.

Christopher Antila and Julie Cumming. "The VIS Framework: Analyzing Counterpoint in Large Datasets", in *Proc. ISMIR*, 2014.

Demos – Audio Analysis

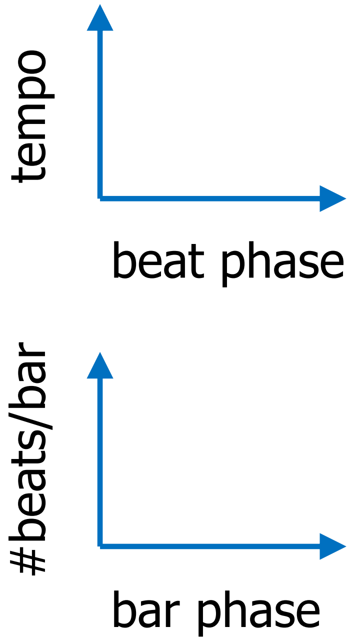
- Score following (i.e., real-time audio-score alignment) and automatic music accompaniment

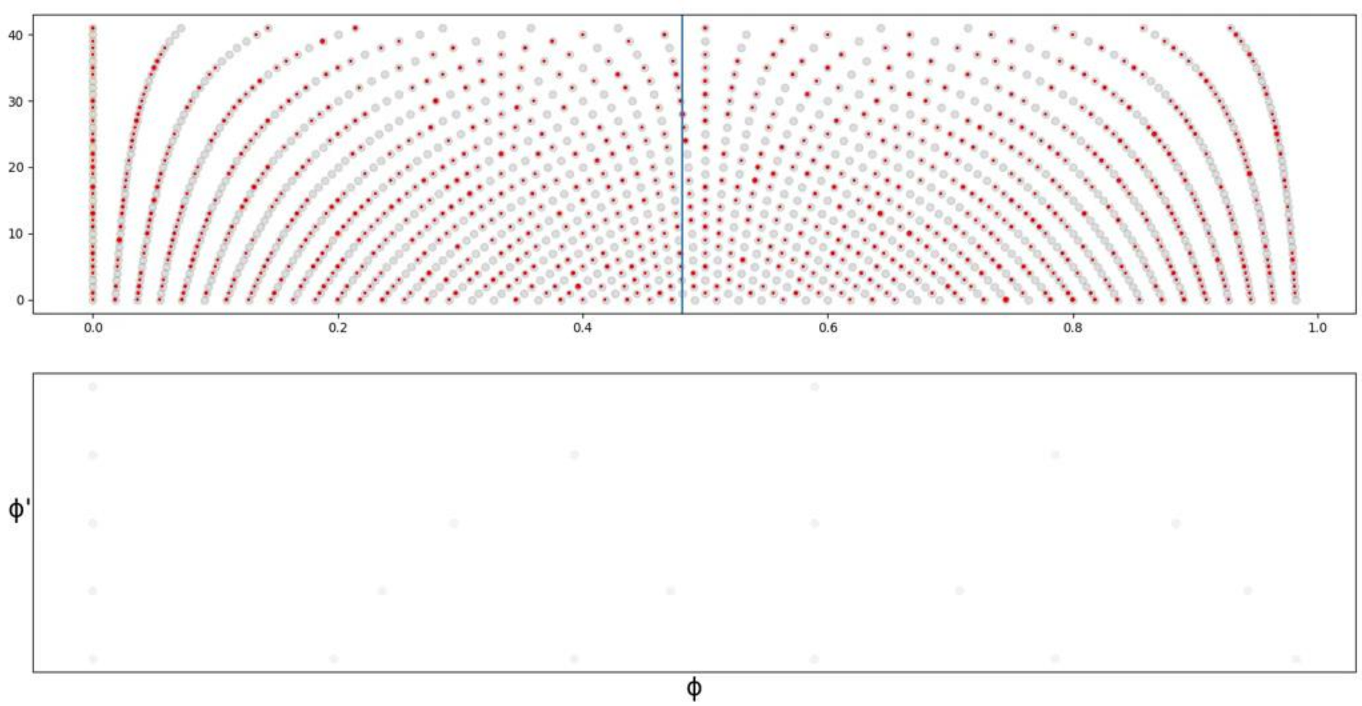
https://music.informatics.indiana.edu/~craphael/music_plus_one/movies/movies.html

Christopher Raphael, "A Bayesian network for real-time musical accompaniment," in Proc. NIPS, 2001.

Demos – Audio Analysis

- Real-time beat tracking
  - BeatNet: https://github.com/mjhydri/BeatNet

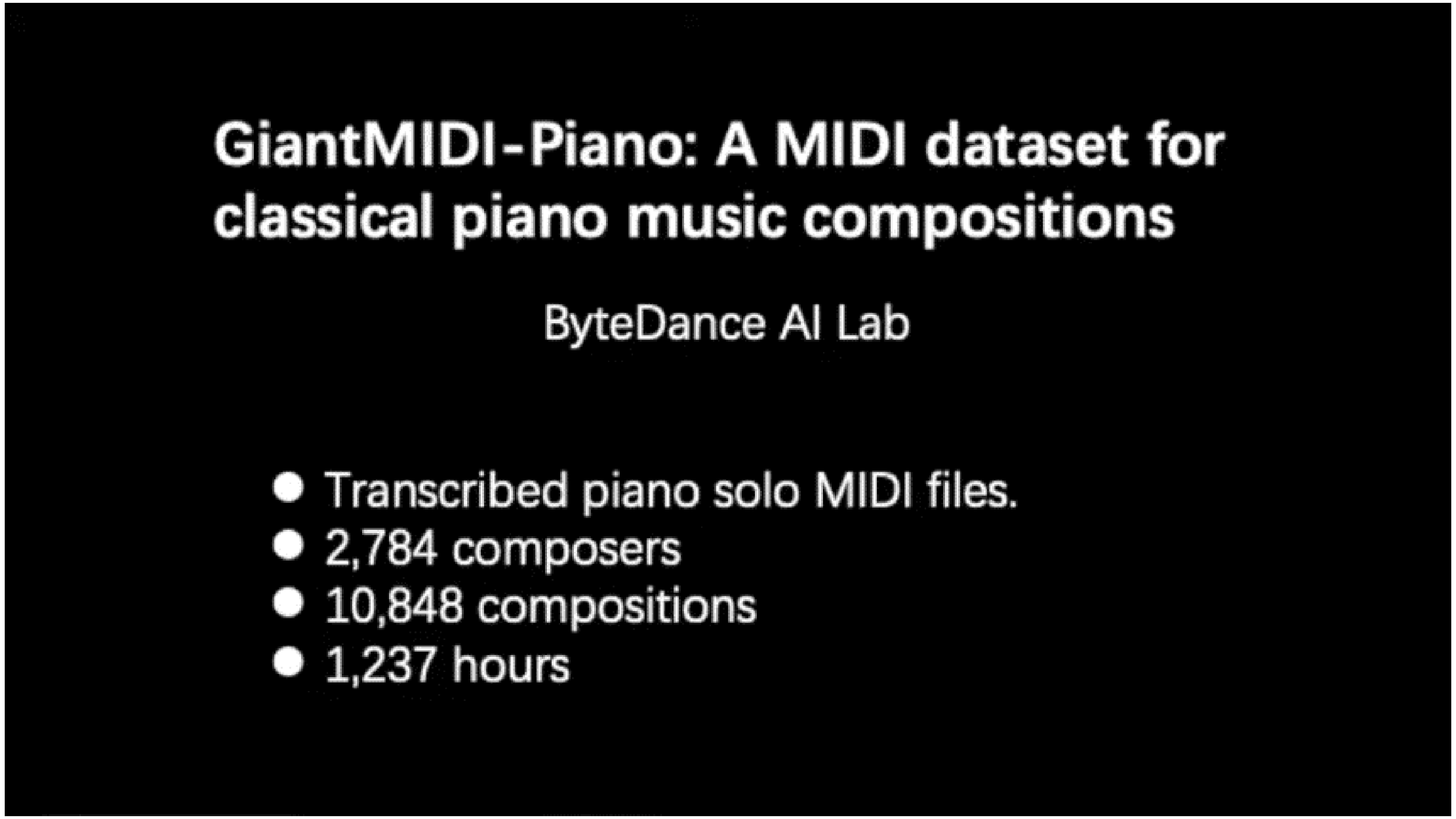




Mojtaba Heydari, Frank Cwitkowitz, and Zhiyao Duan, ”BeatNet: A real-time music integrated beat and downbeat tracker,” in *Proc. ISMIR, 2021*.

Demos – Audio Analysis

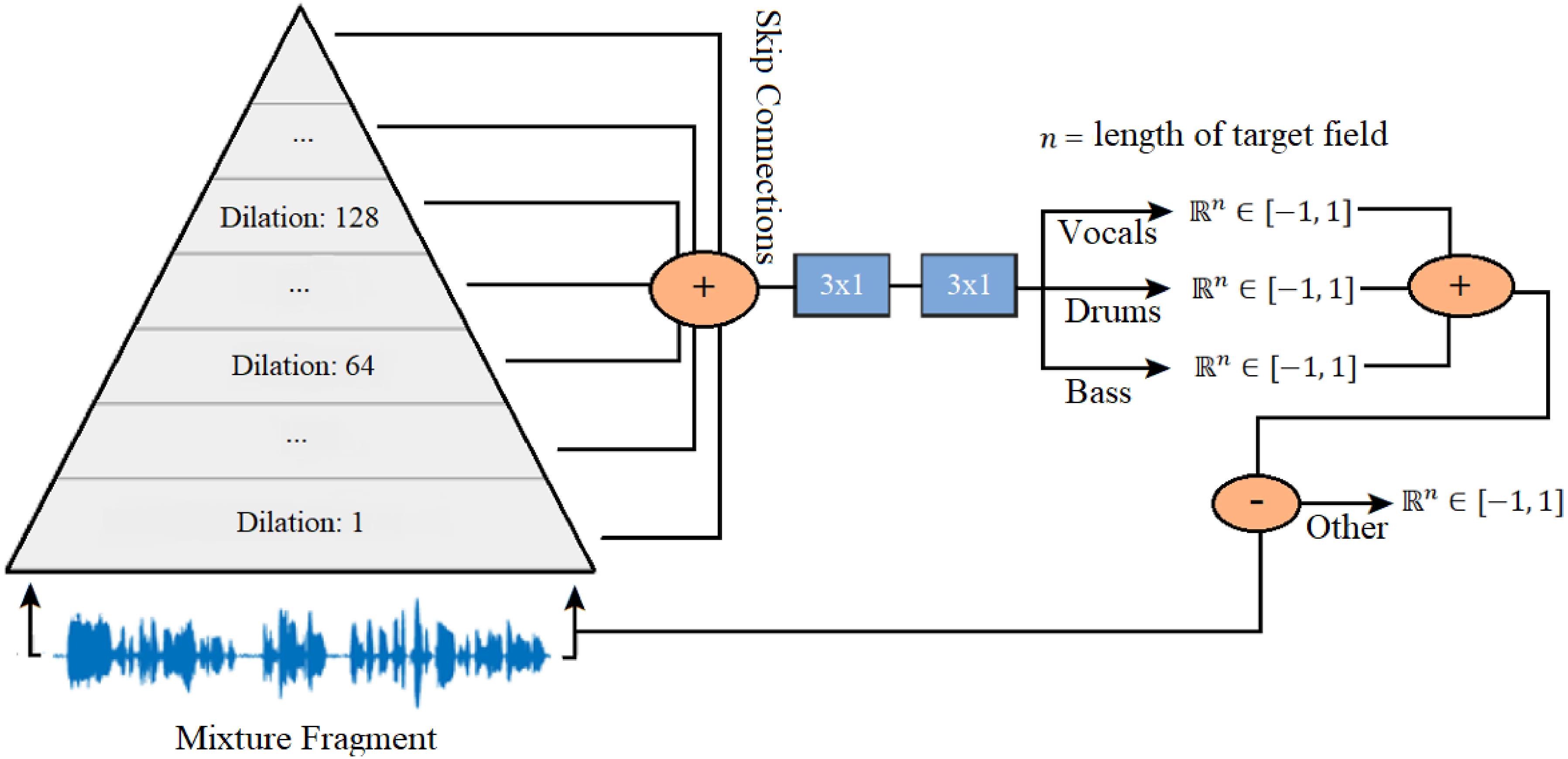
- Piano transcription



Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan and Yuxuan Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," in *IEEE/ACM TASLP, 2021*.

Demos – Audio Analysis

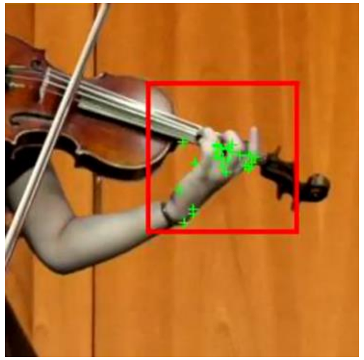
- Source separation

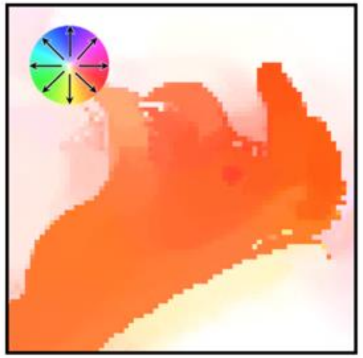


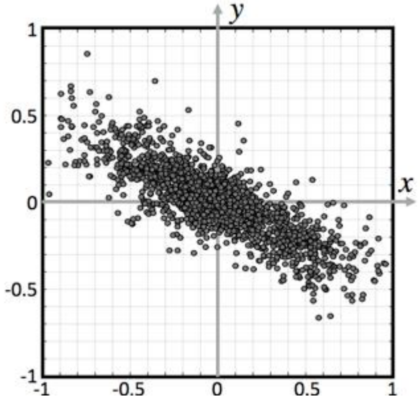
Francesc Lluís, Jordi Pons, Xavier Serra, "End-to-end music source separation: is it possible in the waveform domain?" in Proc. Interspeech, 2019.

Demos – Audiovisual Analysis

- Audiovisual vibrato detection and analysis



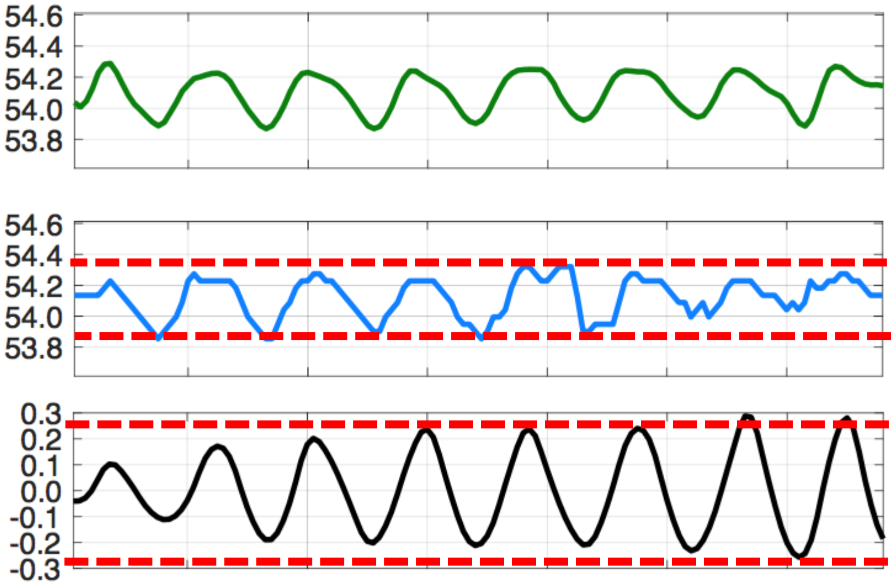




Ground-truth pitch contour

Estimated pitch contour from audio mixture

Motion curve

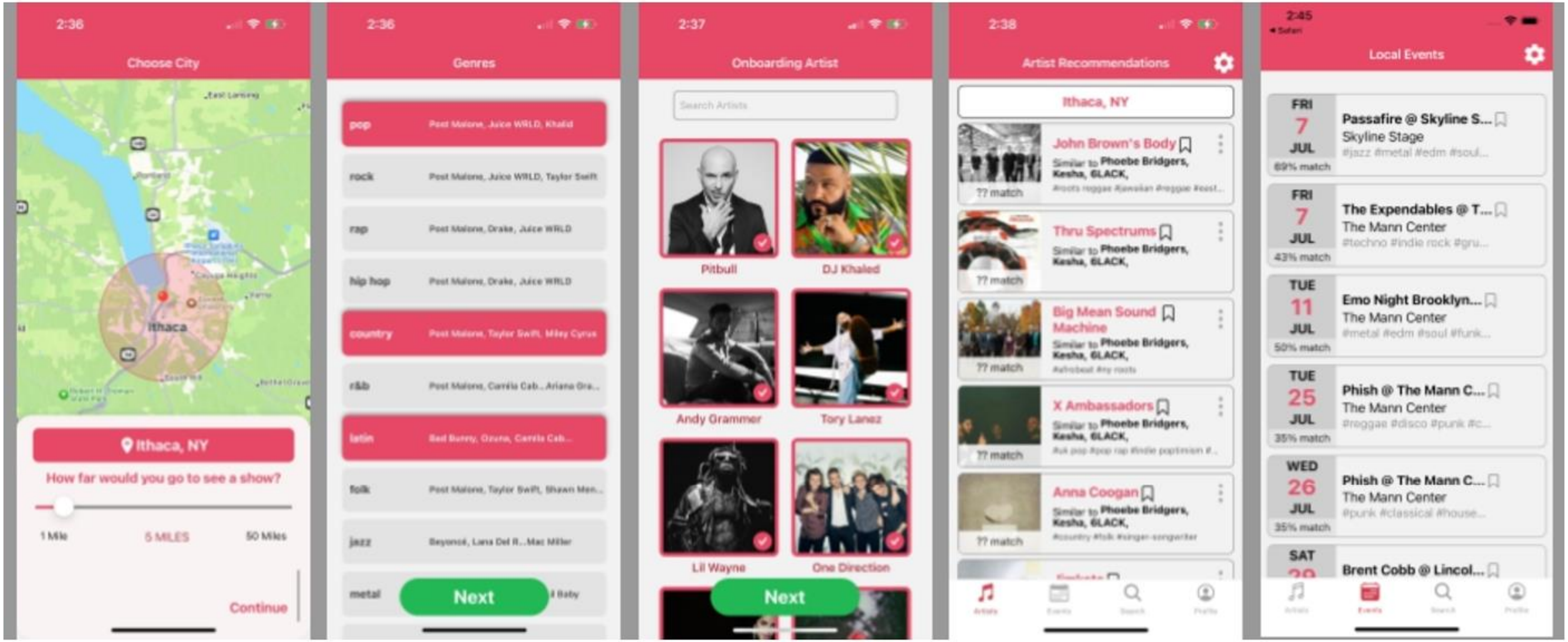


Results
- Vibrato detection accuracy significantly increased
- Very accurate vibrato rate and extent estimation

Bochen Li, Karthik Dinesh, Gaurav Sharma, and Zhiyao Duan, "Video-based vibrato detection and analysis for polyphonic string music," in *Proc. ISMIR*, 2017.

Demos – Metadata Analysis

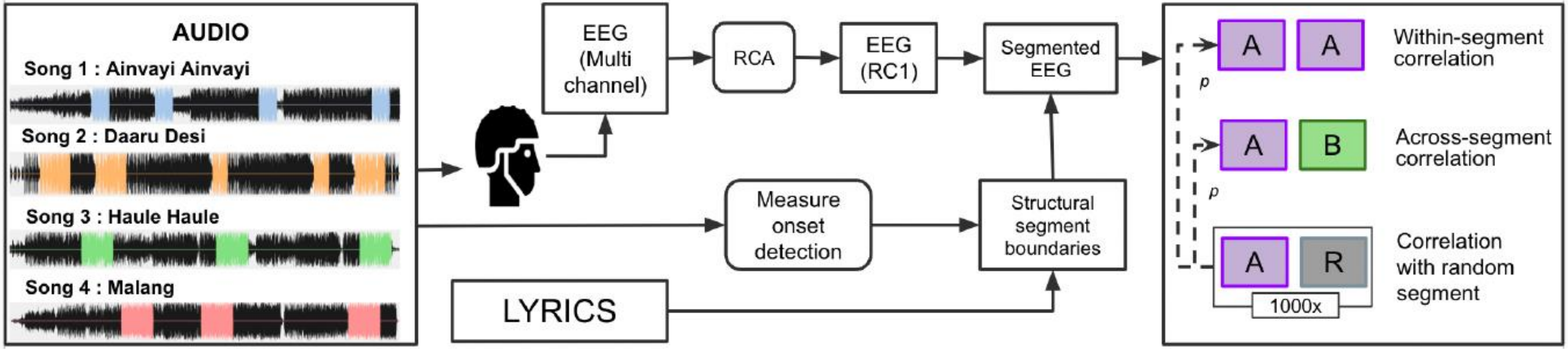
- Local artist recommendation



Douglas Turnbull, April Trainor, Douglas R Turnbull, Elizabeth Richards, Kieran Bentley, Victoria Conrad, Paul Gagliano, Cassandra Raineault, and Thorsten Joachims, "Localify.org: Locally-focus music artist and event recommendation," in *Proc. RecSys, 2023.*

Demos – User Data Analysis

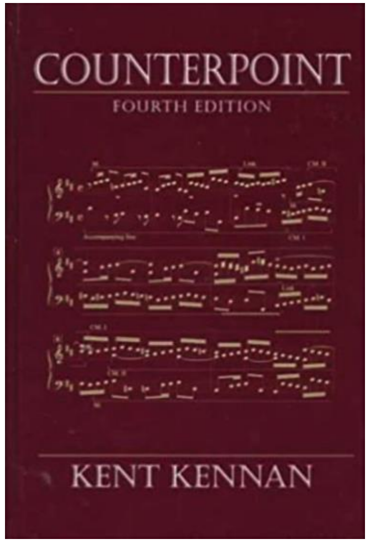
- EEG analysis of music perception (of Bollywood songs)



Neha Rajagopalan and Blair Kaneshiro, “Correlation of EEG responses reflects structural similarity of choruses in popular music,” In *Proc. ISMIR*, 2023.

Demos – Symbolic Generation

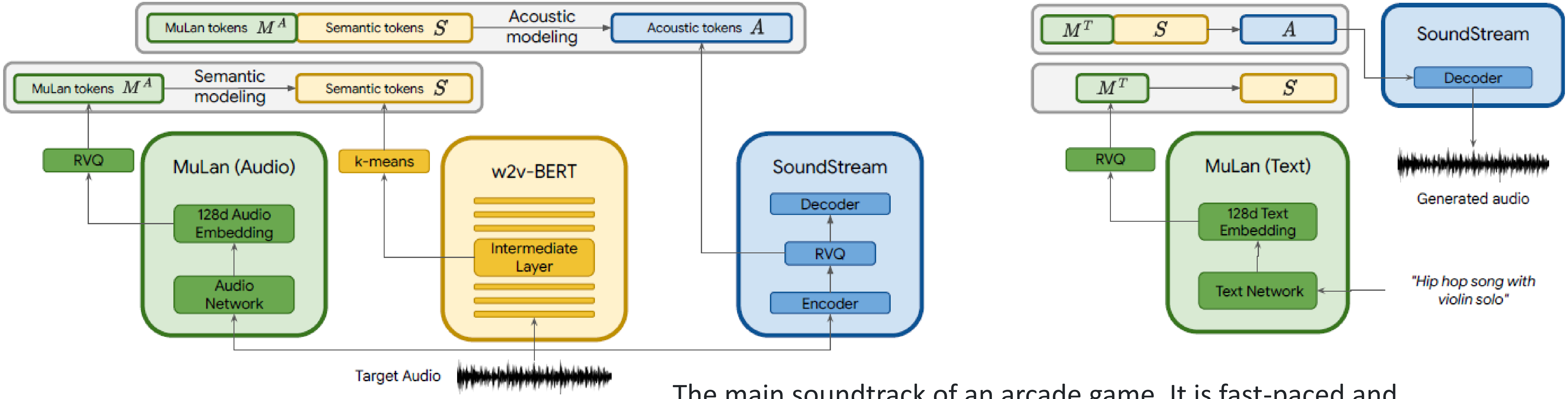
- Countermelody Generation for Chinese Folk Melodies



Nan Jiang, Sheng Jin, Zhiyao Duan, and Changshui Zhang, ”When counterpoint meets Chinese folk melodies,” in *Proc. NeurIPS*, 2020.

Demos – Audio Generation

- Text-conditioned music audio generation

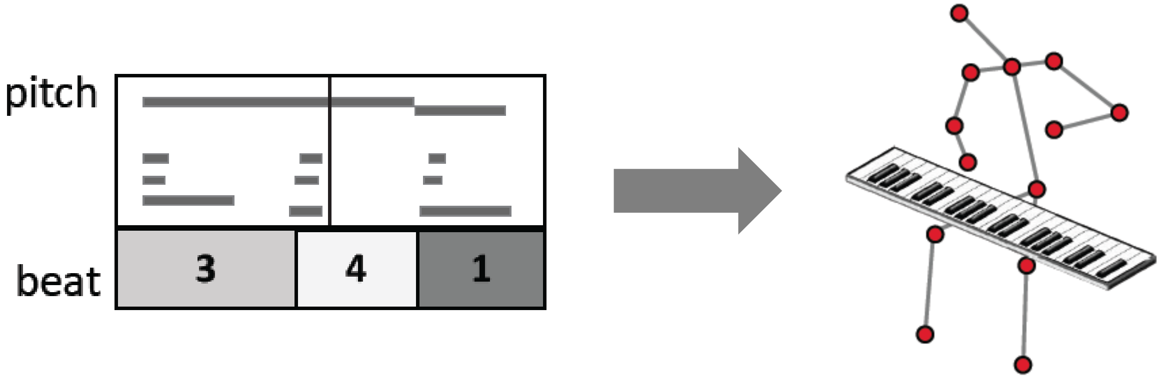


The main soundtrack of an arcade game. It is fast-paced and upbeat, with a catchy electric guitar riff. The music is repetitive and easy to remember, but with unexpected sounds, like cymbal crashes or drum rolls.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, Christian Frank, “MusicLM: Generating music from text,” arXiv:2301.11325v1, 2023.

Demos – Visual Generation

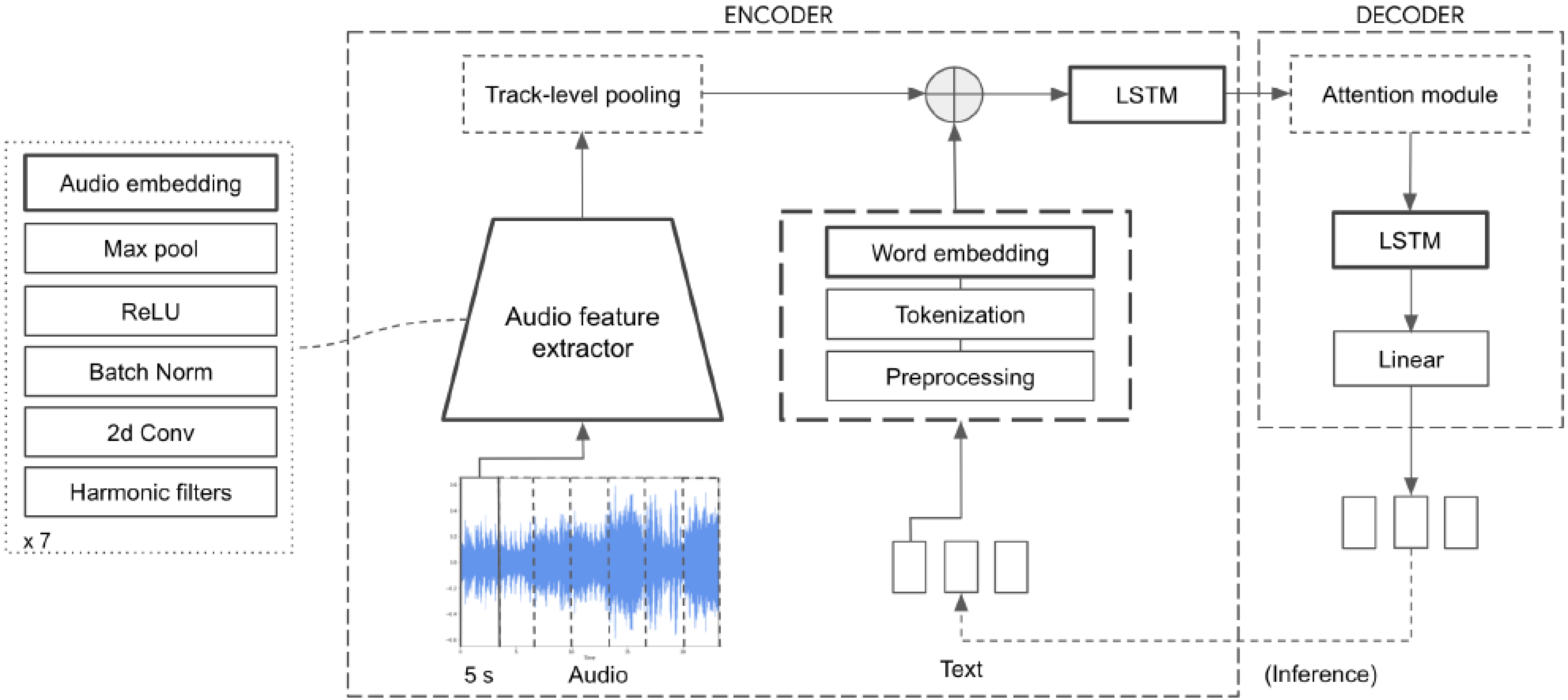
- Skeleton plays the piano



Bochen Li, Akira Maezawa, and Zhiyao Duan, ”Skeleton plays piano: online generation of pianist body movements from MIDI performance,” in *Proc. ISMIR*, 2018.

Demos – Metadata Generation

- Music Captioning



Ilaria Manco, Emmanouil Benetos, Elio Quinton, and Gyorgy Fazekas, “MusCaps: generating captions for music audio,” in Proc. IJCNN, 2021.

Trends on Research Questions in MIR

- MIR started with a strong flavor of "retrieval"
  - Query by humming, music fingerprinting, cover song detection, music tagging, music recommendation, etc.
- Transcription-related tasks took the majority since late 2000's
  - Analysis of pitch, melody, chord, beat, rhythm, structure, etc.
- Source separation has always been a key challenge
  - Singing voice separation, multi-track pop song separation
- The value of multi-modal processing kept increasing
  - Audio-score alignment
  - Audio-visual analysis and generation
  - Text-music linking
- Music generation becomes very popular in recent years
  - Symbolic: melody generation, harmonization, inpainting, continuation
  - Audio: pop song generation, text-to-music

Trends on Techniques in MIR

- 2000s and before: signal processing, hand-crafted features, traditional machine learning (e.g., SVM, HMM, probabilistic models, EM algorithm)

- Late 2000s – early 2010s: Nonnegative Matrix Factorization (NMF), Probabilistic Latent Component Analysis (PLCA)

- Mid 2010s till today: Deep learning (CNN, LSTM, CRNN, GANs, Transformers, Diffusion models)

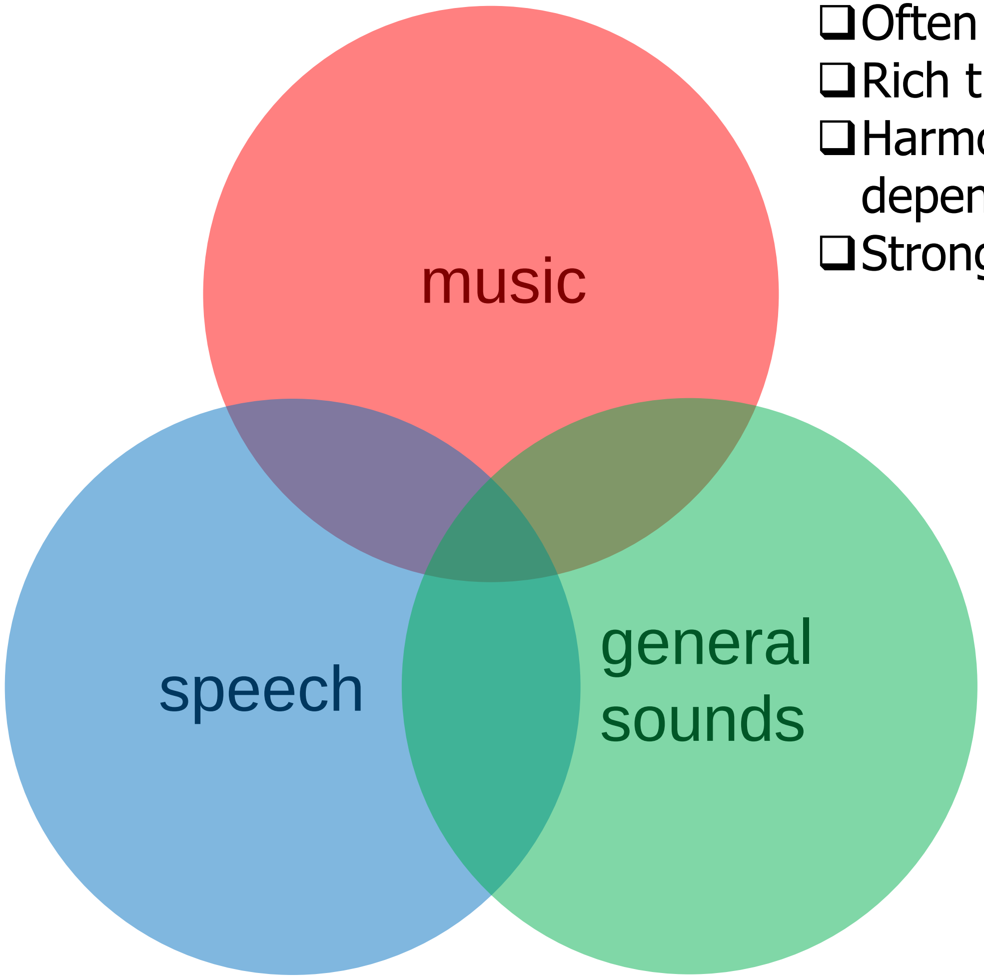
- Emerging: Large Language Models (LLMs)

MIR → Computer Audition

Music	Speech	General sound
User data (e.g., listening activity, EEG)	Style, emotion, expressiveness	Emotion, use case
Metadata (e.g., genre, artist, year)	Gender, speaker, diarization, context	Events, scene, context, location
Video recordings	Audiovisual speech	Video recordings
Audio recordings	Speech recordings	Audio recordings
Symbolic (e.g., sheet music, MIDI)	Text, transcripts	Event labels, captions

MIR → Computer Audition

- ❑ Often polyphonic
- ❑ Rich timbre
- ❑ Harmonic and temporal dependencies
- ❑ Strong reverberation

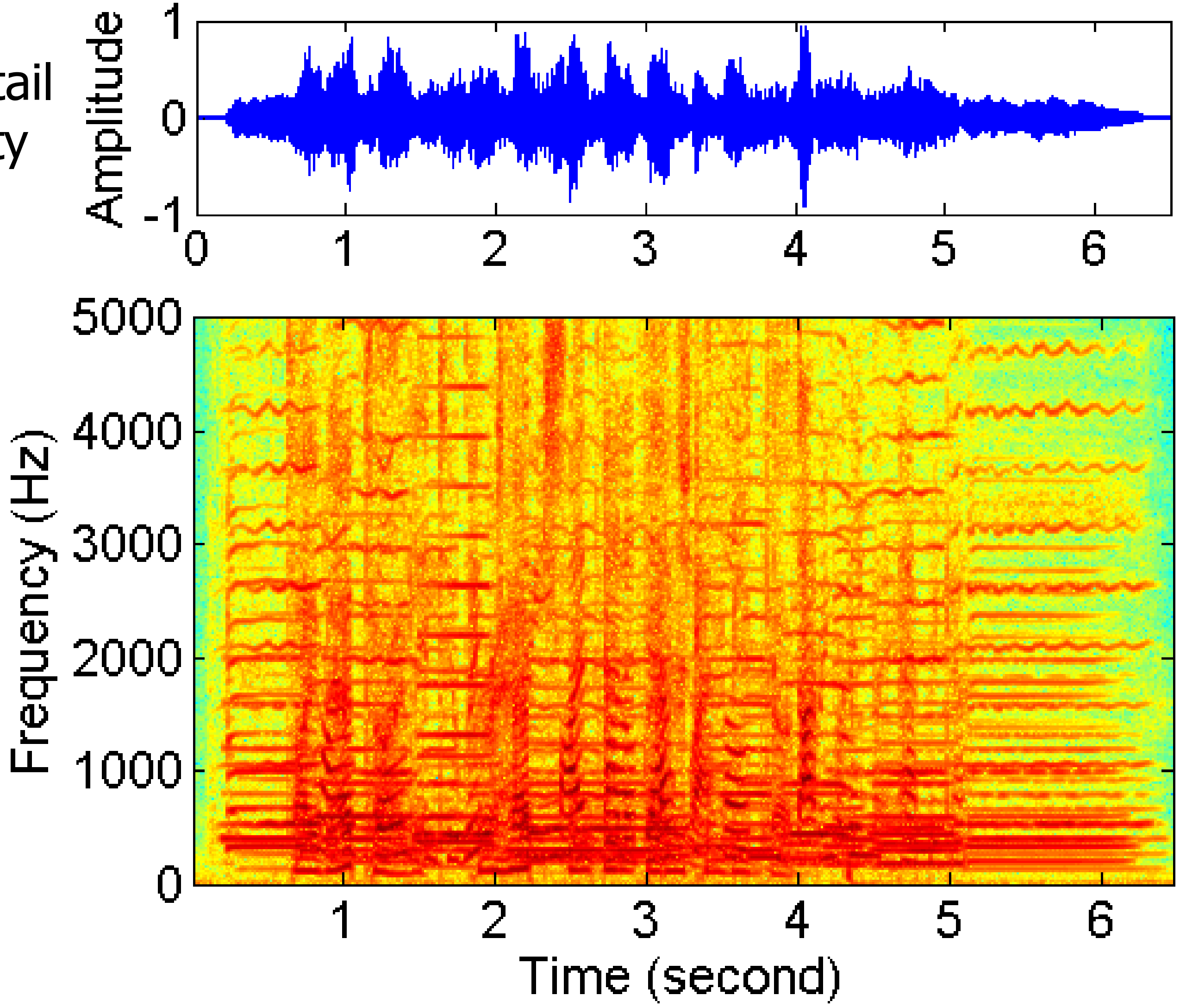


- ❑ Often monophonic but maybe overlapped by noise or other sources
- ❑ Rich semantic meaning
- ❑ Often with reverberation

- ❑ Sporadic but can be polyphonic
- ❑ Long-tail distribution
- ❑ Wide range of duration
- ❑ Less structured

Challenges – Overlapping Sources

cocktail
party



Challenges – Reverberation

- Room Impulse Response (RIR)
  - Reverberation time RT60: time takes for intensity to decay by 60 dB
    - Office ~0.5s, home ~0.7s, classroom ~1s, concert hall ~2s, cathedral ~3.5s
- 1 second is 44,100 samples at 44.1 KHz sampling rate
- Similar to motion blur for images, but with a much large "blurring kernel"

Kernel size
95*95

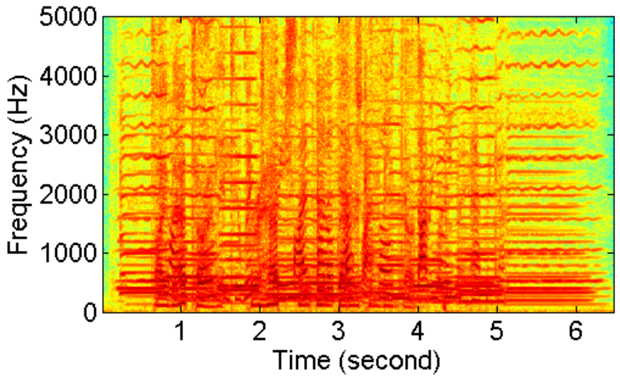
(images from http://www.cse.cuhk.edu.hk/~leojia/projects/robust_deblur/)

Challenges – Annotation

- Approach 1: annotate a real recording directly
  - Time consuming to listen through
  - Difficult to attend to simultaneous sound sources
- Approach 2: record each source in isolation and then mix them and add effects
  - Difficult to ensure synchronization and coordination
  - Still needs to annotate each source
- Approach 3: mix sound events (musical note samples) based on a transcript (musical score)
  - Requires a concatenative synthesis engine
  - Costly to obtain authentic sound samples
  - Less realistic room acoustics

Vision vs. Audition

- Visual scenes mainly describe objects that reflect light
  - Shape, color, brightness, texture, motion, etc.
- Audio scenes mainly describe sources that emit sound
  - Time, frequency, loudness, location, temporal evolution, etc.
- Visual objects occlude; auditory objects overlap
  - Analyzing audio scenes is like computer vision where
    - Objects are half-transparent
    - Objects change transparency over time
    - Objects disappear and reappear unexpectedly
    - (if with reverb) objects are all strongly motion blurred



Music is organized sound

How about contemporary music??

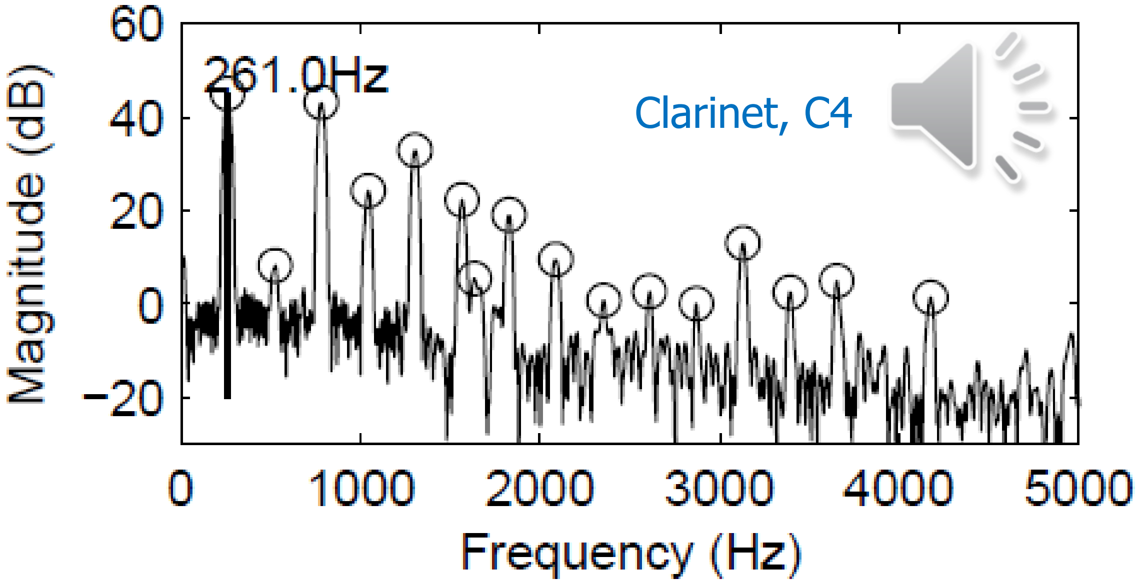
General sound (e.g., field recordings) | Animal sounds | Speech | Music

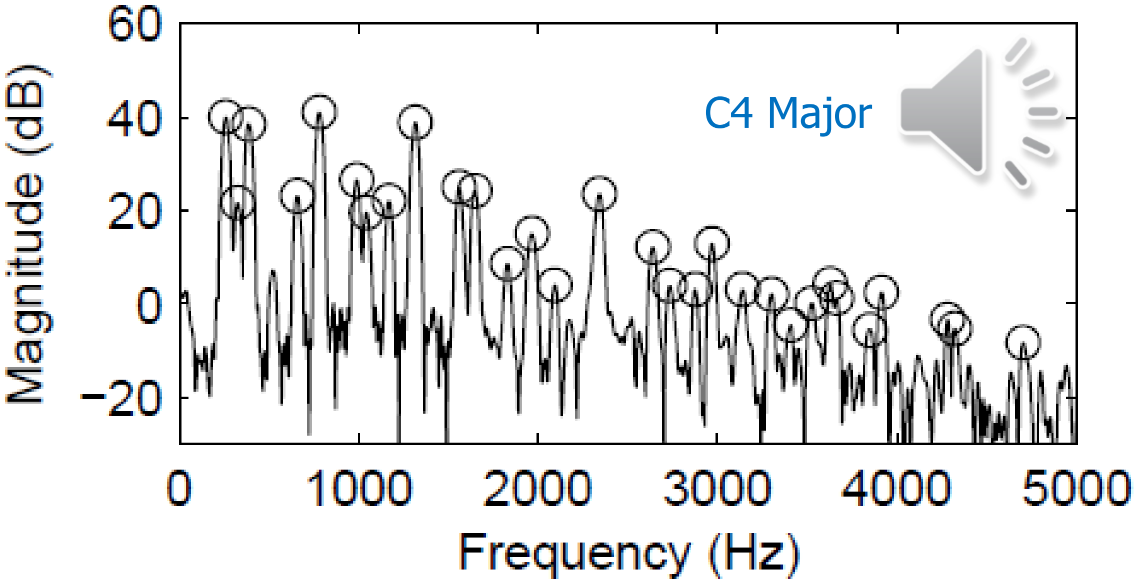
Less organized ⟷ More organized

- Harmonic structures
- Temporal structures
- Stream structures

Music is rich of structures

- Harmonic structures
  - Fundamental frequencies of simultaneous notes are often of small integer ratios, causing many harmonics of different notes to overlap with each other
    - E.g., C4:C5 = 1:2, C4:G4 = 2:3, C4:F4 = 3:4, C4:E4 = 4:5
    - For C4-E4-G4 major chord, harmonic overlap ratios are: C4 (46.7%), E4 (33.3%), G4 (60%)

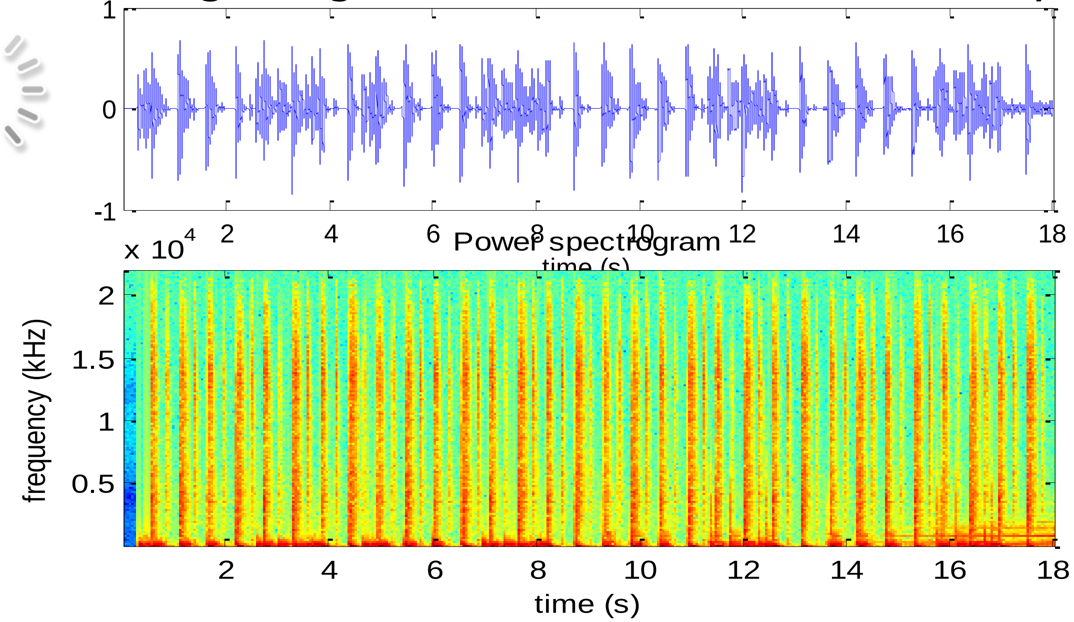


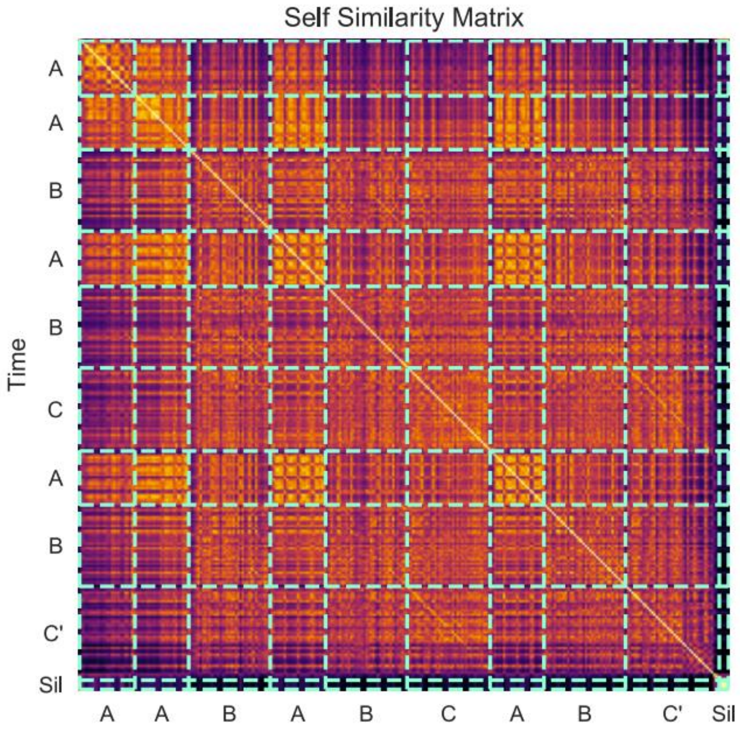


Music is rich of structures

- Temporal structures
  - Repetitions and variations at different time scales: section, phrase, measure, beat

Beginning of *Another one bites the dust* by





Oriol Nieto, Gautham J. Mysore, Cheng-i Wang, Jordan B. L. Smith, Jan Schlüter, Thomas Grill and Brian McFee, “Audio-based music structure analysis: Current trends, open challenges, and applications,” TISMIR, 2020.

Music is rich of structures

– Transformations of motifs: transposition, inversion, retrograde (reverse), etc.

Music is rich of structures

- Stream structures: “grouping of the notes of a polyphonic texture into melodic lines (also called streams, voices)”

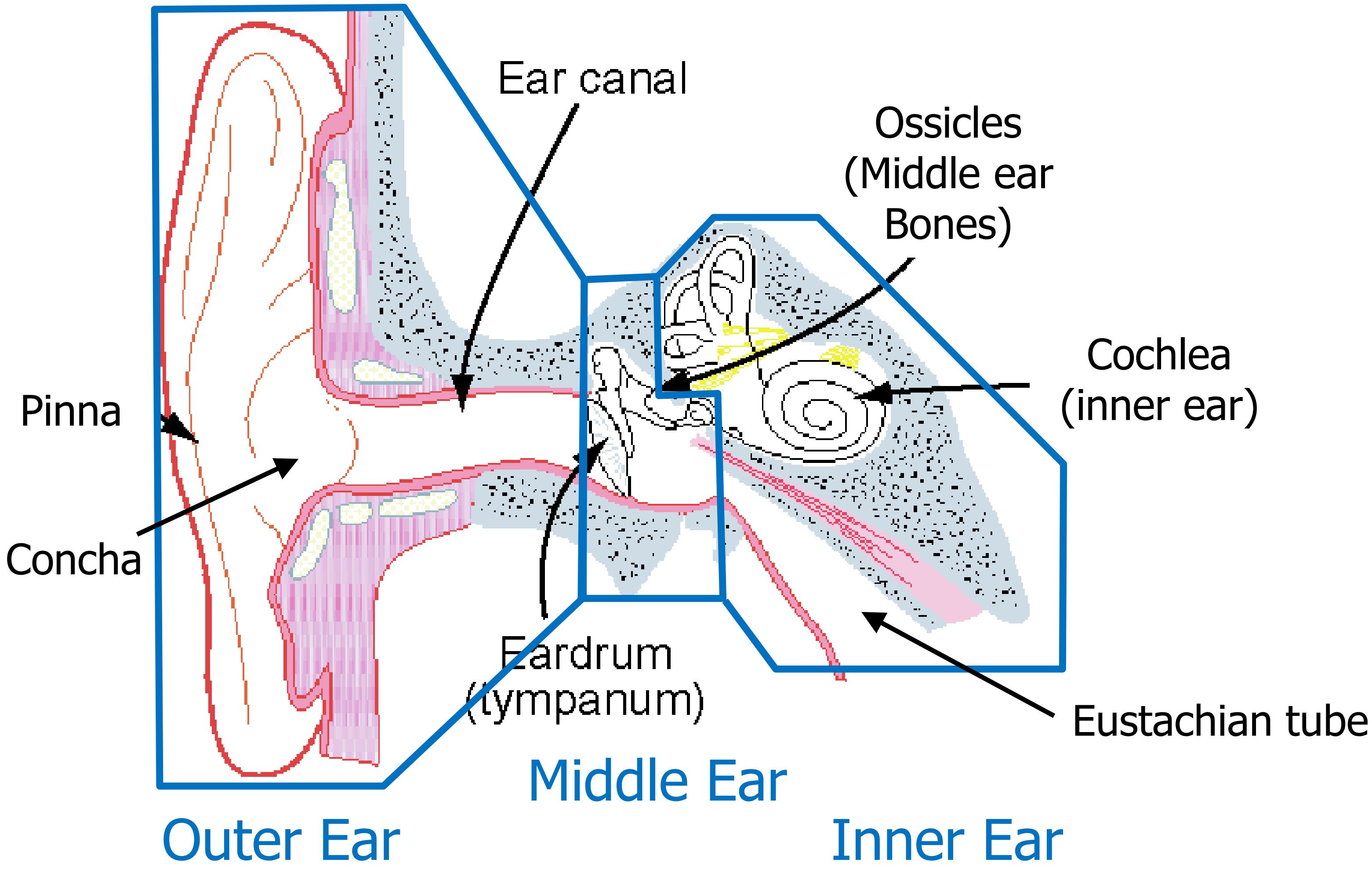
Minuet in G from *Notebook for Anna Magdalena Bach*

David Temperley, “A Unified Probabilistic Model of Polyphonic Music Analysis,” Journal of New Music Research vol. 38, pp. 3-18, 2009.

Outline

- MIR overview
- Auditory sensation
- Psychoacoustic inspirations
- Music audio features

The Ear

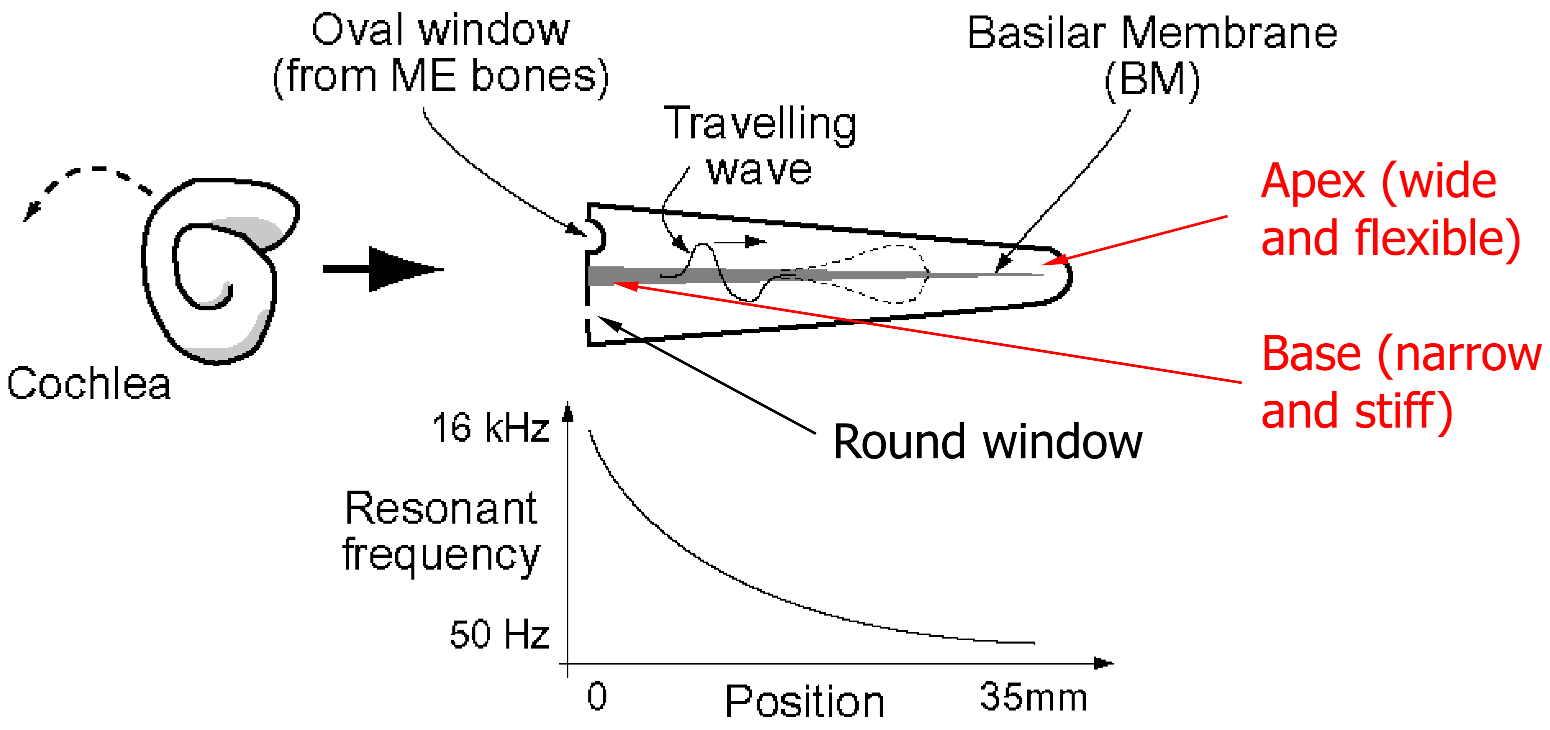


Ear canal
Ossicles
(Middle ear
Bones)
Cochlea
(inner ear)
Pinna
Concha
Eardrum
(tympanum)
Eustachian tube
Middle Ear
Outer Ear
Inner Ear

Function of the Ear

- Outer ear: shapes the sound spectrum
  - Torso, head, pinnae: head-related transfer function (HRTF). Interaural difference.
  - Concha, canal: increase sound level of about 10-15dB between 1.5k-7kHz, due to resonances
- Middle ear: effective and efficient transfer
  - Eardrum: effective area about 55 mm^2 (where the oval window is about 3 mm^2 size.
  - Three ossicles: a lever system
  - The last ossicle is called stapes, the smallest bone in the human body

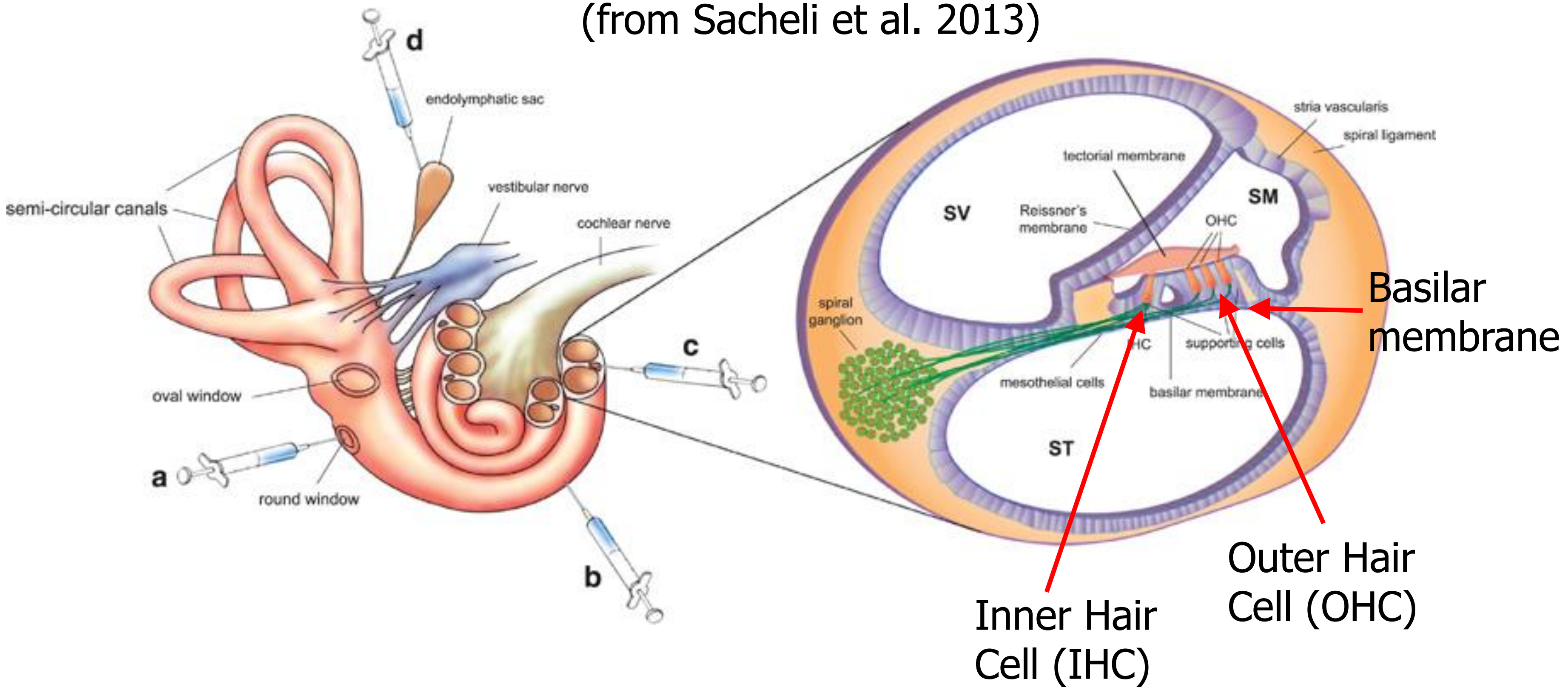
The Cochlea



- Each point on the Basilar membrane resonates to a particular frequency
- At the resonance point, the membrane moves

Cross Section of Cochlea

(from Sacheli et al. 2013)



- When the membrane moves, it moves hairs.
- When hairs move, they fire nerve impulses.

A Movie!

(thanks to Howard Hughes Medical Institute)

Hair Cells

- Inner hair cell: the actual transducer

- Outer hair cell: feedforward amplifiers, making small but very fast movements

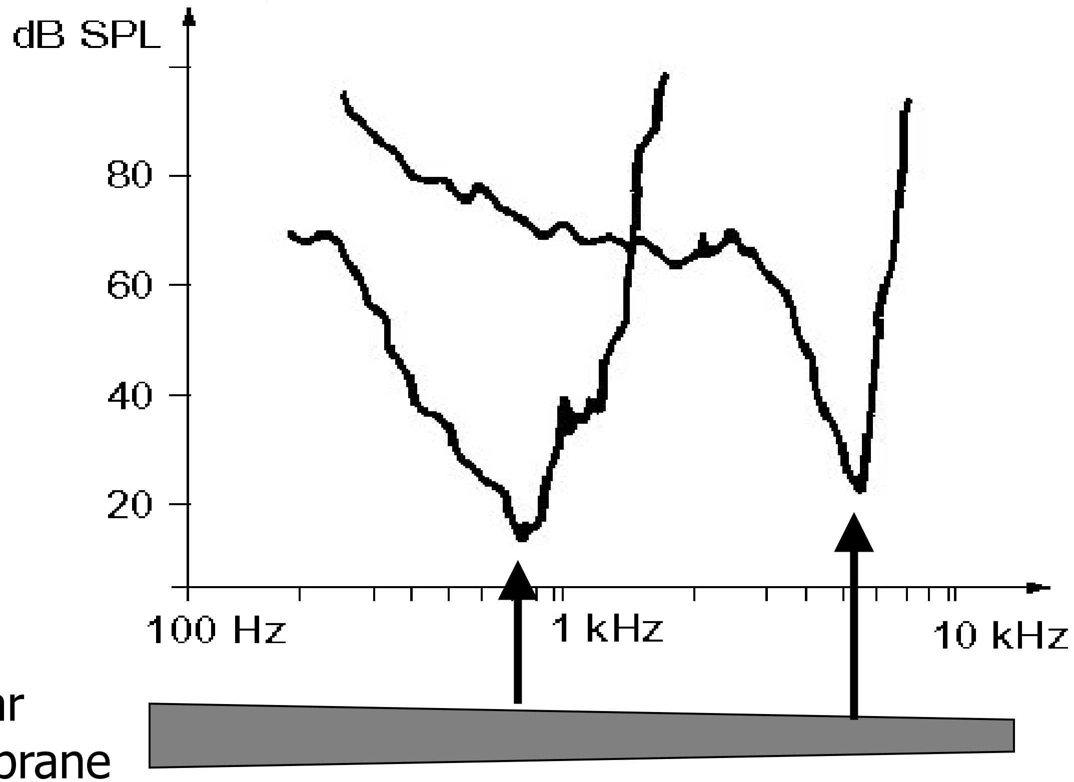
- They are damaged by age and hard to regrow

Dance of an outer hair cell of pig

https://www.youtube.com/watch?v=pij8a8aNpWQ

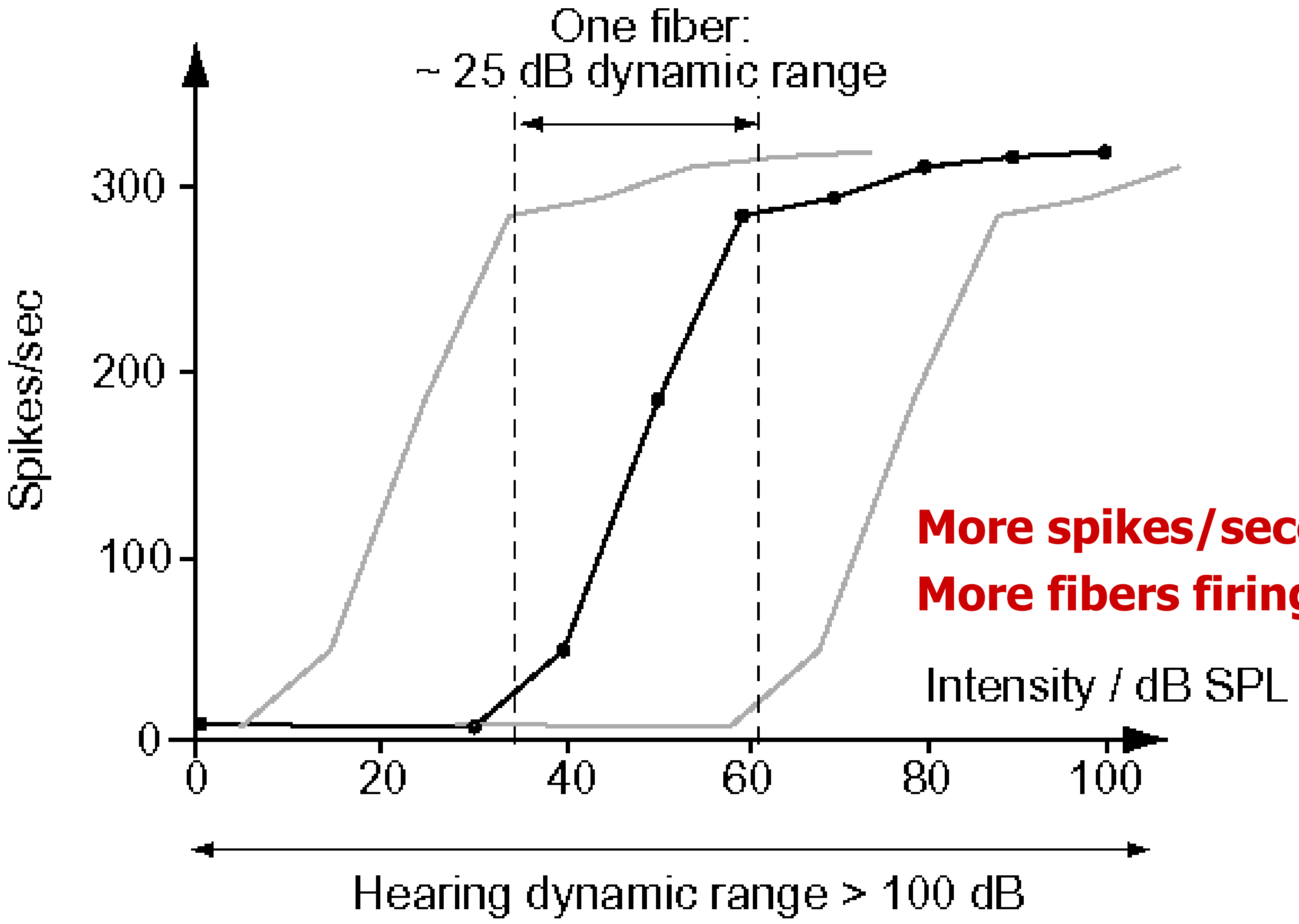
Frequency Sensitivity

- single nerve measurements
- (roughly) symmetric in log of frequency



Basilar membrane

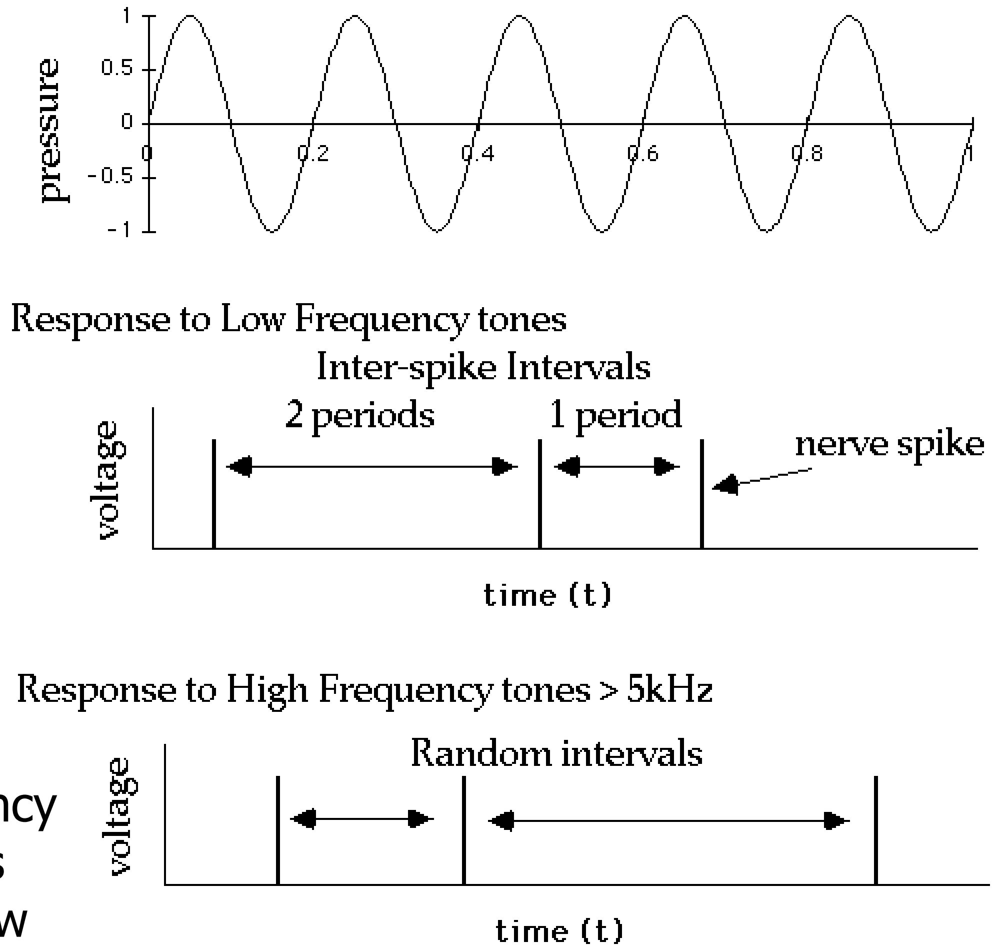
Encoding Loudness



More spikes/second = louder

More fibers firing = louder

Phase Locking



Half-wave rectification

For high frequency tones, the fibers phase lock to low frequency modulations.

(from Chris Darwin)

Measuring Signal Strength

- Acoustical

Average intensity

$$I = \frac{1}{\rho c}\frac{1}{T_D}\int_0^{T_D} x^2(t)dt$$

View $x(t)$ as sound pressure

density

sound speed

- Electrical

Average power

$$P = \frac{1}{R}\frac{1}{T_D}\int_0^{T_D} x^2(t)dt$$

View $x(t)$ as electric voltage

resistance

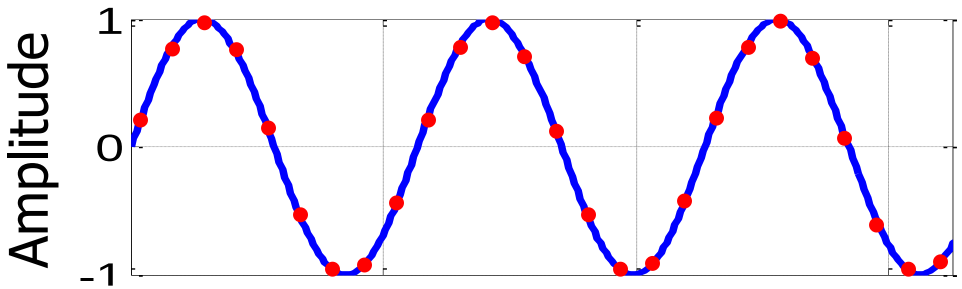
Root-Mean-Square (RMS)

$$x_{RMS} = \sqrt{\frac{1}{T_D}\int_0^{T_D} x^2(t)dt}$$

- T_D should be long enough.
- $x(t)$ should have 0 mean, otherwise the DC component will be integrated.
- For sinusoids

$$x_{RMS} = \sqrt{\frac{1}{T}\int_0^{T} A^2 \sin^2(2\pi f t)dt} = \sqrt{A^2/2} = 0.707A$$

Root-Mean-Square (RMS)



The red dots form the discrete signal $x[n]$

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]}$$

The Decibel Scale

- Softest audible sound intensity 0.000000000001 watt/m^2
- Threshold of pain is around 10 watt/m^2
- 13 orders of magnitude difference
- A log scale helps with this
- The decibel (dB) scale is a log scale, with respect to a reference value

$$L = 10\log_{10}\left(\frac{I}{I_{\mathrm{ref}}}\right)$$
$$= 20\log_{10}\left(\frac{x_{RMS}}{x_{\mathrm{ref},RMS}}\right)$$

Lots of references!

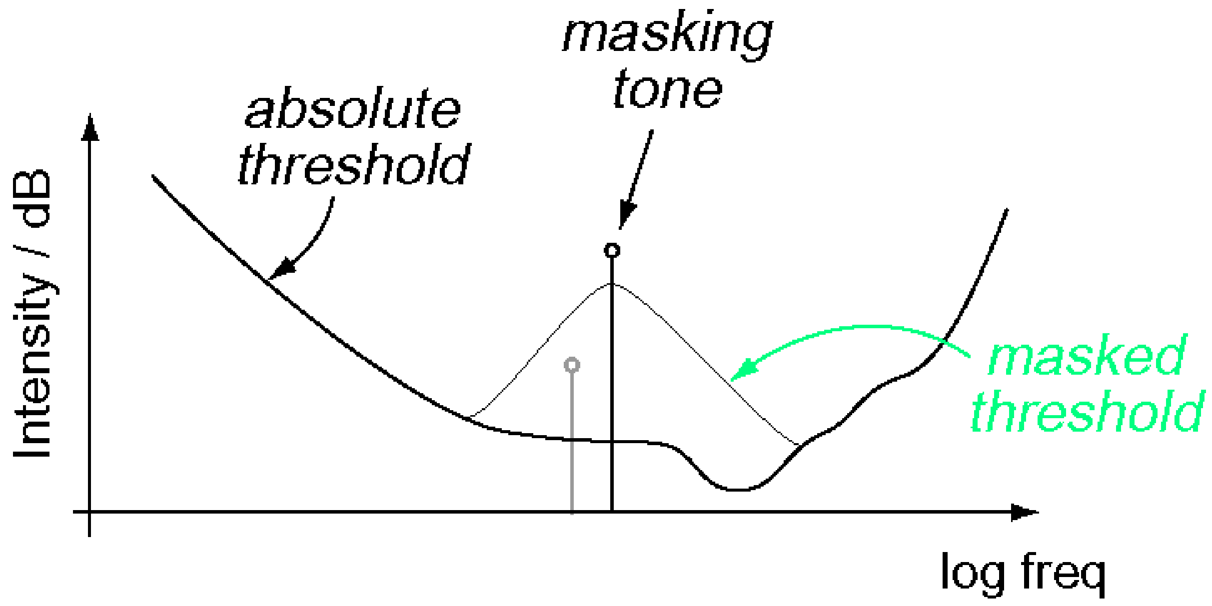
- **dB-SPL** – A measure of sound pressure level. 0dB-SPL is approximately the quietest sound a human can hear, roughly the sound of a mosquito flying 3 meters away.

- **dbFS** – relative to digital full-scale. 0 dbFS is the maximum allowable signal. Values are typically negative.
- **dBV** – relative to 1 Volt RMS. 0dBV = 1V.
- **dBu** – relative to 0.775 Volts RMS with an unloaded, open circuit.
- **dBmV** – relative to 1 millivolt across 75 Ω. Widely used in cable television networks.
- ……

Typical Values

• Jet engine at 3m	140 db-SPL
• Pain threshold	130 db-SPL
• Loud motorcycle, 5m	110 db-SPL
• Vacuum cleaner	80 db-SPL
• Quiet restaurant	50 db-SPL
• Rustling leaves	20 db-SPL
• Human breathing, 3m	10 db-SPL
• Hearing threshold	0 db-SPL

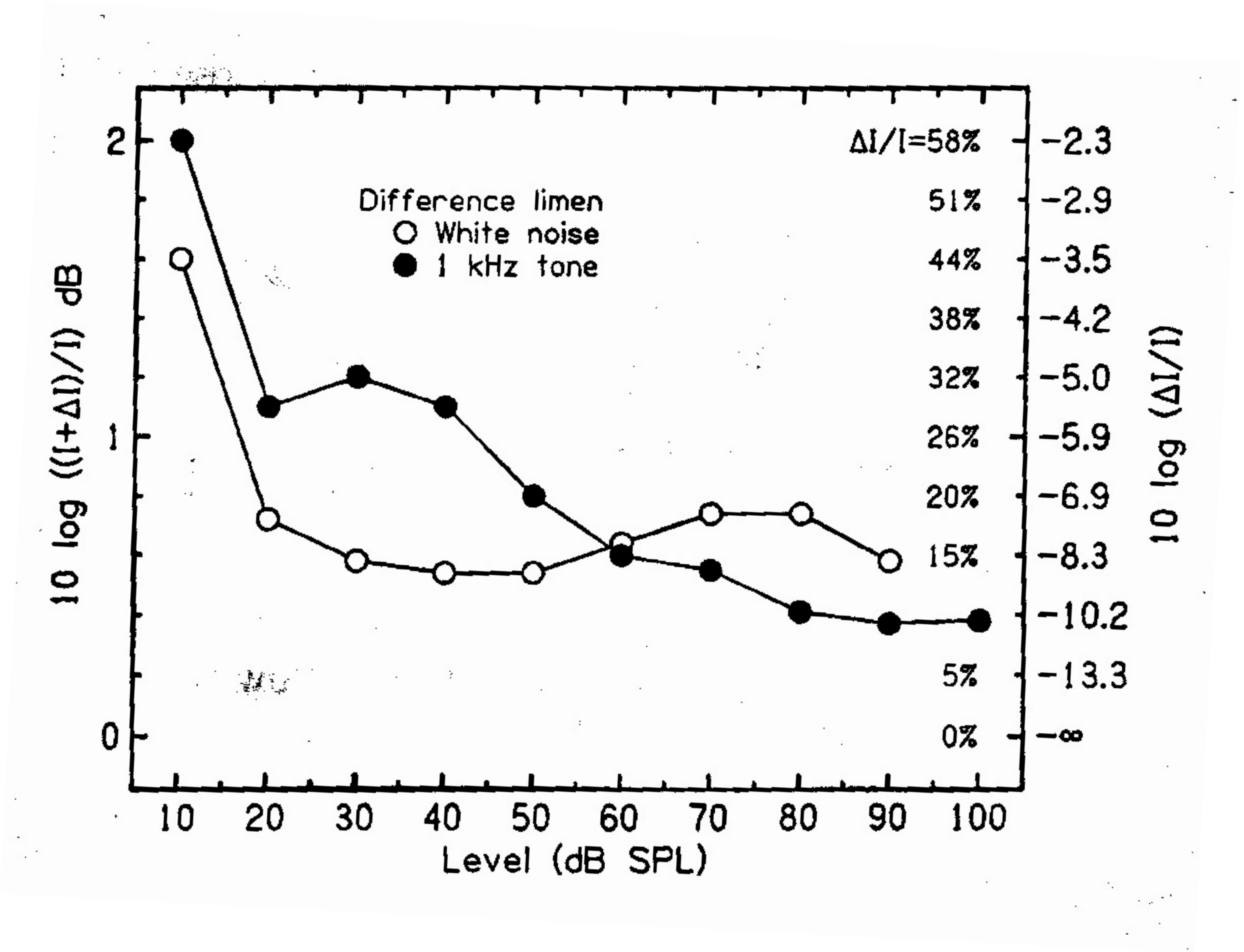
Masking



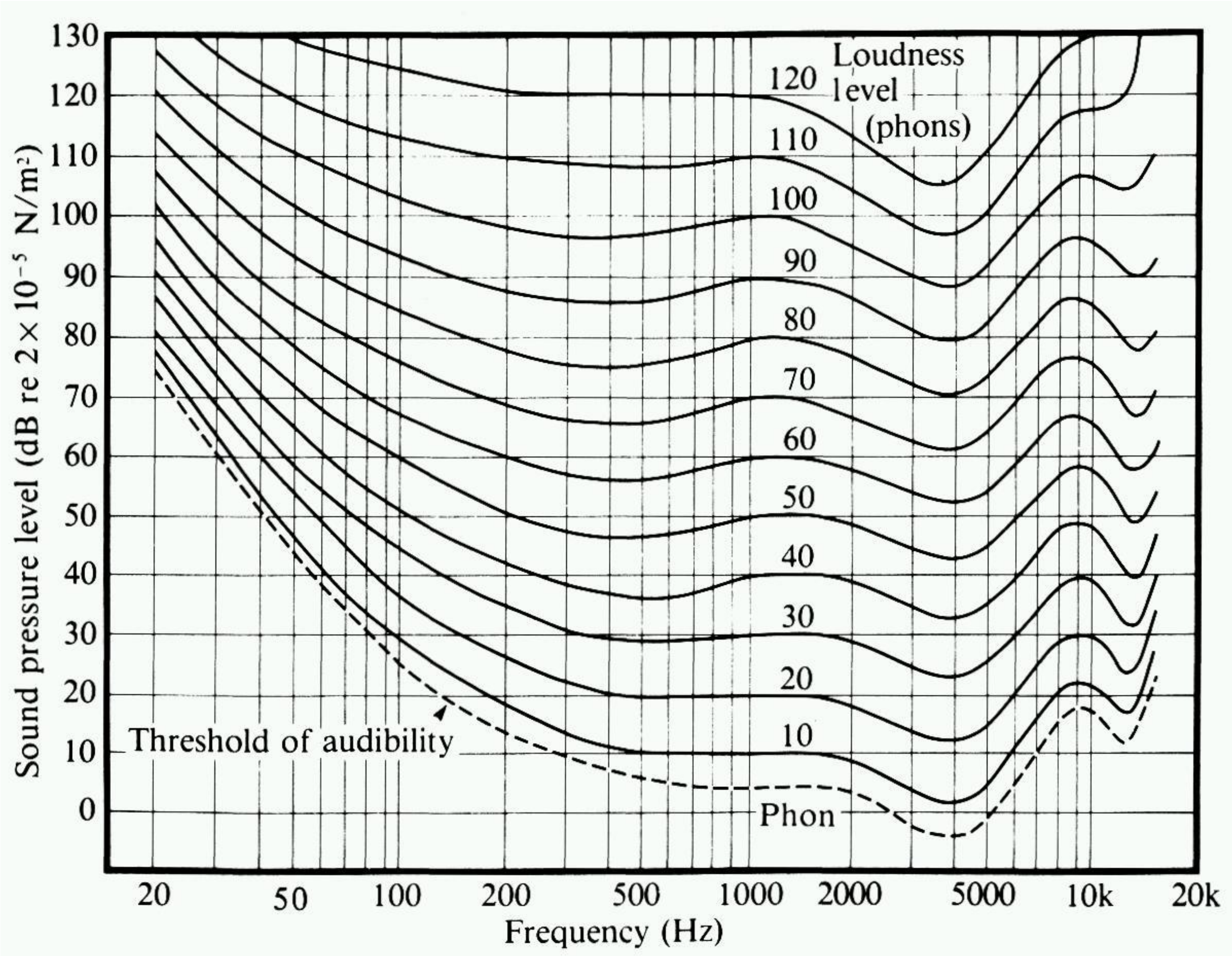
- A loud tone masks perception of tones at nearby frequencies

1000 Hz   1000_975_20dB   1000_975_6dB   1000_475_20dB

Loudness Difference Limens



Equal Loudness Contours (dB-SPL ⬄ Phon)

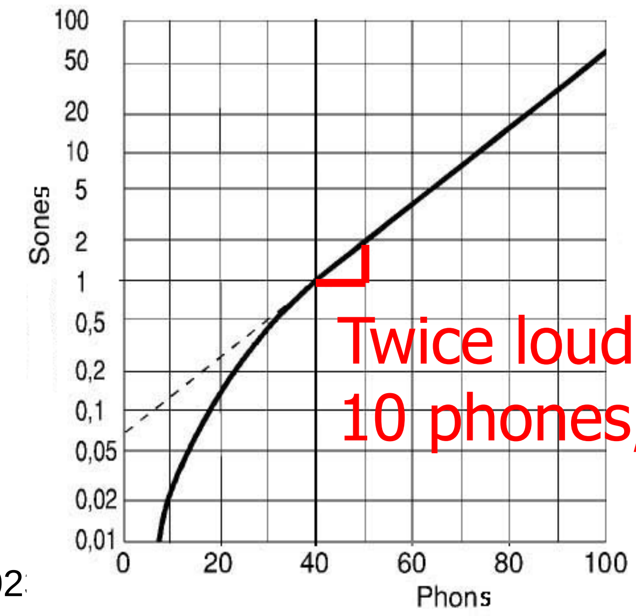


Phon ⇔ Sone

- The **phon** is a unit of perceived loudness to compensate for the effect of frequency on the perceived loudness of tones.
  - By definition, 1 phon is equal to 1 db-SPL at 1000 Hz.
- The **sone** is a unit of perceived loudness scale
  - At 1kHz, 1 sone = 40 phons = 40 dB-SPL
  - A stimulus that is *n* sones loud is judged to be *n* times as loud as 1 sone.

Relation between Phon and Sone

$$P_2 - P_1 = 10(\log_2 S_2 - \log_2 S_1)$$

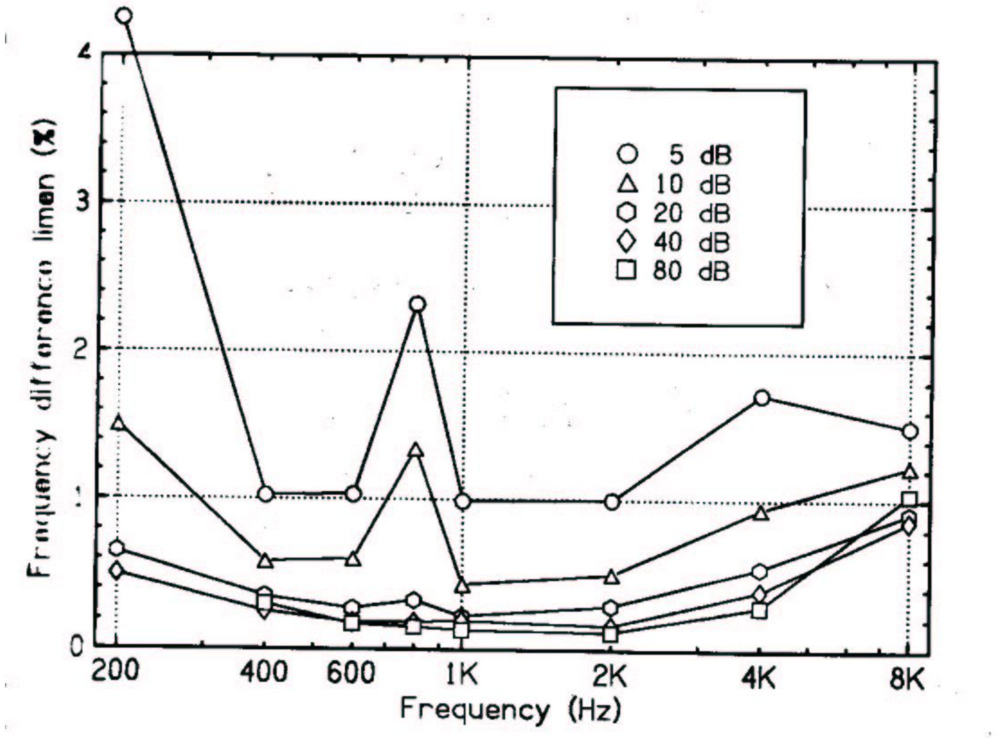


Intensity ⇔ Sone

- Tone at 1kHz with intensity > 40 dB SPL
- To make the tone *n* times as loud, how many times should we increase the intensity?
  - We want to have $\frac{S_{new}}{S} = n$.
  - Therefore, we need $P_{new} - P = 10 \log_2 n$.
  - That is, we need $10 \log_{10} \frac{I_{new}}{I} = 10 \log_2 n$.
  - So $\frac{I_{new}}{I} = 10^{\log_2 n} = 10^{\frac{\log_{10} n}{\log_{10} 2}} = n^{\log_2 10} \approx n^{3.32}$.
- This is why $\sqrt[3]{I}$ was used to describe perceived loudness

Frequency Difference Limen

- The smallest difference between the frequencies of two sine tones that can be discriminated correctly 75% of the time.



1000Hz

1001Hz

1002Hz

1005Hz

Pitch (ANSI 1994 Definition)

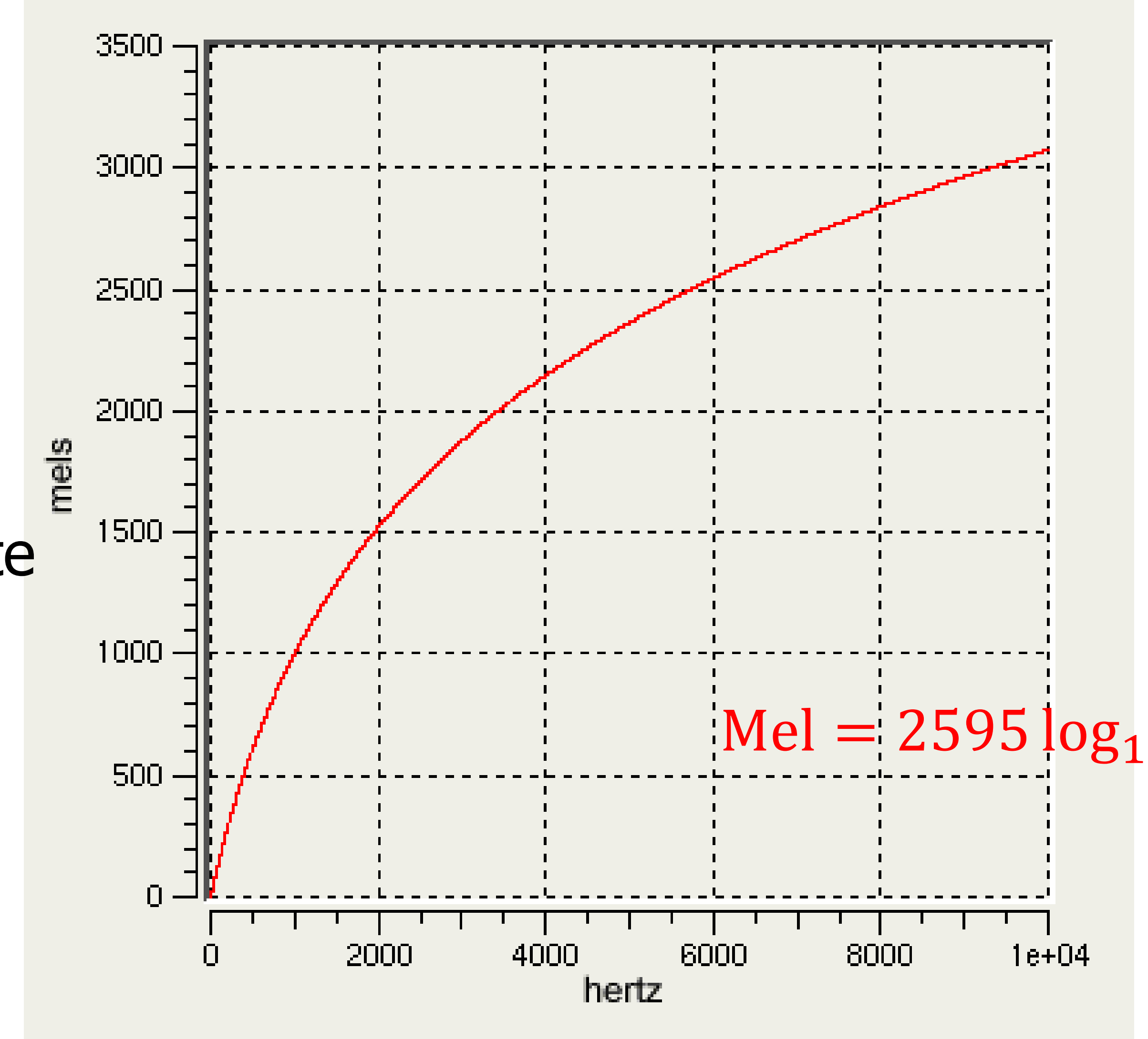
- That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch **depends mainly on the frequency** content of the sound stimulus, but **also depends on the sound pressure and waveform** of the stimulus

- (Operational) A sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at 40 dB SPL

Mel Scale

- Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments.

- The name **mel** comes from the word **melody** to indicate that the scale is based on pitch comparisons.



$$\text{Mel} = 2595\log_{10}(1 + \frac{f}{700})$$

Physical vs. Psychological

Frequency	Pitch	Low - high
Intensity	Loudness	Soft - loud
?	Timbre	Warm Bright Rough Violin-like ...

Timbre (tone quality, tone color)

"That attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

---- ANSI, 1960.

Oboe

Clarinet

- OK, but…, what is it?
- "The word timbre...is empty of scientific meaning, and should be expunged from the vocabulary of hearing science."

---- Keith Martin, PhD thesis, 2000.

Timbre and Physics

- “Quality of tone [timbre] should depend on the manner in which the motion is performed within the period of each single vibration”

---- Helmholtz, 1877.

- “Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus.”

---- ANSI, 1960.

Examples

- Spectral energy distribution
  - The clarinet and oboe example
- Attack (onset)

Without attack

With attack

- Temporal evolution

Time reverse

Timbre Definition Revisit

"That attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

---- ANSI, 1960.

- Does not mention the role timber plays in cases where pitch and/or loudness are different.
  - Two notes played by the same instrument have similar timbre, even if they have different pitch and/or loudness.

Timbre Representation

- Can be computed from the signal
- Can discriminate different sound sources (e.g., different musical instruments, different talkers)
- Approximately invariant to pitch/loudness changes for the same source

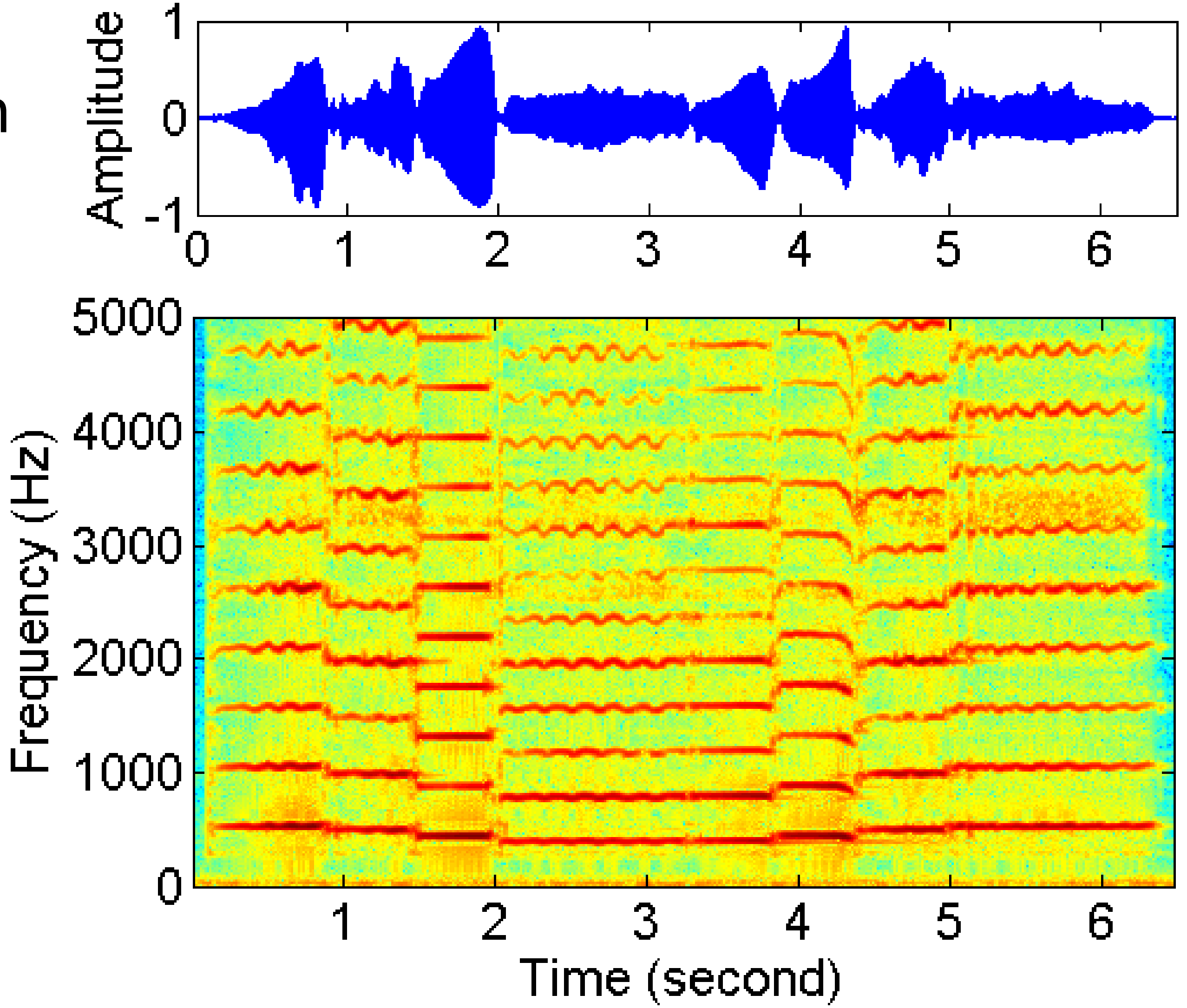
- Most audio features are related to timbre (e.g., spectral statistics, MFCC, wav2vec), but there is not a single feature that can completely characterize timbre

Outline

- MIR overview
- Auditory sensation
- Psychoacoustic inspirations
- Music audio features

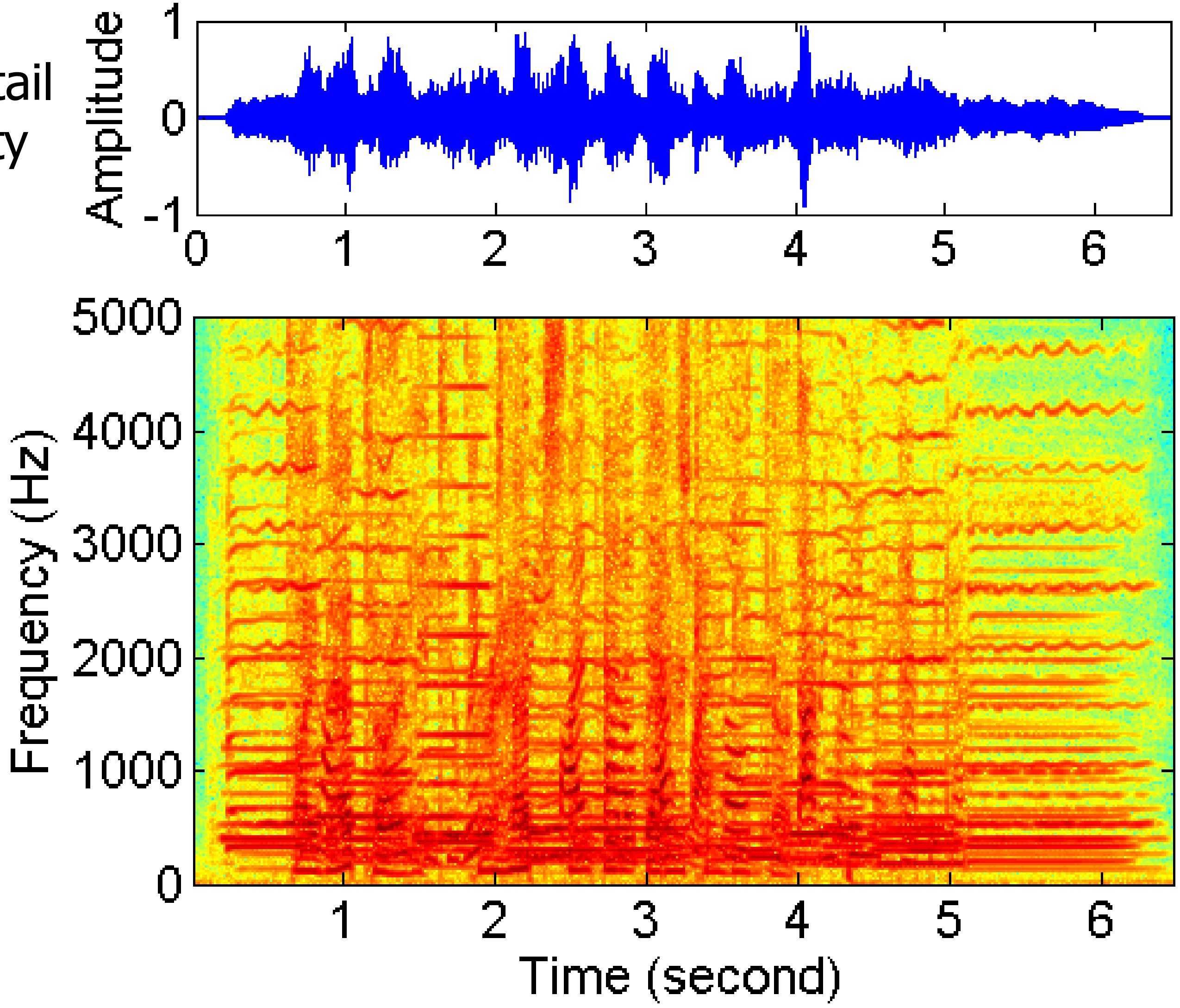
Spectrogram

violin

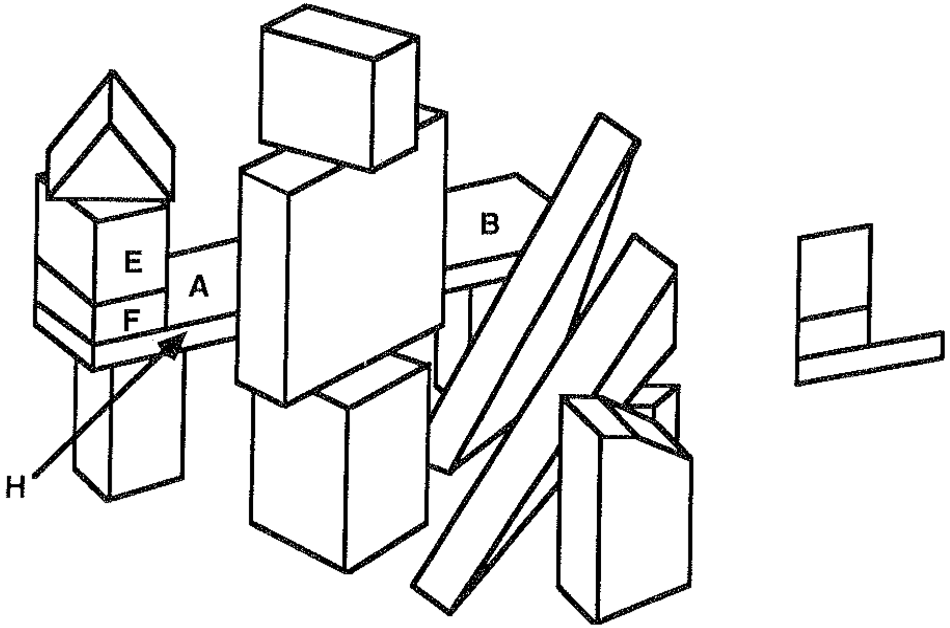


How about this?

cocktail party



Auditory Scene Analysis

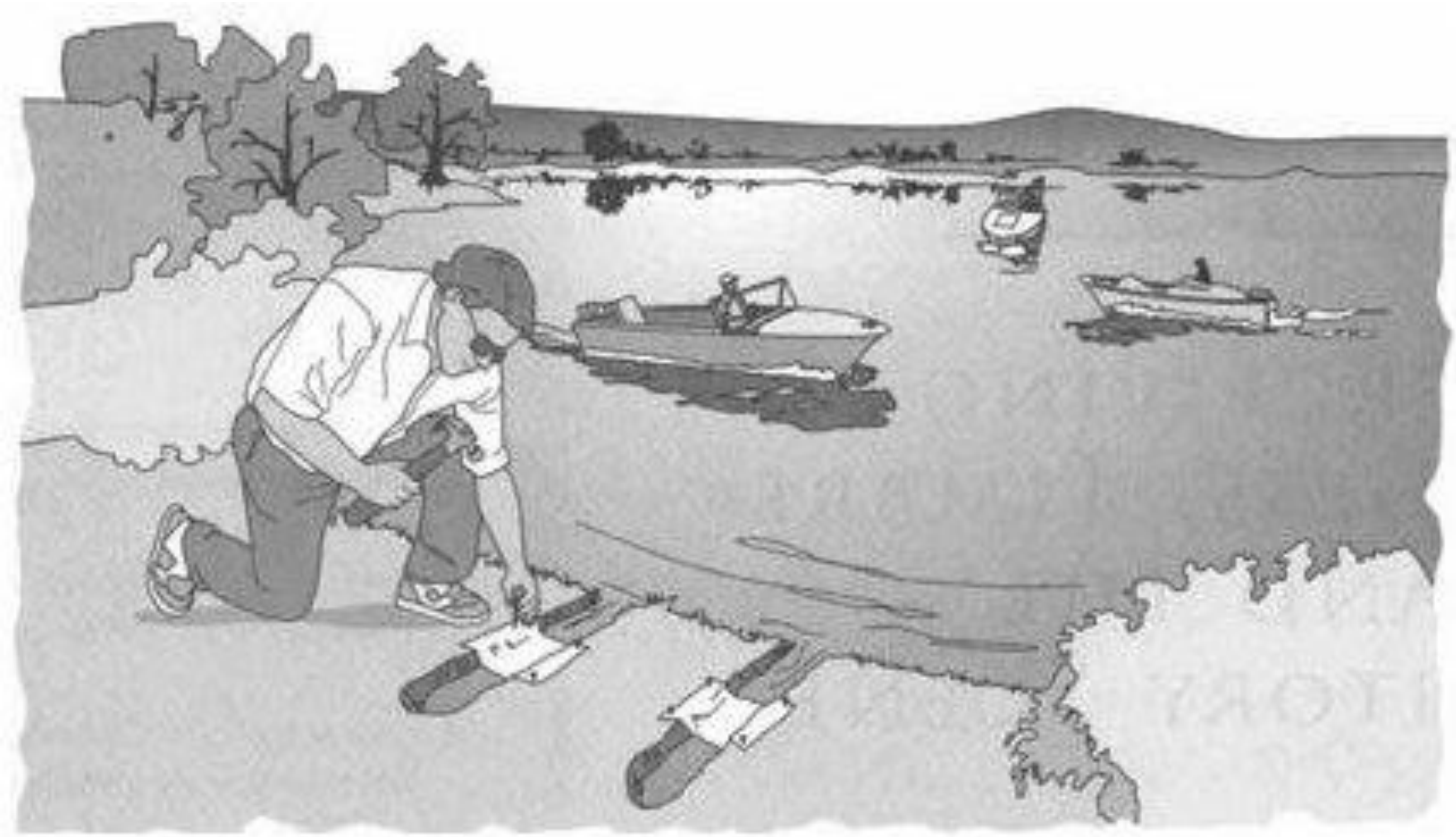


(from Bregman’s ASA book, Figure 1.2)

The cocktail party problem

(From http://www.justellus.com/)

It's very difficult!

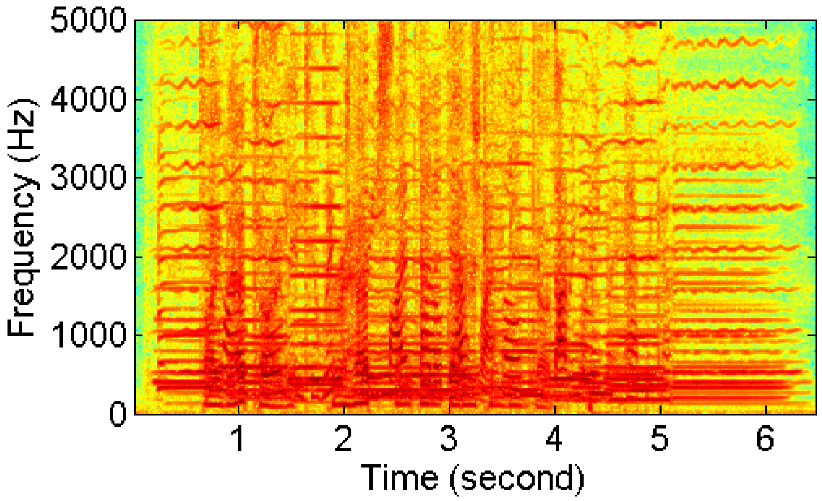


Auditory Scene Analysis

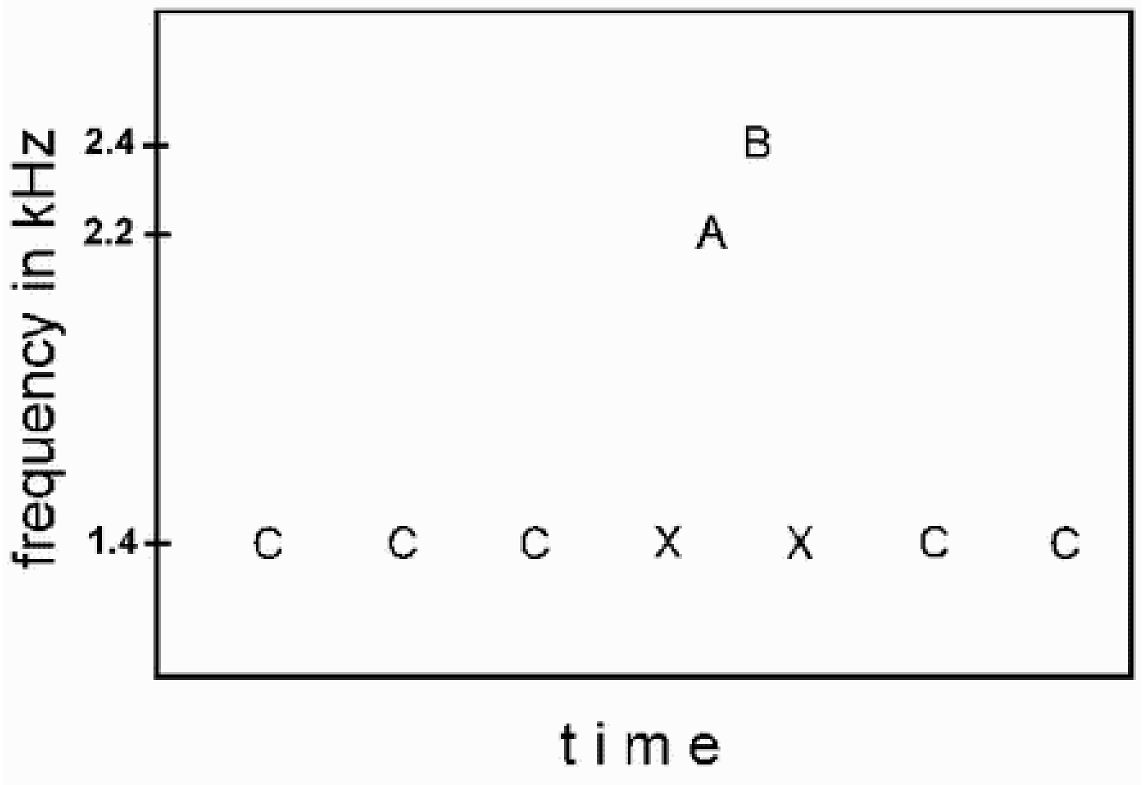
- Studies how human auditory systems solve problems like
  - How many sources at a time?
  - Which frequency components belong to the same source?
  - How does a source evolve?
  - Where are the sources?

The Analysis-Synthesis Process

- Decompose the acoustic scene into a collection of segments
- Group segments into streams
  - Simultaneous vs. sequential
  - This is the main concern of ASA



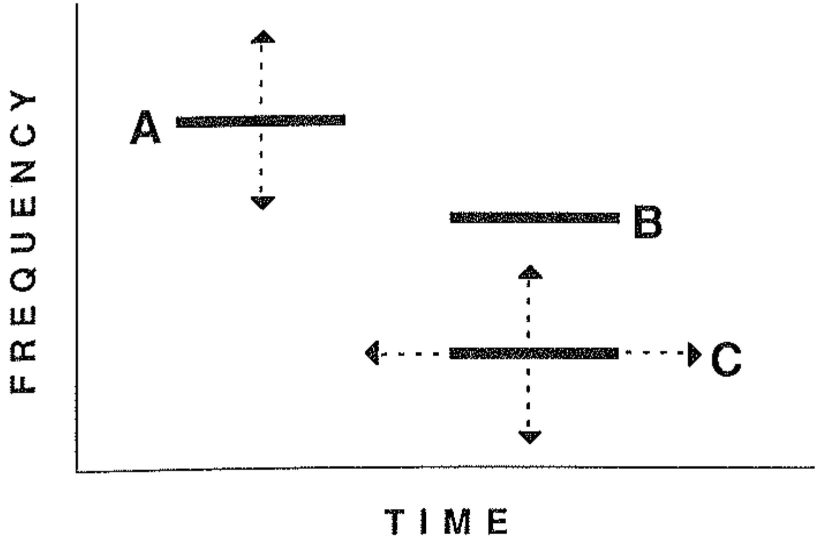
Exclusive Allocation



Audio downloaded from https://themusiclab.github.io/bregman-archive/downloadsdl.htm

- The allocation of the X tones are different when the C tones are played or not, and it affects our perception of the A and B tones.

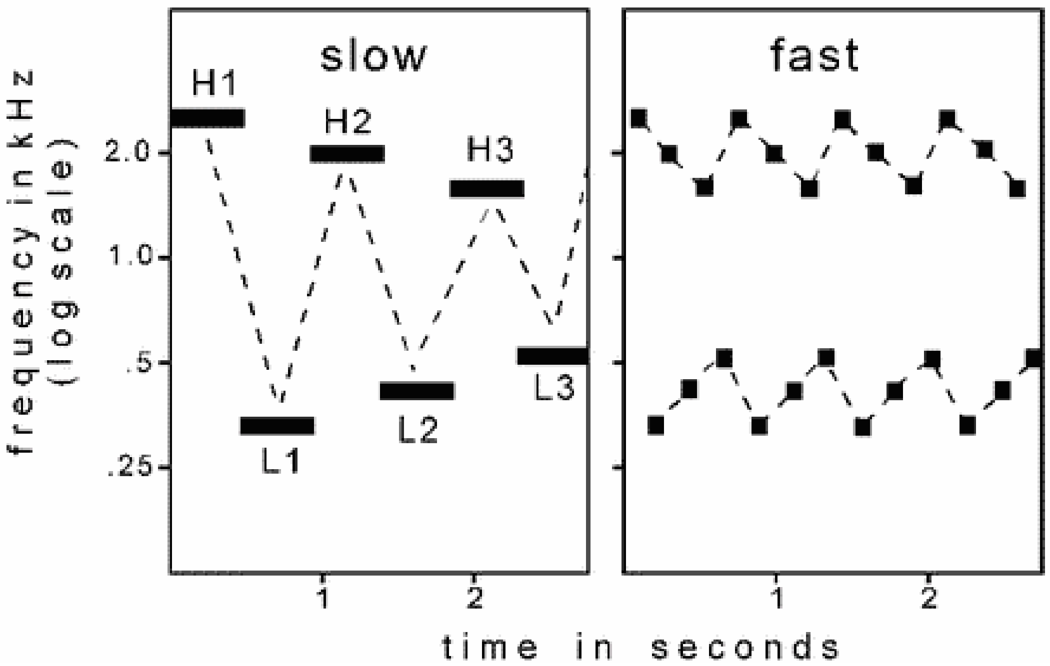
Simultaneous vs. Sequential



Audio downloaded from https://themusiclab.github.io/bregman-archive/downloadsdl.htm

- Things that affect the grouping of ABC tones
  - Frequency difference between A and B
  - Frequency difference between B and C
  - Synchronization between B and C

Stream Segregation



Audio downloaded from https://themusiclab.github.io/bregman-archive/downloadsdl.htm

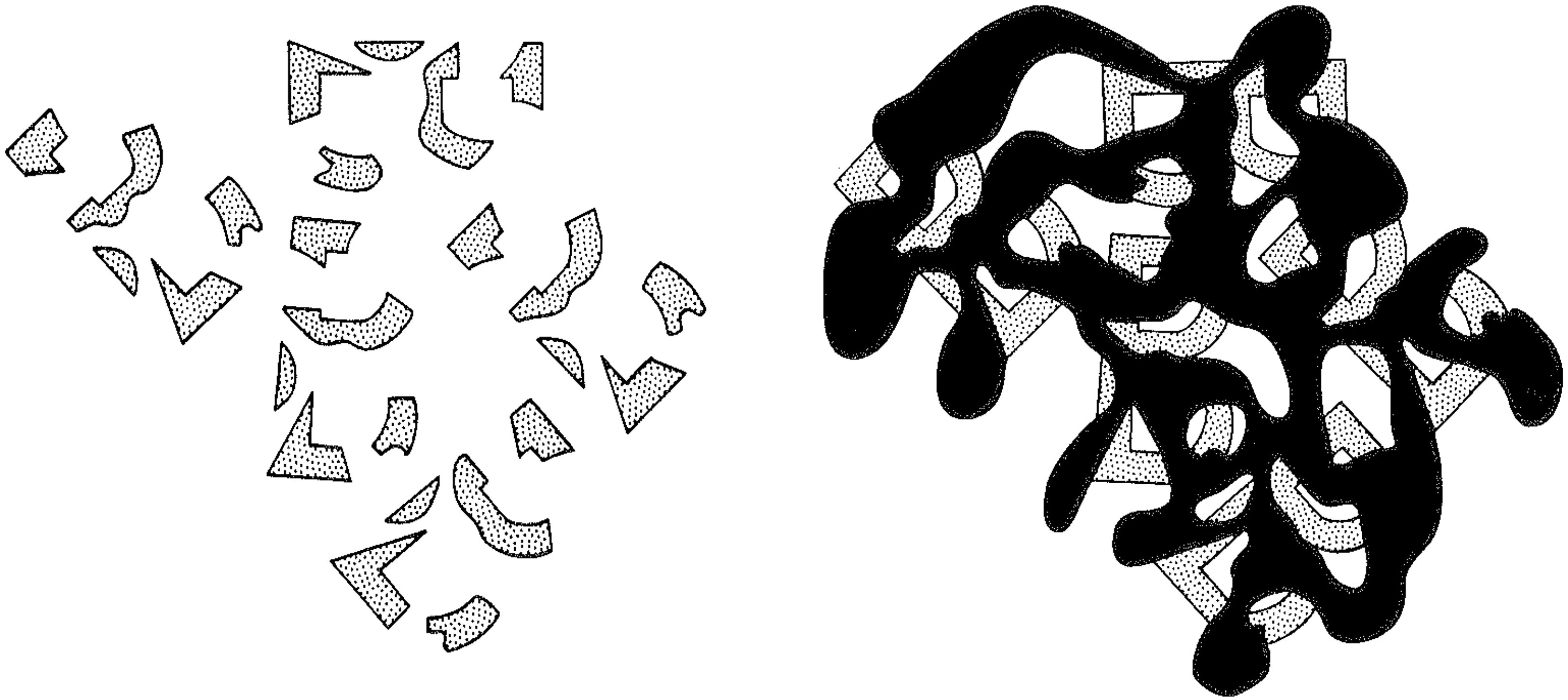
- High and low tones are segregated when played fast
- Can you tell the order of the tones?

Stream Segregation in Music

Toccata and Fugue in D minor, J.S. Bach

https://www.youtube.com/watch?v=R_tu63ypB6I

Occlusions in Vision

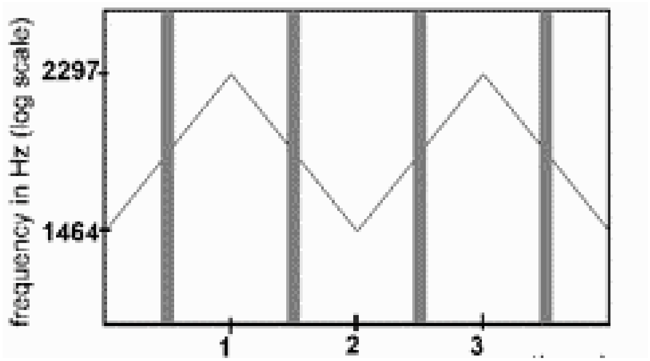


- The occlusion in this example helps with the grouping of the fragments

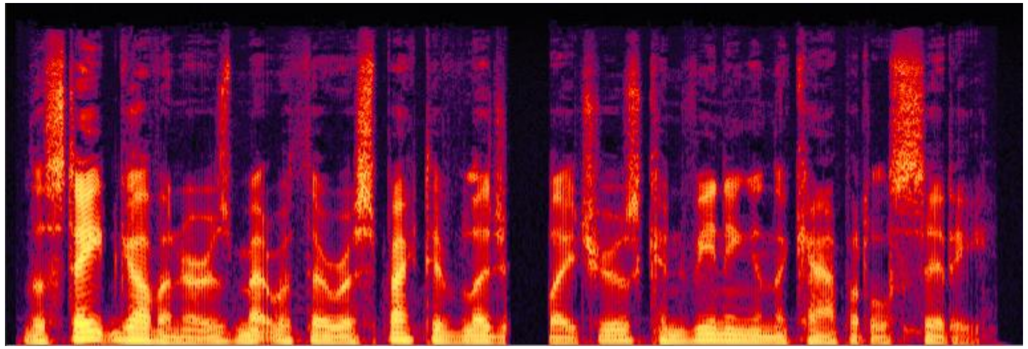
(from Bregman’s ASA book)

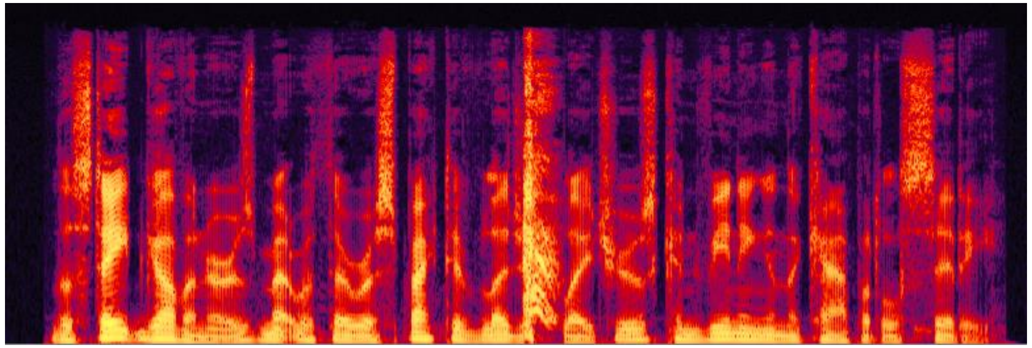
Masking in Audition

Sinusoids



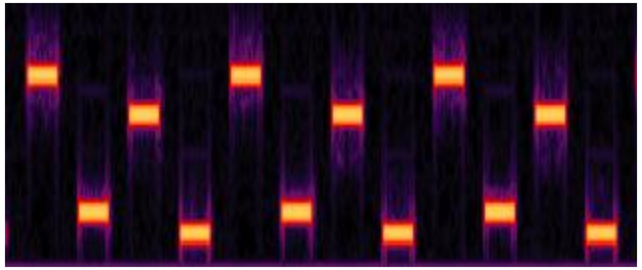
Speech



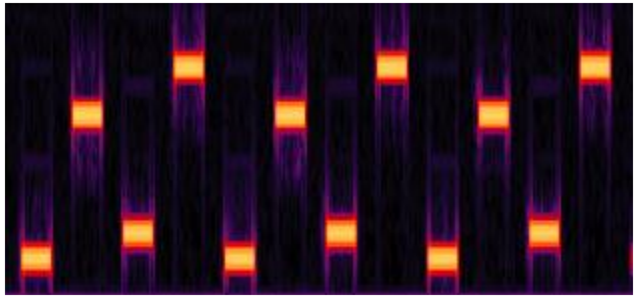


Primitive vs. Learned

H1-L1-H2-L2



L2-H2-L1-H1



- Infants cannot discriminate the two stimuli, which indicates that they perform stream segregation of the high and low tones.

Audio downloaded from https://themusiclab.github.io/bregman-archive/downloadsdl.htm

Primitive Grouping Mechanisms

- For simultaneous grouping
  - Periodicity
  - Common onset and offset
  - Common amplitude and frequency modulation
- For sequential grouping
  - Proximity in frequency and time
  - Continuous or smooth transition
  - Related rhythm
- Common spatial location

Learning Improves ASA Performance

- Repeated listening to the stimulus can improve performance in ASA tasks
- Easier to follow a friend's than a stranger's voice in a noisy environment
  - Prior knowledge of timbre helps
- Music training helps analyzing music audio scene
  - Prior knowledge of music theory, composition rules, music style, etc. helps

Extreme Capability in Music ASA

- "In Rome, he (14 years old) heard Gregorio Allegri's *Miserere* **once** in performance in the Sistine Chapel. He wrote it out **entirely from memory**, only returning to correct **minor errors**…"

  -- Gutman, Robert (2000). *Mozart: A Cultural Biography*

Wolfgang Amadeus Mozart

- "MIR grant challenge": can algorithms compete against Mozart?

Outline

- MIR overview
- Auditory sensation
- Psychoacoustic inspirations
- Music audio features

Time-Domain Features

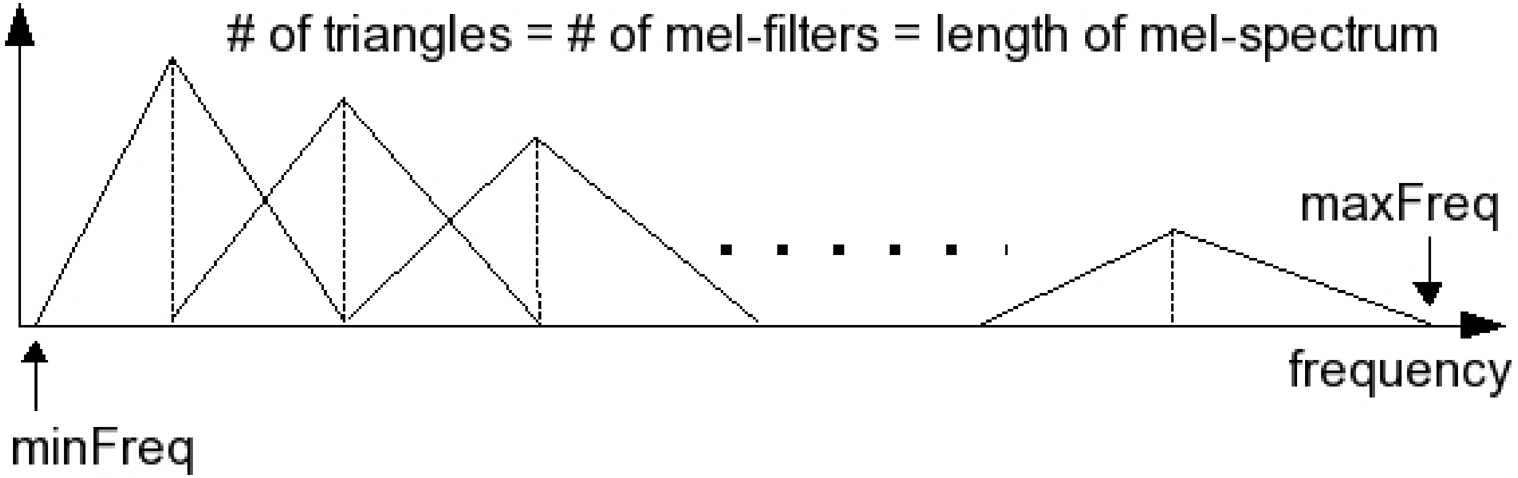
- RMS
  - Used to discriminate silence/non-silence

- Zero crossing rate (ZCR)
  - How often the time-domain signal changes its sign
  - Describes the amount of high-frequency energy
  - Correlates strongly with spectral centroid
  - Quite discriminative for percussion instruments

$$ZCR(n) = \frac{1}{2N}\sum_{i=1}^{N}|\mathrm{sign}(x[n+i]) - \mathrm{sign}(x[n+i-1])|$$

Spectral Features

- Can be calculated from either the linear frequency magnitude spectrum, or the mel-scale filter bank responses



- From now on, let $X[k]$ be either a linear frequency scale magnitude spectrum or a mel-scale filter bank response.

Spectral Features

- Spectral centroid

$$C_f = \frac{\sum_k kX[k]}{\sum_k X[k]}$$

- Spectral spread

$$S_f^2 = \frac{\sum_k \left(k - C_f\right)^2 X[k]}{\sum_k X[k]}$$

Spectral Features

- Spectral skewness
  - How asymmetric of the frequency distribution around the spectral centroid

$$\gamma_1 = \frac{\sum_k \left(k - C_f\right)^3 X[k]}{S_f^3 \sum_k X[k]}$$

- Spectral kurtosis
  - The peakiness of the frequency distribution

$$\gamma_2 = \frac{\sum_k \left(k - C_f\right)^4 X[k]}{S_f^4 \sum_k X[k]}$$

Spectral Features

- Spectral flatness
  - How flat (i.e., “white-noisy”) the spectrum is

$$SFM = 10\log_{10}\left(\frac{\left(\prod_{k=1}^{K} X[k]\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} X[k]}\right)$$

- Spectral irregularity
  - The jaggedness of the spectrum

$$SI = \frac{\sum_k (X[k] - X[k+1])^2}{\sum_k X[k]^2}$$

Spectral Features

- Spectral roll-off
  - The frequency index R below which a certain fraction γ of the spectral energy resides

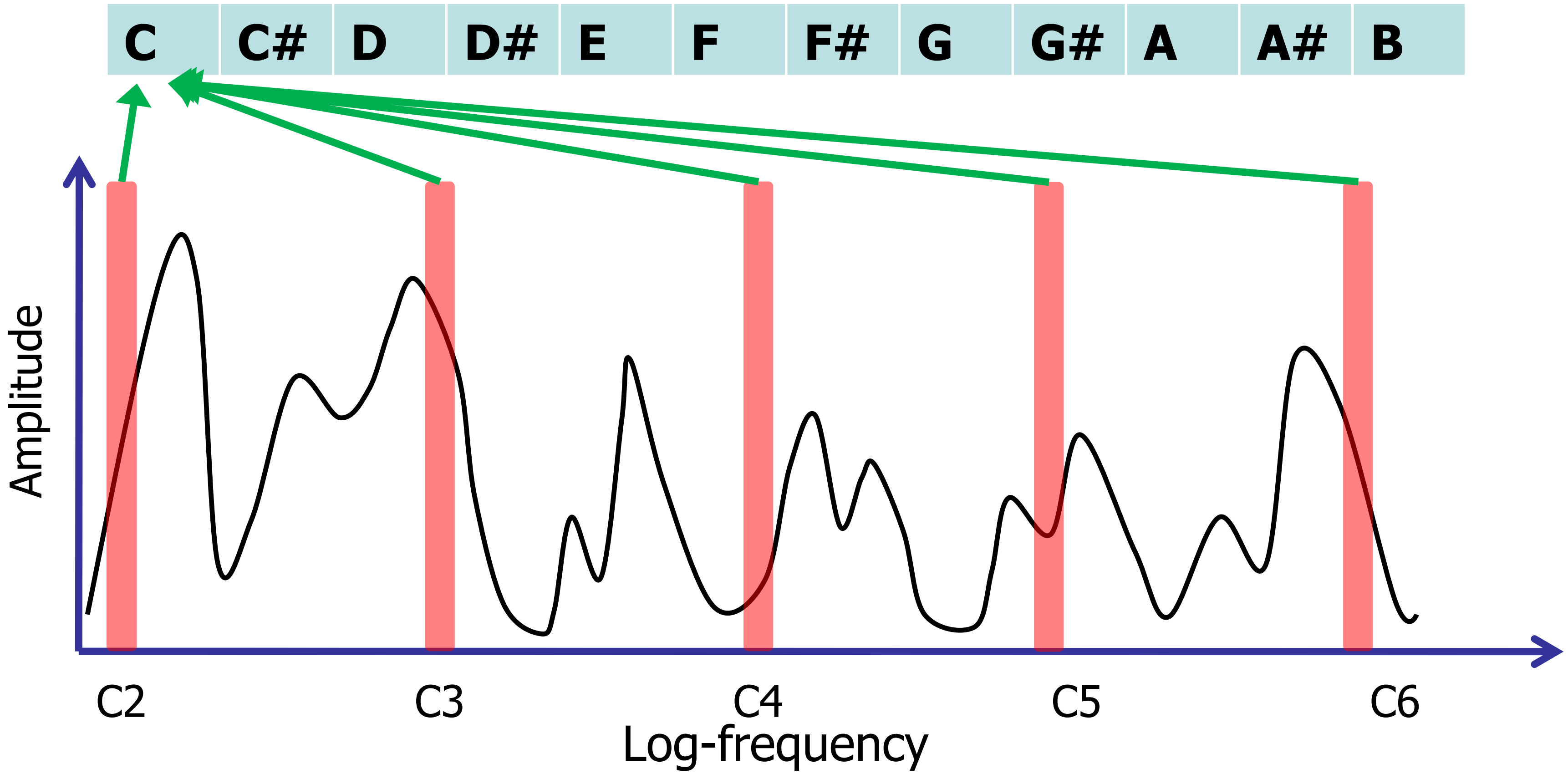
$$\sum_{k=1}^{R} X[k]^2 \geq \gamma \sum_{k} X[k]^2$$

- Spectral flux (delta spectrum magnitude)
  - Measure of local spectral change

$$SFX(t) = \sum_{k} \left(\frac{X_t[k]}{\sum_k X_t[k]} - \frac{X_{t-1}[k]}{\sum_k X_{t-1}[k]} \right)^2$$

Chroma Feature

- Spectral energy of the 12 pitch classes
  - 12-d vector



Spectrogram

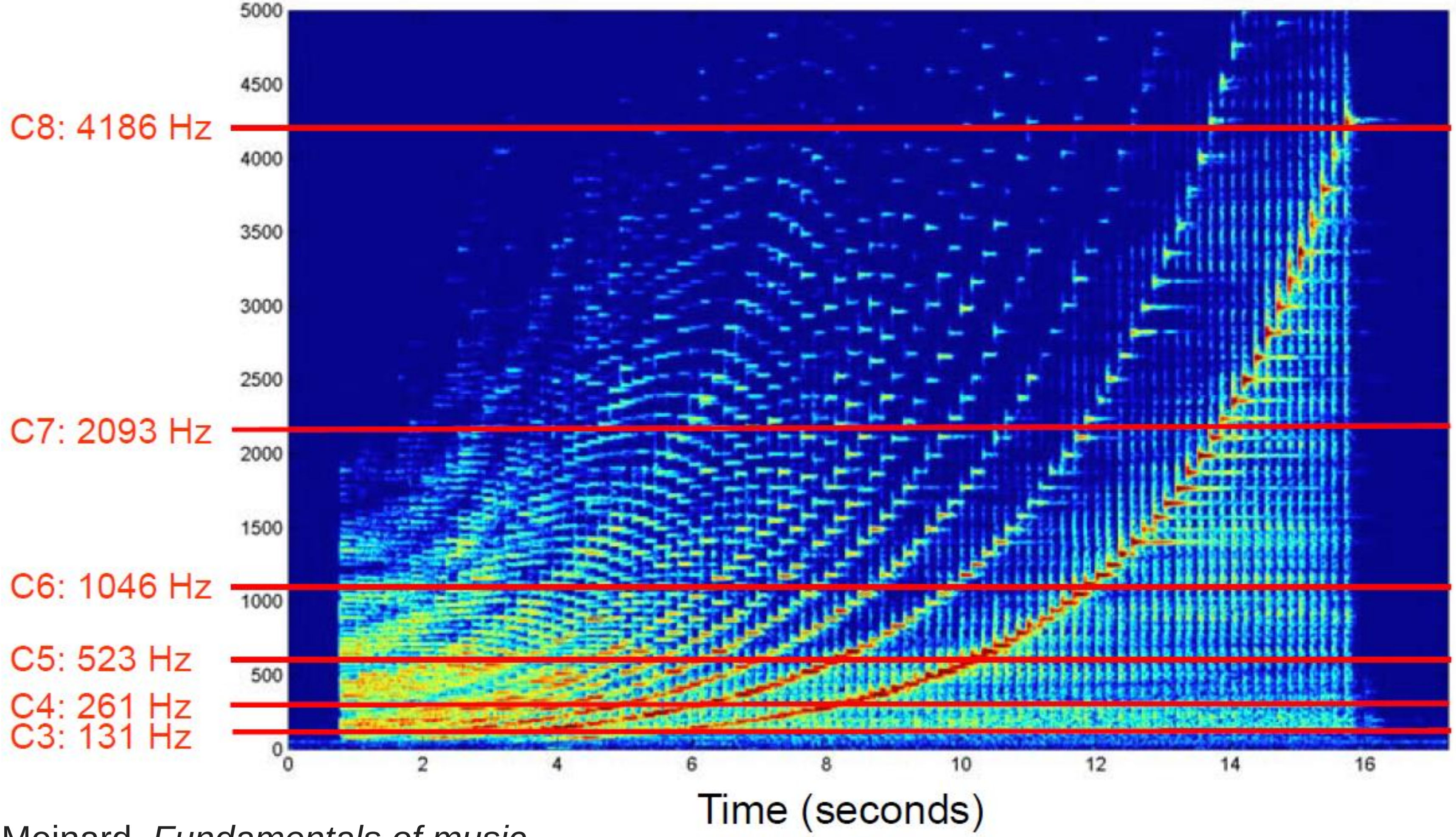


Figure from Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Log-frequency Spectrogram

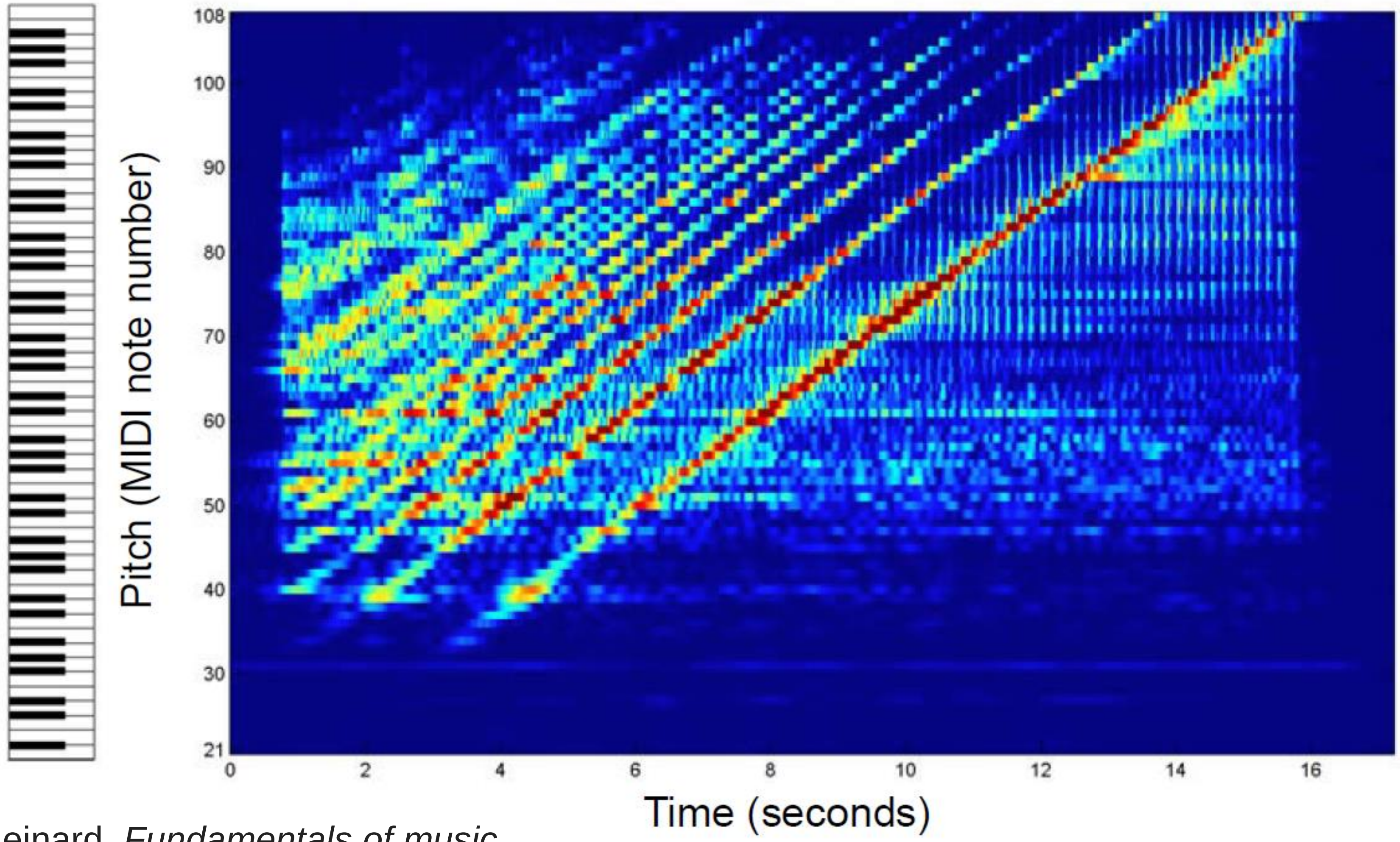


Figure from Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Chromagram

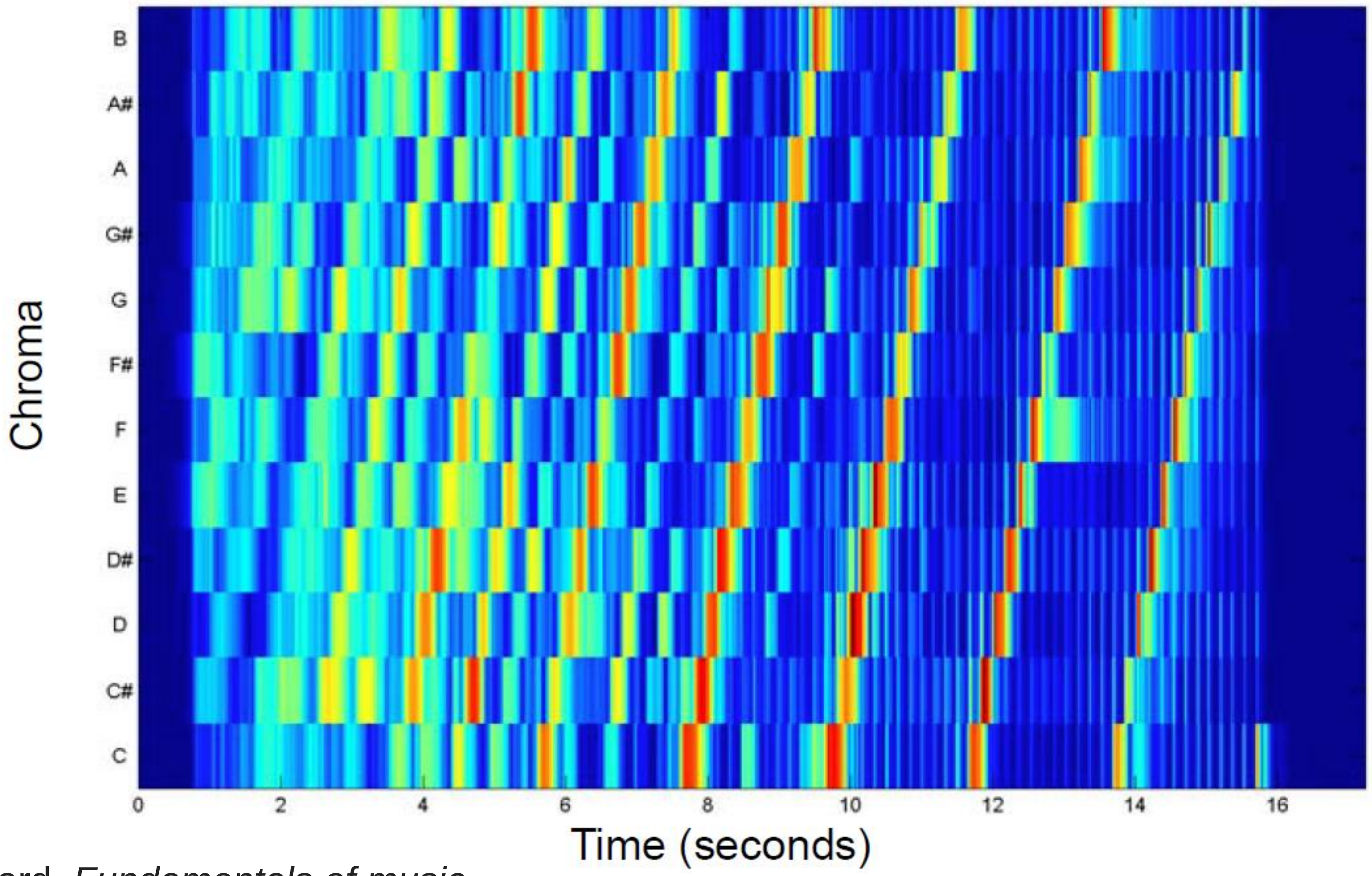


Figure from Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Normalized Chromagram

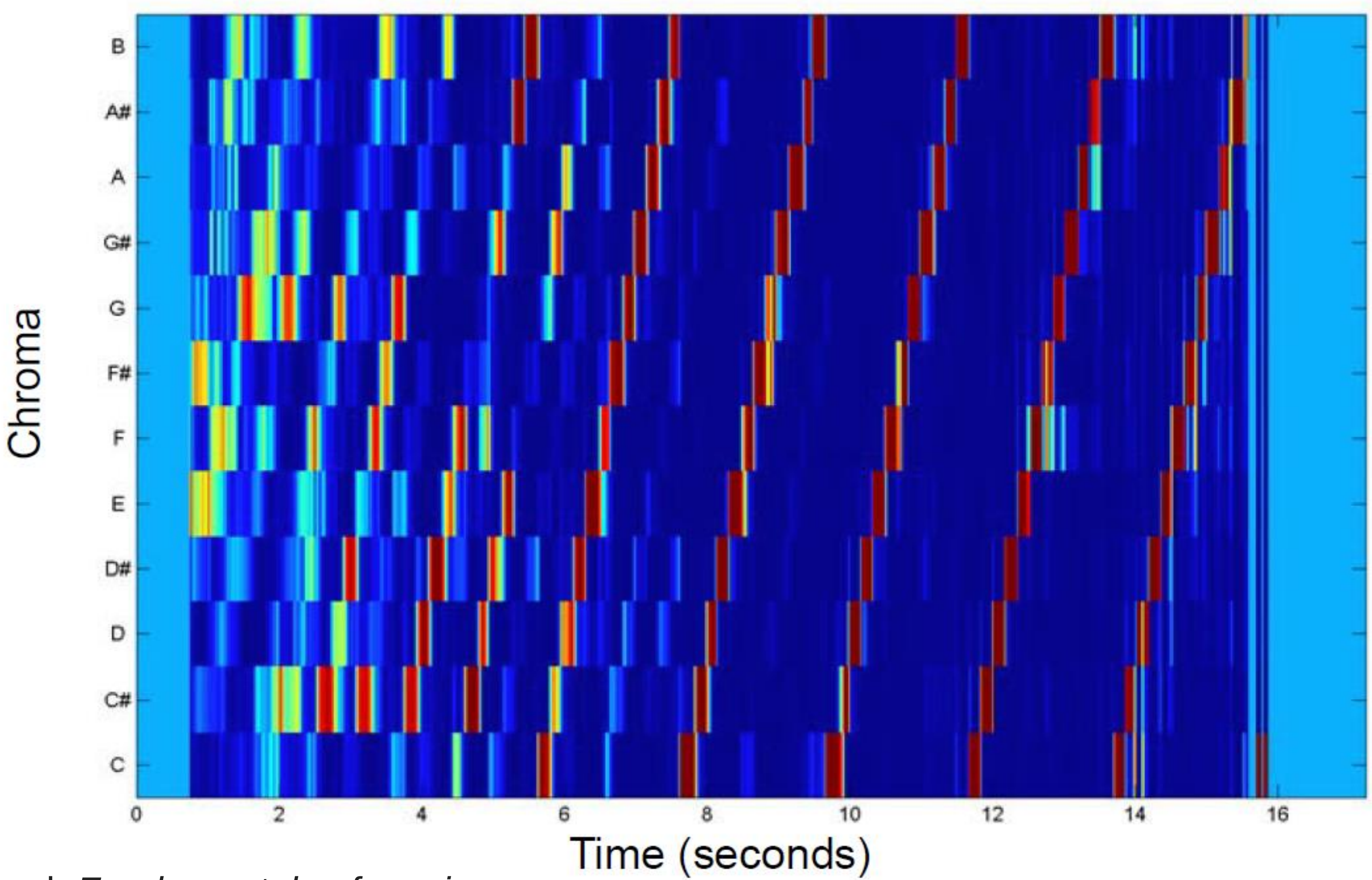


Figure from Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Chromagram of Polyphonic Music

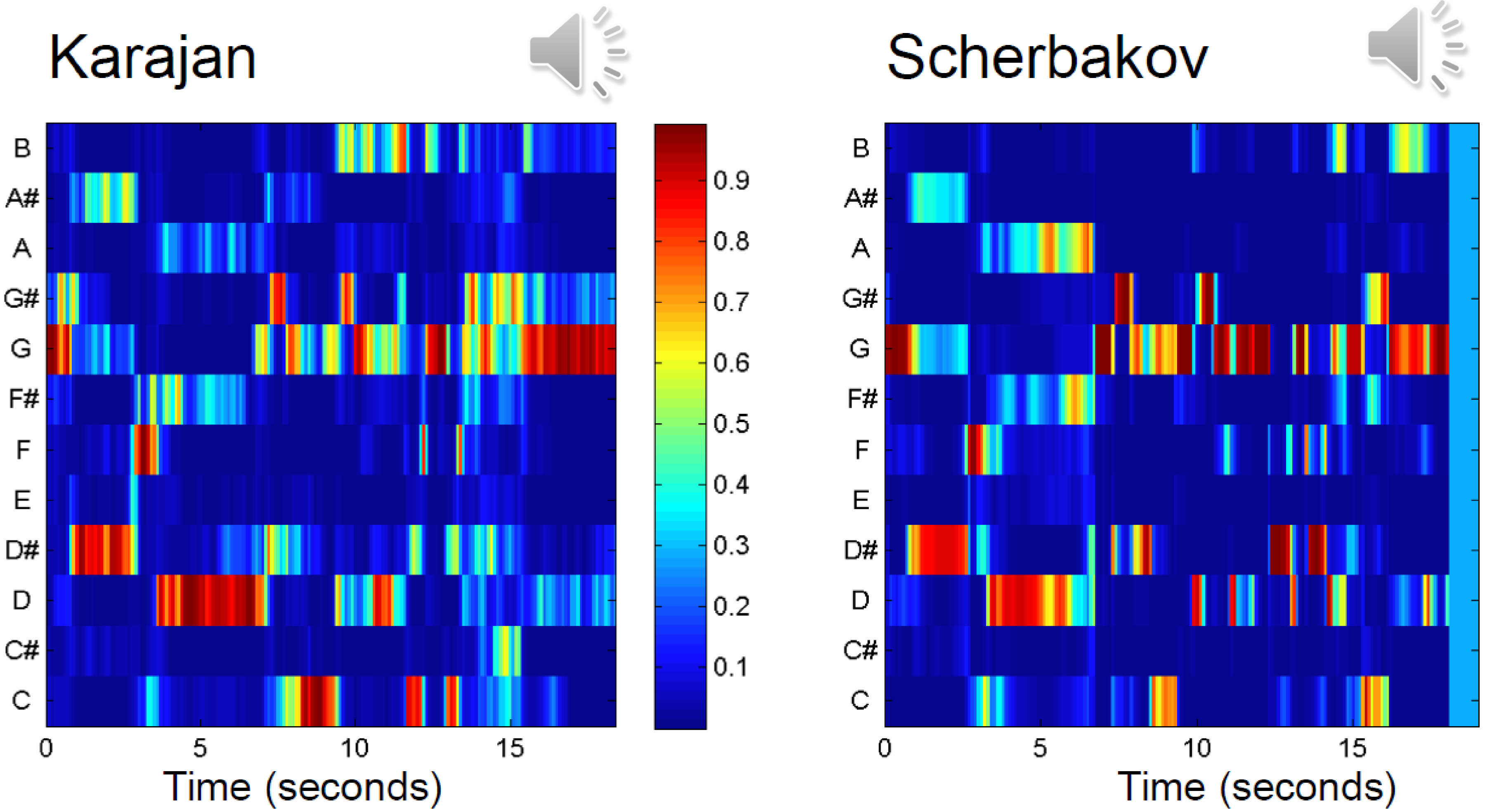


Figure and audio from
Müller, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Harmonic Features

- Fundamental frequency F_0

- Inharmonicity
  - Average deviation of spectral components from perfectly harmonic positions

$$IH = \frac{2}{F_0} \times \frac{\sum_{h=1}^{H} |f_h - hF_0| \times a^2(h)}{\sum_{h=1}^{H} a^2(h)}$$

- Odd-to-even ratio

$$OER = \frac{\sum_{h\ \mathrm{odd}} a^2(h)}{\sum_{h\ \mathrm{even}} a^2(h)}$$

Harmonic Features

- Tristimulus
  - Relative weights of low and high harmonics

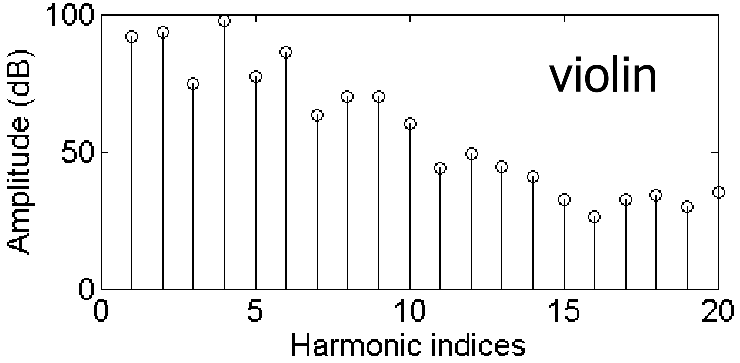
$$T1 = \frac{a^2(1)}{\sum_{h=1}^{H} a^2(h)}$$

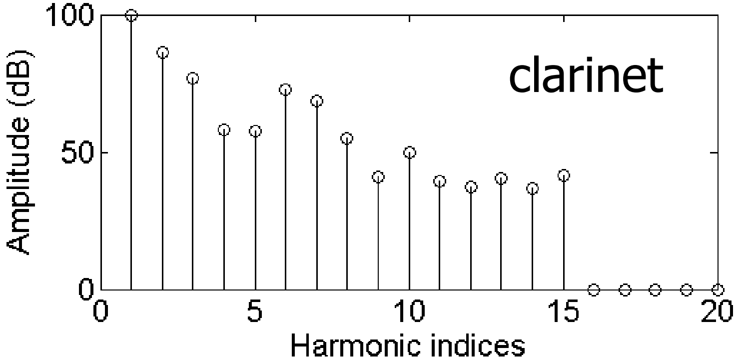
$$T2 = \frac{a^2(2) + a^2(3) + a^2(4)}{\sum_{h=1}^{H} a^2(h)}$$

$$T3 = \frac{\sum_{h=5}^{H} a^2(h)}{\sum_{h=1}^{H} a^2(h)}$$

Harmonic Features

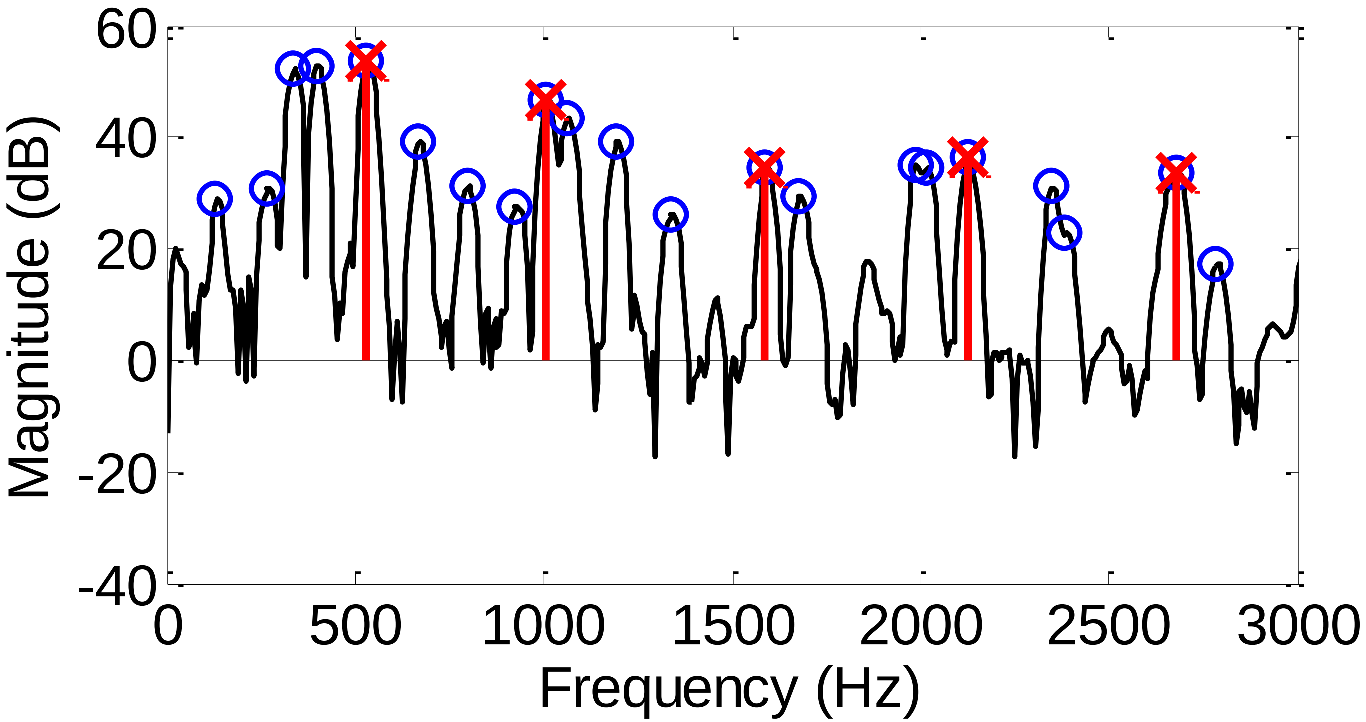
- Harmonic structure
  - Relative normalized amplitudes (dB) of harmonics





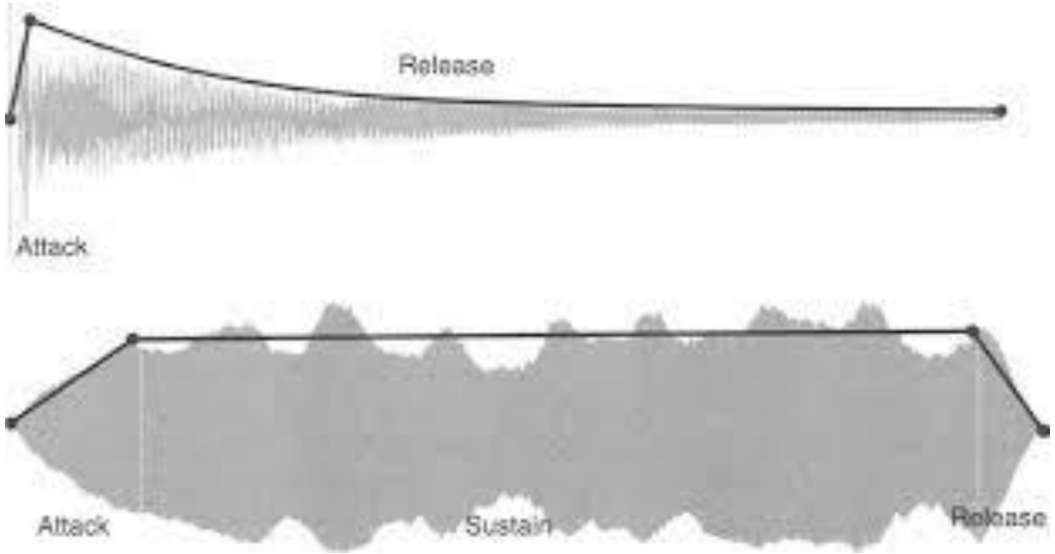
Harmonic Structure Calculated From Mixture

- Assume the F0 of the source is given
- Detect the closest peak for each harmonic



Temporal Features

- Amplitude envelope



- Attack time

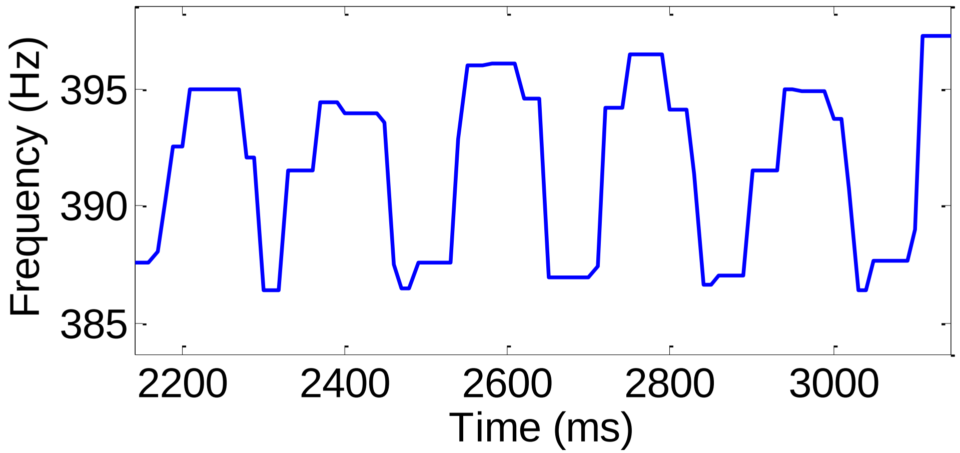
$$LAT = \log_{10}(t_{80} - t_{20})$$

Figure from Anssi Klapuri, and Manuel Davy, editors. Signal Processing Methods for Music Transcription. Springer, 2006.

Temporal Features

- Vibrato rate and depth
  - How fast and how much the pitch changes

Pitch contour of
a violin note



  - Around 5-6Hz
  - How to calculate its period and amplitude?
- Tremolo
  - Amplitude changes periodically
  - Perform FFT on the RMS contour

Cepstral Features

- Mel-Frequency Cepstral Coefficients (MFCC)
  - 1. Calculate magnitude spectrum
  - 2. Calculate the mel-scale filterbank response (e.g., 40-d)
  - 3. Take log of the filterbank response
  - 4. Perform discrete cosine transform (DCT) on the 40-d vector in 3.
  - 5. Choose the several (e.g., 15) lowest-order DCT coefficients

Deltas of MFCC

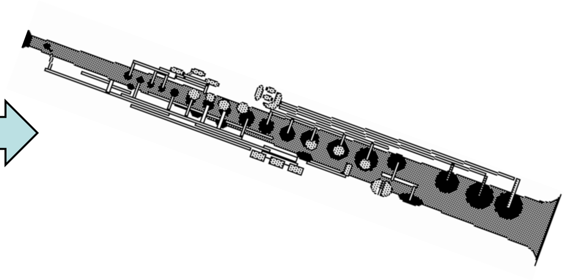
- Capture the temporal evaluation of MFCC
- Delta:
  - "velocity", the local slope. M=1 or 2.

$$\Delta \mathrm{Cep}_i(t) = \frac{\sum_{m=-M}^{M} m \mathrm{Cep}_i(t+m)}{\sum_{m=-M}^{M} m^2}$$

- Delta-delta
  - "acceleration"
- Broadly used in speech/speaker recognition, instrument recognition, etc.

Source-Filter Model

Source Signal $x(t)$ → **Filter** $h(t)$ → **Output Signal** $y(t)$



- Filtering is convolution in time domain, i.e., multiplication in frequency domain.

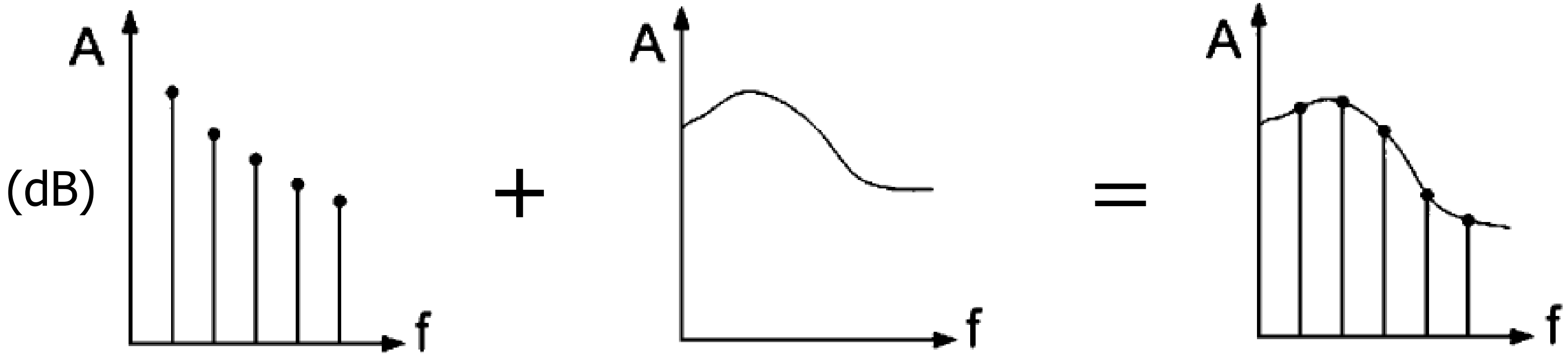
$$x(t) * h(t) = y(t)$$

$$X(f) \times H(f) = Y(f)$$

$$|X(f)| \times |H(f)| = |Y(f)|$$

Harmonic Sounds

Source Signal $x(t)$ → **Filter** $h(t)$ → **Output Signal** $y(t)$



Excitation spectrum (harmonic comb)

Resonance spectrum (spectral envelope)

Harmonic sound spectrum

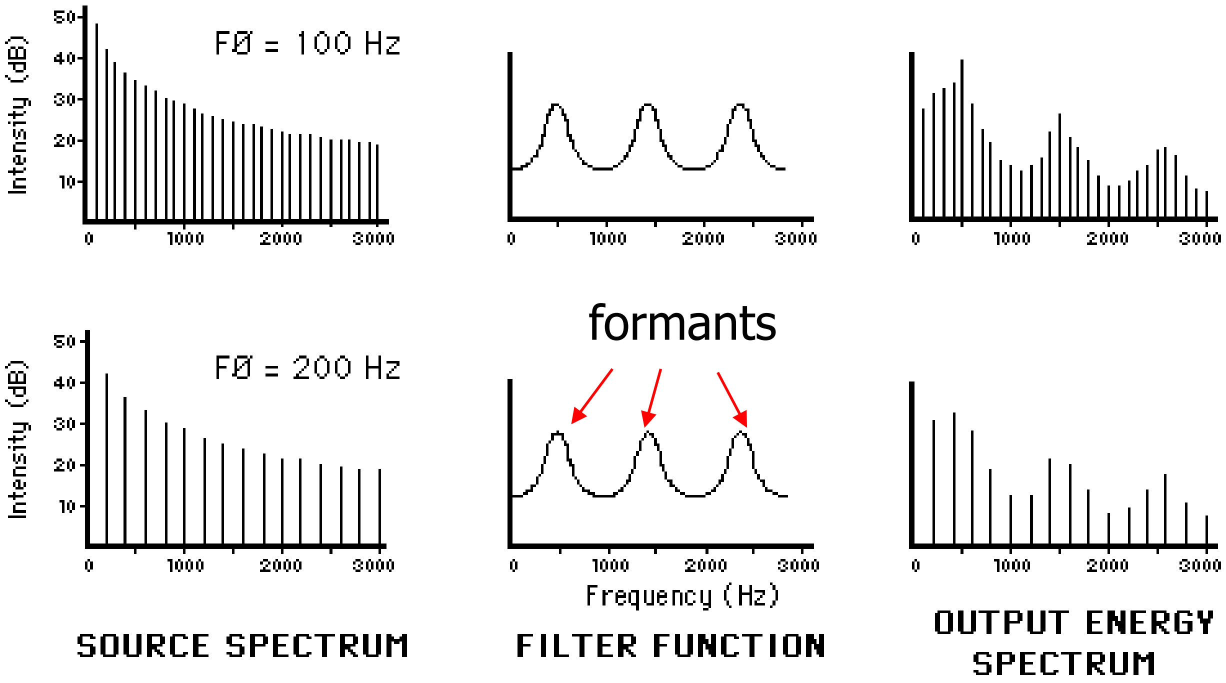
- For log-amplitudes, multiplication becomes addition

$$\log_{10}|X(f)| + \log_{10}|H(f)| = \log_{10}|Y(f)|$$

Spectral envelope → timbre

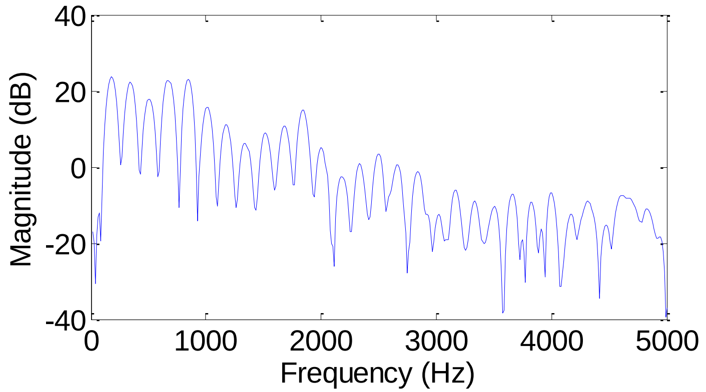
- The excitation spectrum changes with pitch
- The spectral envelope changes with the shape, material, etc. of the resonance body
  - It does not change much with pitch.

Speech production (from Haskins Lab at Yale)



How to characterize the envelope?

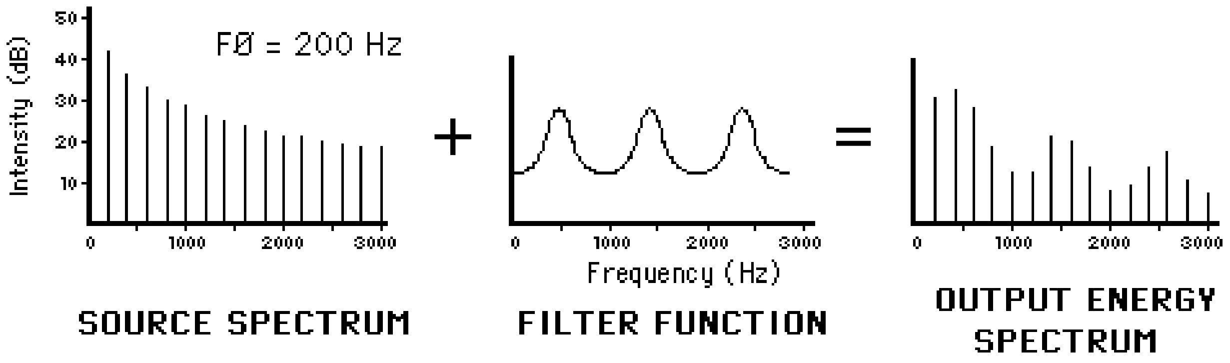
Harmonic sound magnitude spectrum



- First thought
  - Detect peaks
  - Draw a smooth line connecting the peaks
  - This line is the envelope
- How to represent the envelope?
  - Non-parameterized? Very high dimension
  - Parameterized. How?
    - Polynomial?
    - Sinusoidal?

Basic Idea of Cepstrum

- View the log-magnitude spectrum as a mixture of two signals, one high-frequency and one low frequency.



- What if we perform Fourier analysis on the mixture?
  - Fourier transform is linear!
  - Fourier transform separates low/high frequencies!
- Higher Fourier coefficients ⇔ excitation spectrum
- Lower Fourier coefficients ⇔ spectral envelope

Formal Definition of Cepstrum

- Bogert et al. 1963, heuristically

$$\text{power cepstrum} = \left|\mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|^2\}\right|^2$$

- Digital version
  - Use DFT and IDFT to replace Fourier transforms.

- Why IDFT?
  - Well, it actually doesn't matter for real signals.

IDFT or DFT? It doesn't matter.

- Remember IDFT

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j2\pi kn/N}$$

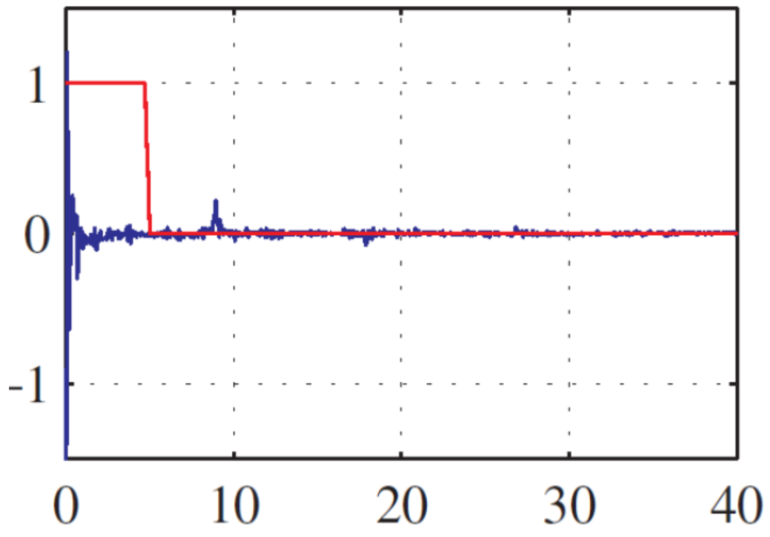
$$= \frac{1}{N}\sum_{k=0}^{N-1} X[k]\left\{\cos\left(\frac{2\pi kn}{N}\right) + \overbrace{j\sin\left(\frac{2\pi kn}{N}\right)}^{\text{Cancelled out}}\right\}$$

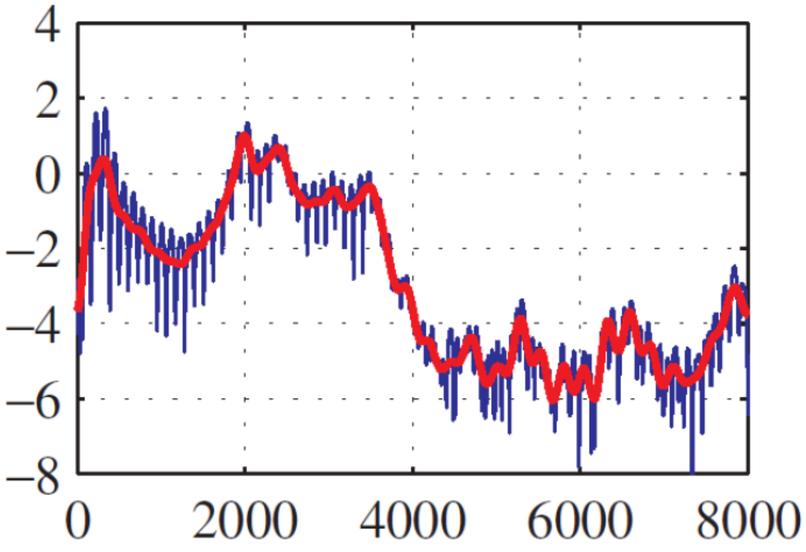
- Now, substitute $\mathrm{a}[k] = \log|X[k]|$ (symmetric, real) as $\mathrm{X}[k]$ into the equation

$$c[n] = \underbrace{\frac{1}{N}\left(a[0] + (-1)^n a\left[\frac{N}{2}\right]\right)}_{\text{DC} \quad \text{Nyquist}} + \underbrace{\frac{2}{N}\sum_{k=1}^{\frac{N}{2}-1} a[k]\cos\left(\frac{2\pi kn}{N}\right)}_{\text{Positive frequencies}}$$

- This is DCT of the positive frequency part of the log-magnitude spectrum

Liftering





Cepstrum

Spectrum

Liftering

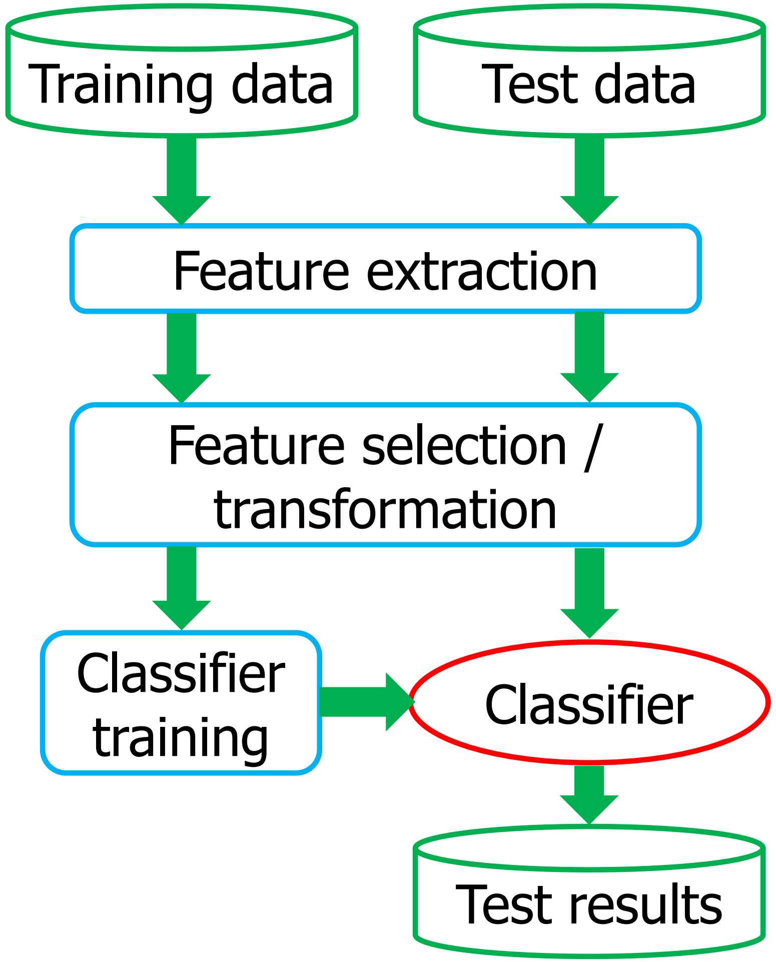
Filtering

Quefrency

Frequency

We extract audio features for downstream tasks

- For example, a classification task
  - Music genres, mood, artist, composer, instrument classification
  - Chord recognition
  - Acoustic event detection
  - Speech/speaker recognition
- General flowchart

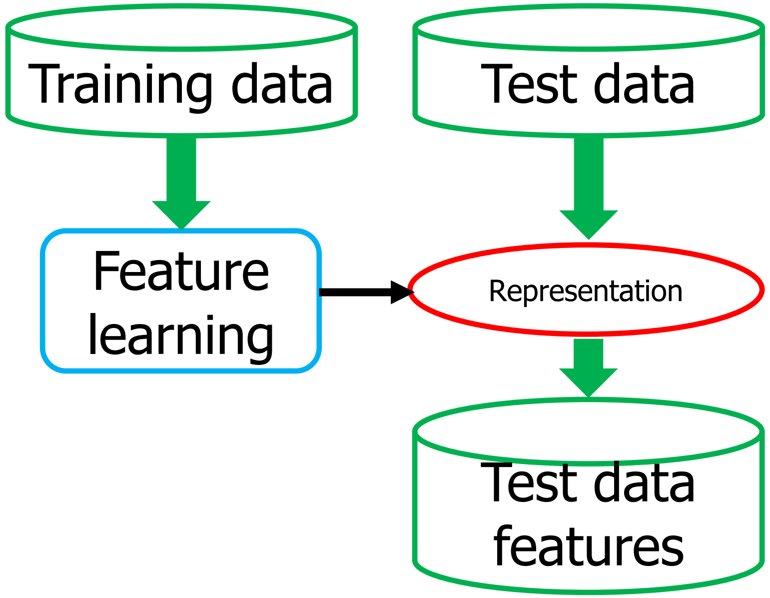


Features Presented Earlier

- Hand-crafted / engineered / pre-defined
- Hard to decide what features to use for a task
- Question: can computers learn features directly from data?

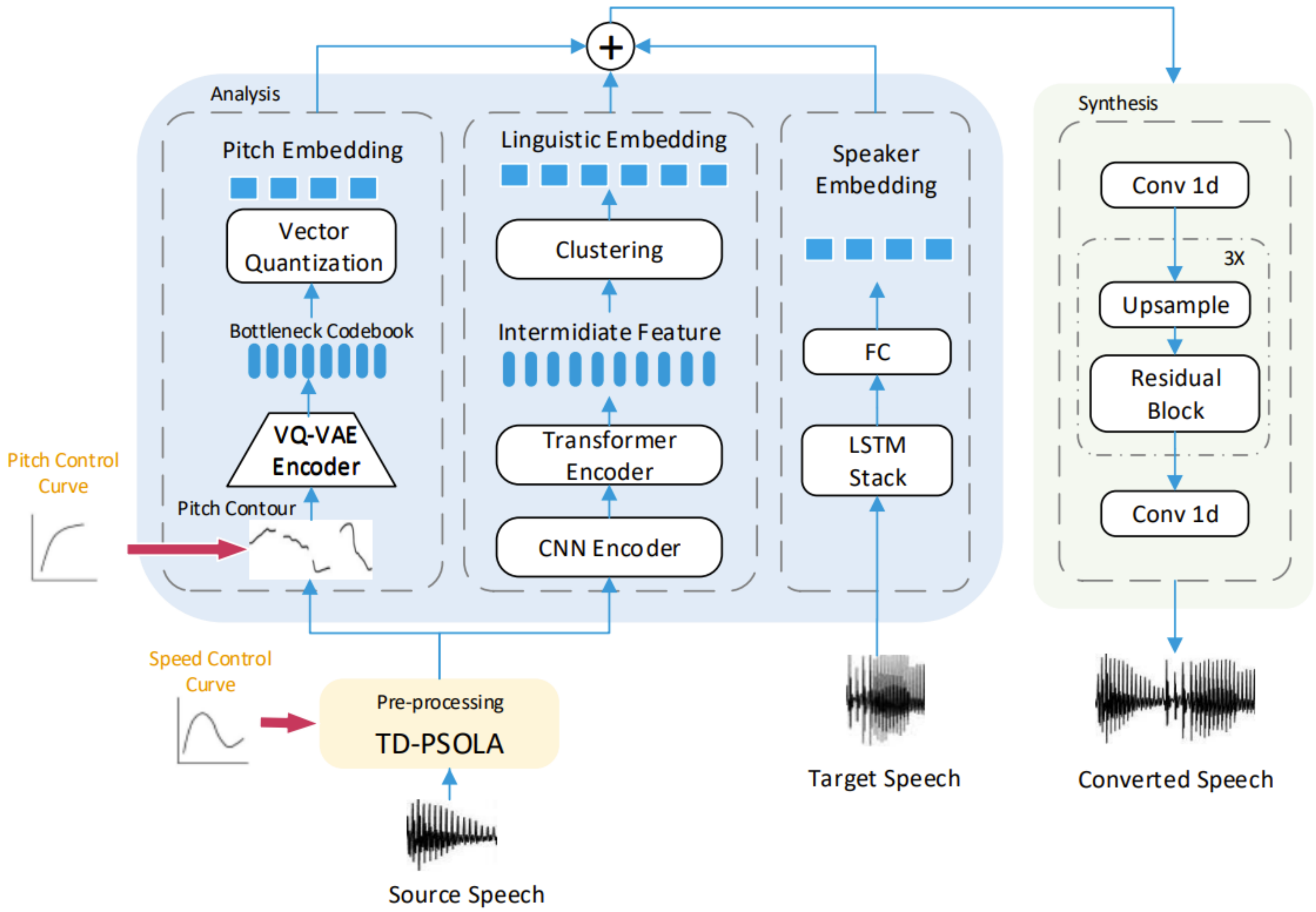
Feature / Representation Learning

- Learn a transformation from "raw" inputs to a representation that can be effectively exploited in a task
- Automatic / not hand-crafted
- Can be adapted for a specific task



Different Ways for Representation Learning

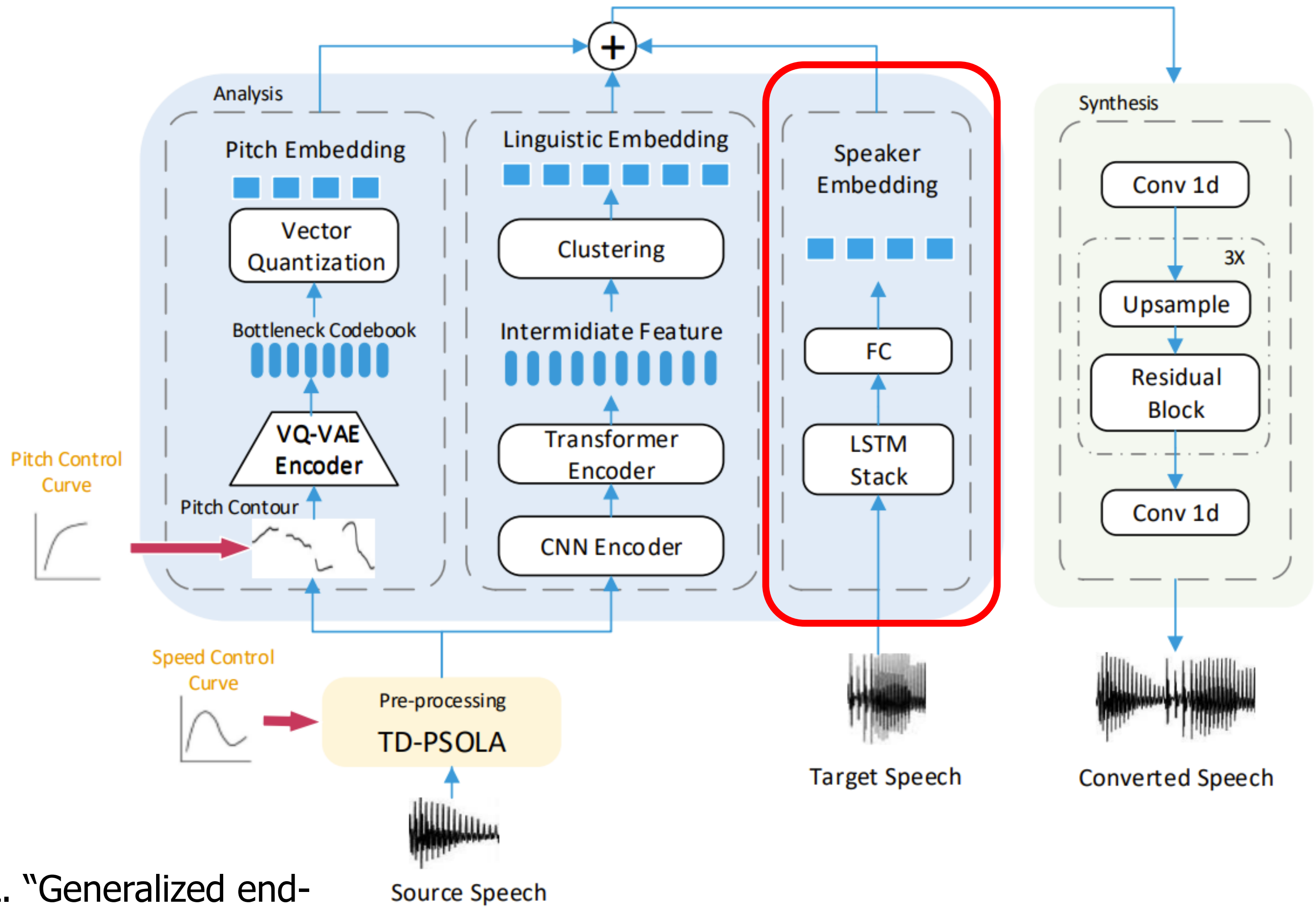
- Let's take this controllable voice conversion system as an example
- Goals:
  - Converting source speaker's voice to target speaker's
  - Allowing time-varying controls on pitch and speed



Meiying Chen and Zhiyao Duan, "ControlVC: Zero-shot voice conversion with time-varying controls on pitch and speed," in *Proc. Interspeech*, 2023.

Supervised Representation Learning

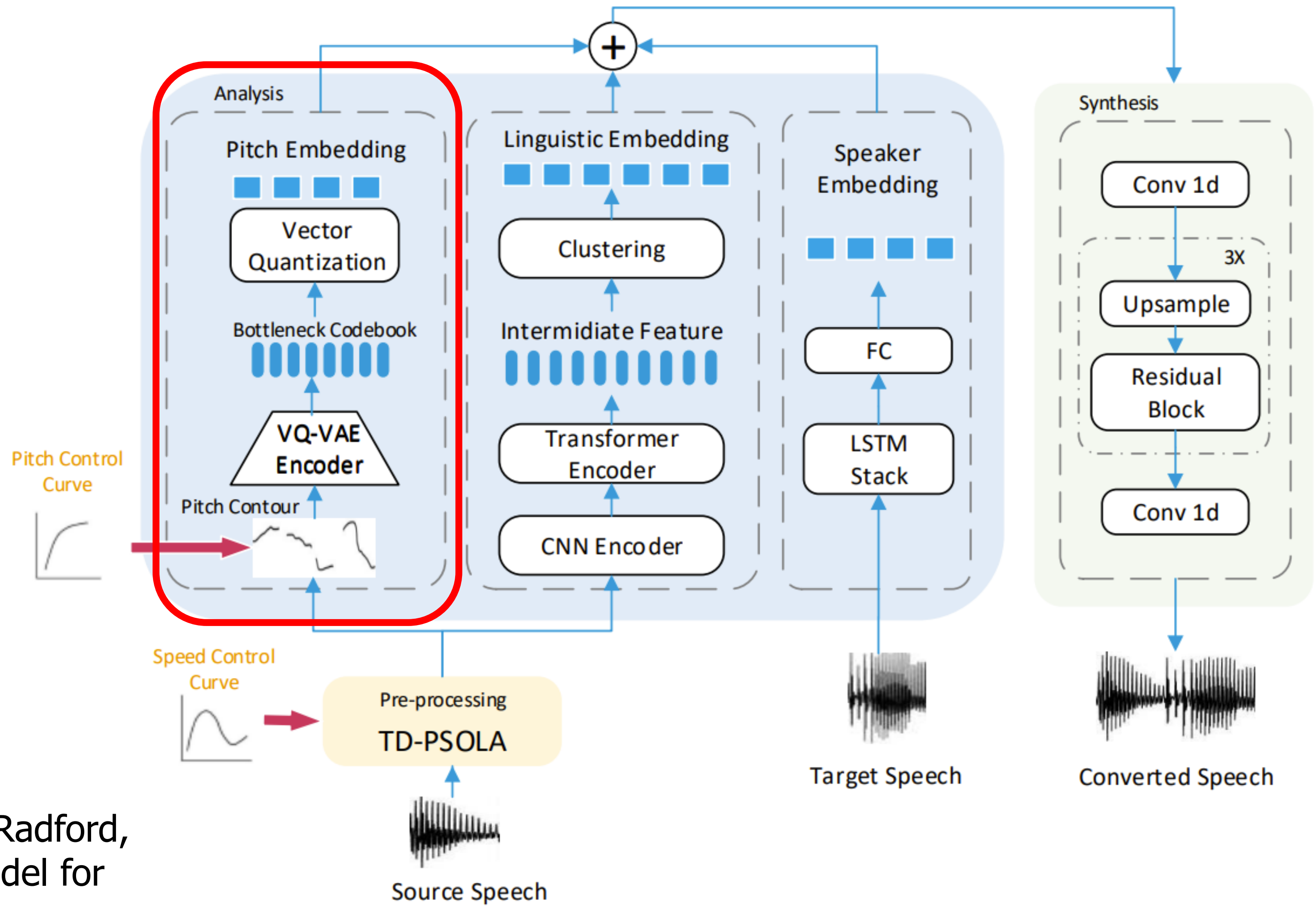
- Training on a supervised learning task (e.g., classification)
  - Training data are <x, y> pairs
- Speaker classification tasks
  - <utterance, speaker ID> pairs



Wan, L., Wang, Q., Papir, A., and Moreno, I. L. "Generalized end-to-end loss for speaker verification," In *Proc. ICASSP*, 2018.

Unsupervised Representation Learning

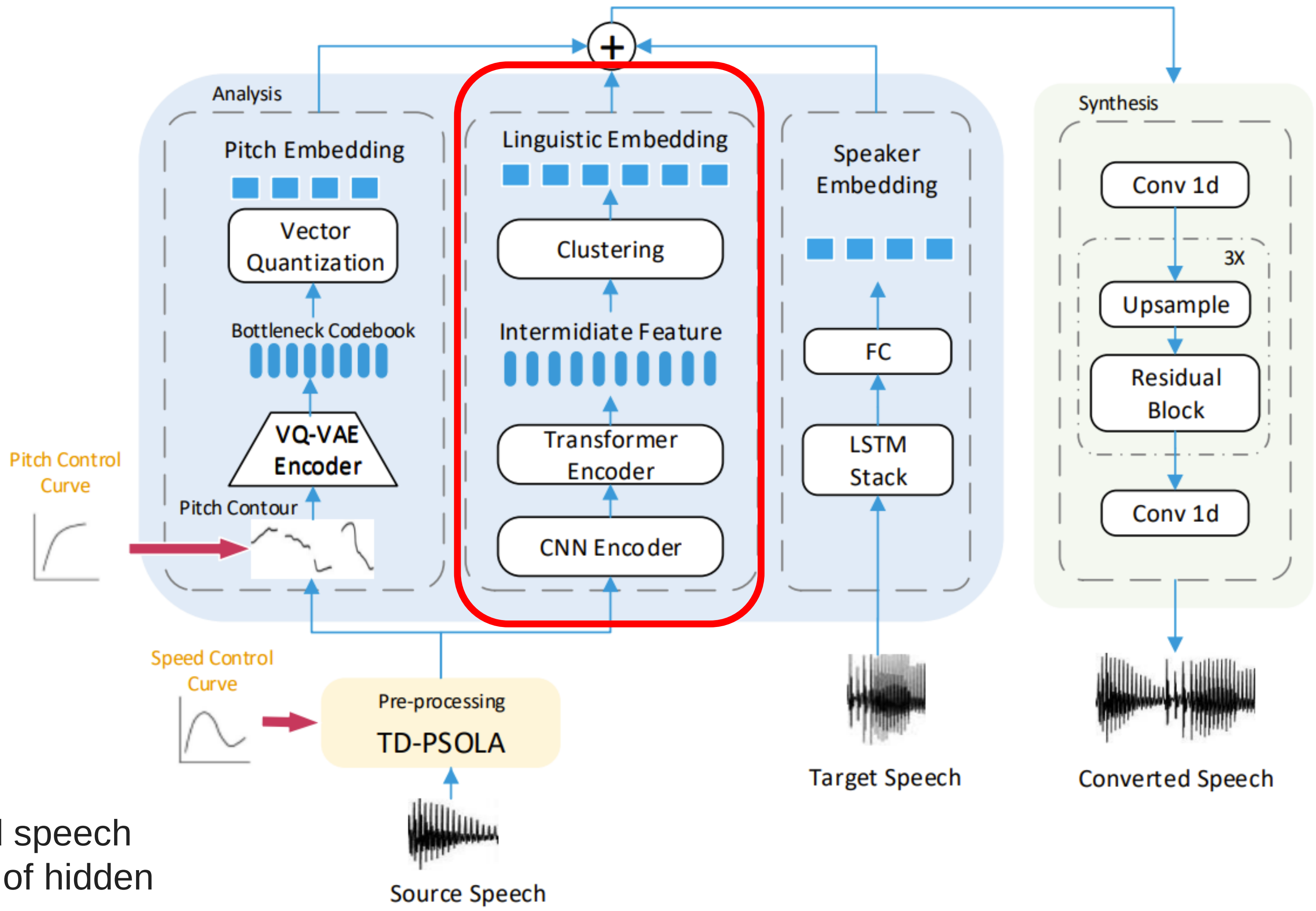
- Train on an unsupervised tasks (e.g., reconstruction)
  - No labels are required during training
- Vector Quantized-Variational AutoEncoder (VQ-VAE)



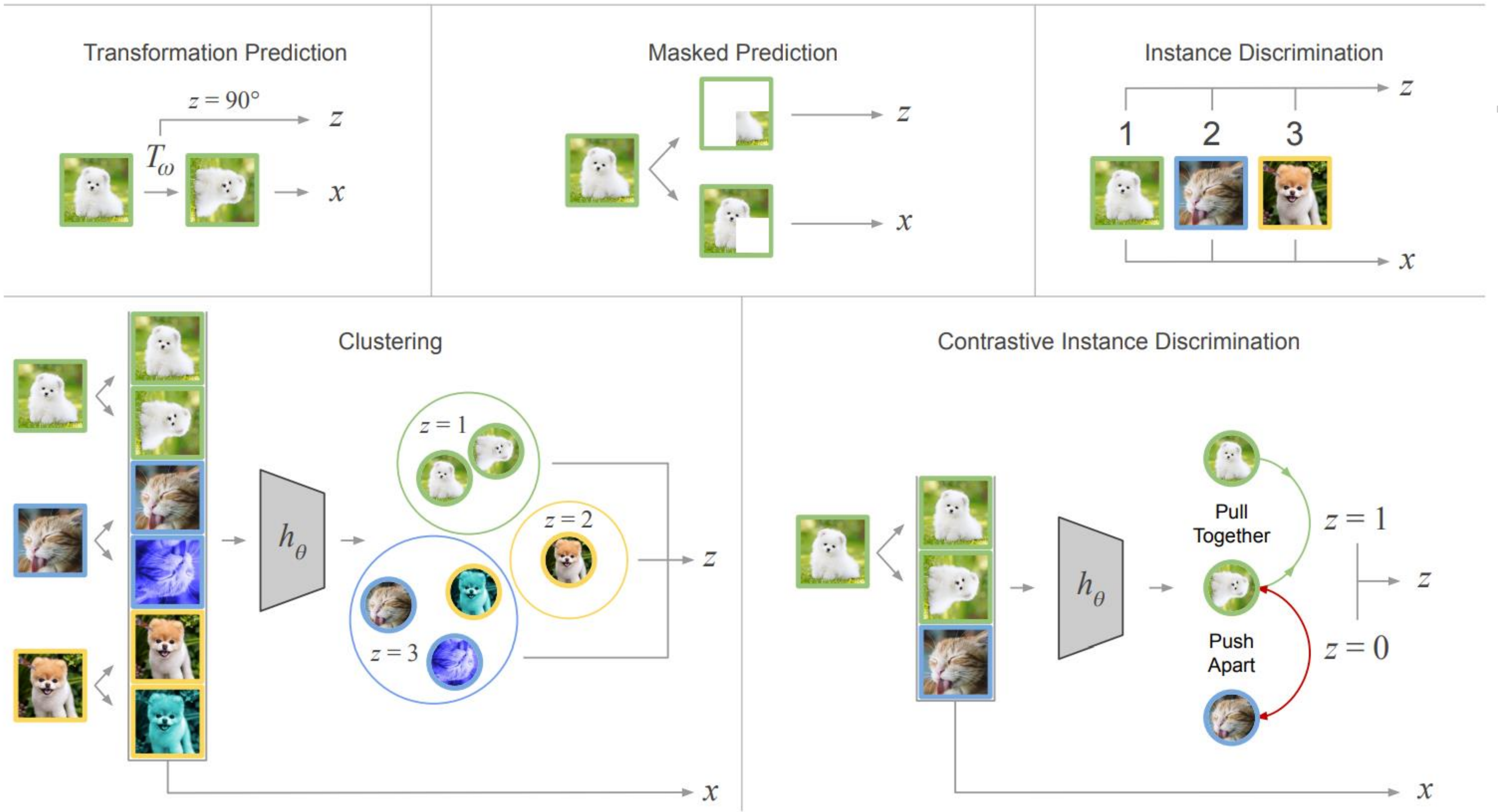
P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," arXiv:2005.00341, 2020.

Self-Supervised Representation Learning

- Train on a self-supervised learning task
  - No labels are required during training
- HuBERT: Hidden Unit BERT



Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM TASLP, 2021*.



Ericsson, Linus, et al. "Self-supervised representation learning: Introduction, advances, and challenges." *IEEE Signal Processing Magazine* 39.3 (2022): 42-62.

Summary

- MIR overview
  - Different types of music data and MIR tasks
  - Relation between MIR and computer audition
- Auditory sensation
  - Auditory system
  - Auditory percepts (loudness, pitch, timbre)
- Psychoacoustic inspirations
  - Auditory scene analysis
- Music audio features
  - Hand-crafted features
  - Representation learning


Thank You!