Sound and speech retrieval

Stanislav Protasov

Agenda

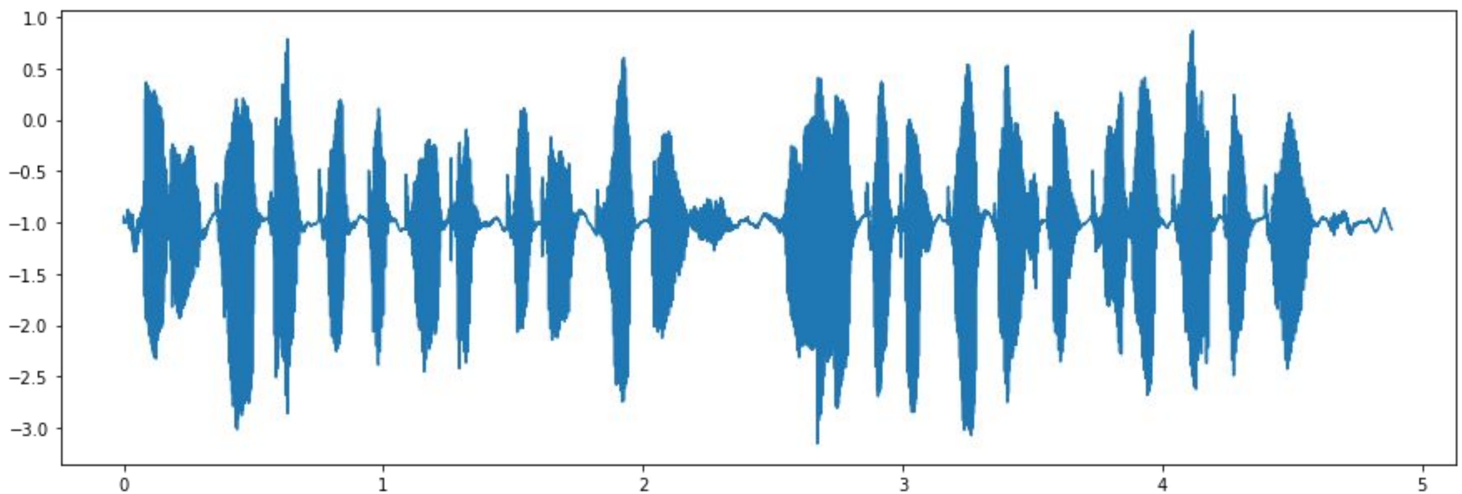
- Sound as a wave
- Music search
- Speech recognition

What is sound and
how humans perceive it?

Hint*: frequencies*

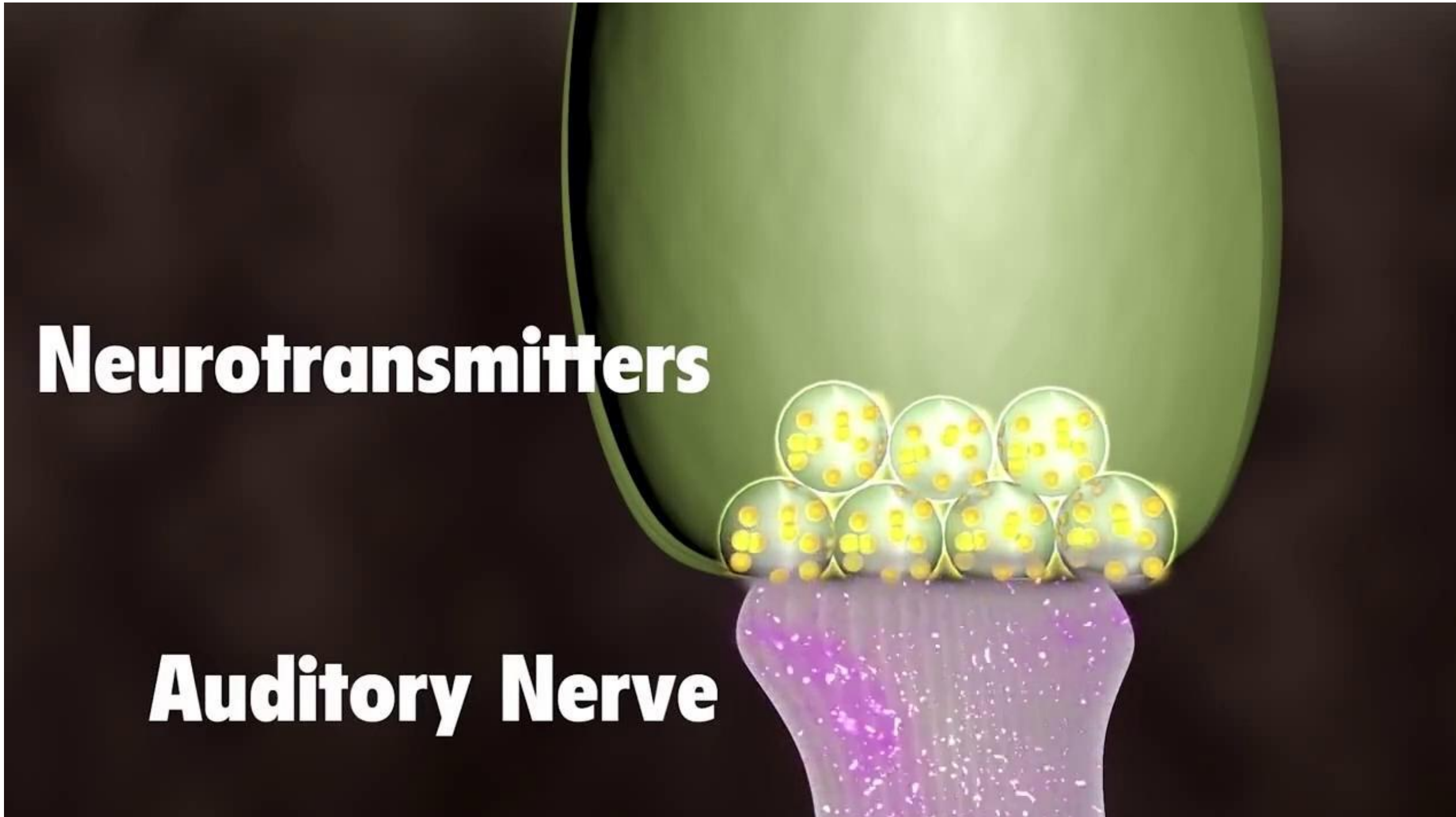
What is the sound?

Sound is a **vibration** that propagates through a transmission medium such as a gas, liquid or solid.



Vinyl player

How ear works

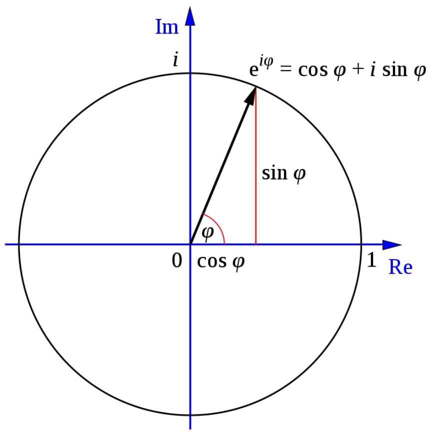


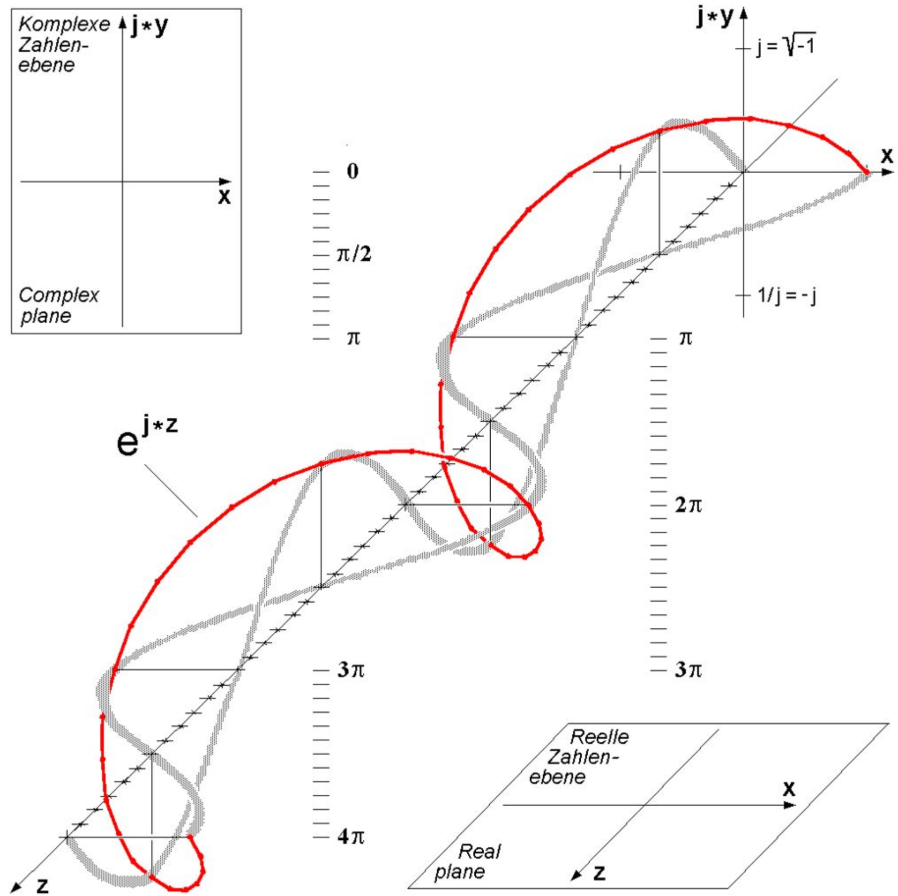
How to repeat this in math?

Hint: *FT*

Euler's identity to link complex exponent with frequencies

$$e^{ix} = \cos x + i \sin x,$$





Fourier Transforms

$$\text{FT}: \quad \hat{f}(w) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i x w} dx$$

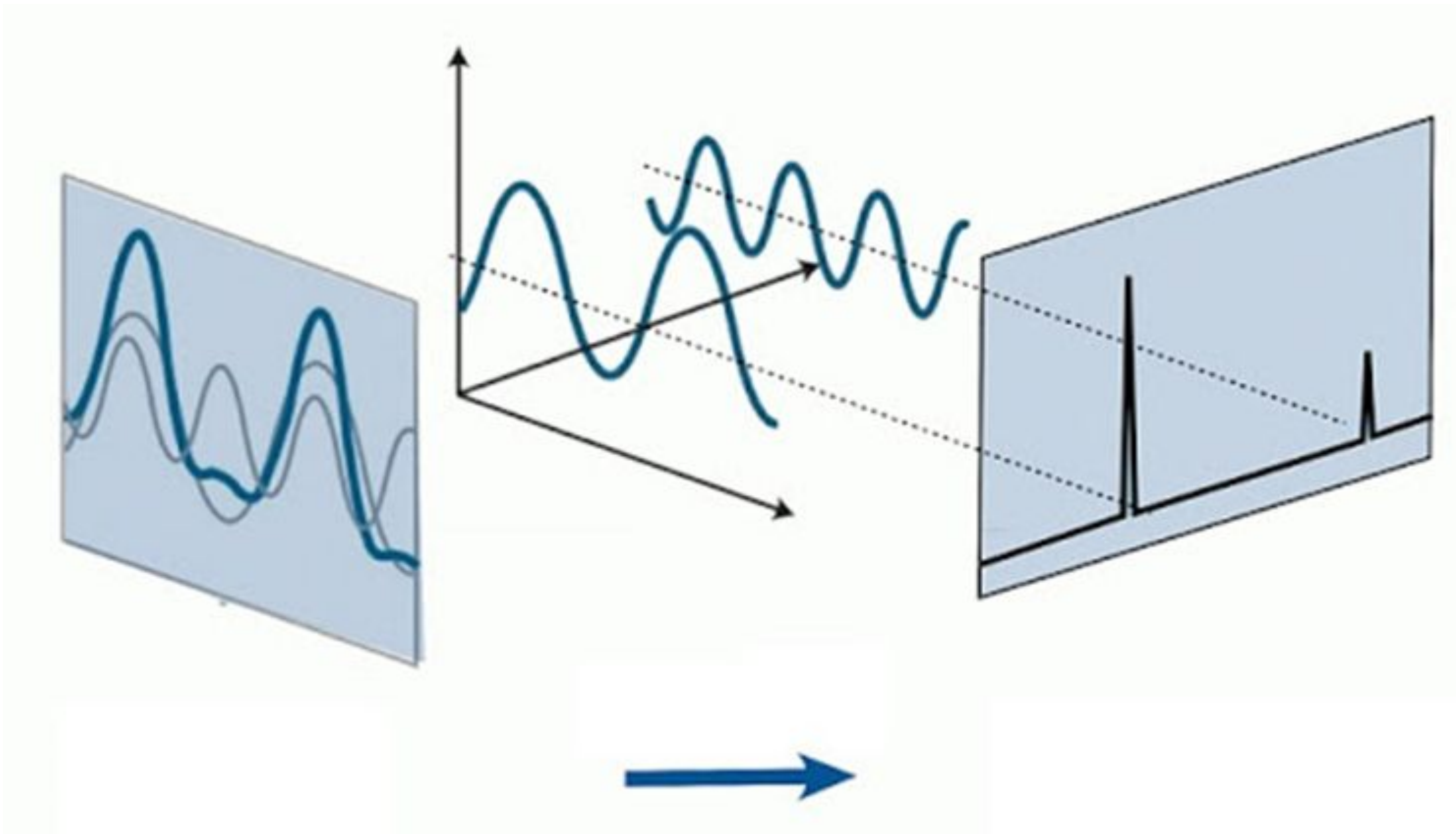
$$\text{DTFT}: \quad X_T(\omega) = \sum_{n=-\infty}^{+\infty} f(nT) e^{-2\pi i \omega n T}$$

$$\text{DTFT + window}: \quad X_T(\omega) = \sum_{n=0}^{M} f(nT) w\left(\frac{n}{M}\right) e^{-2\pi i \omega n T}$$

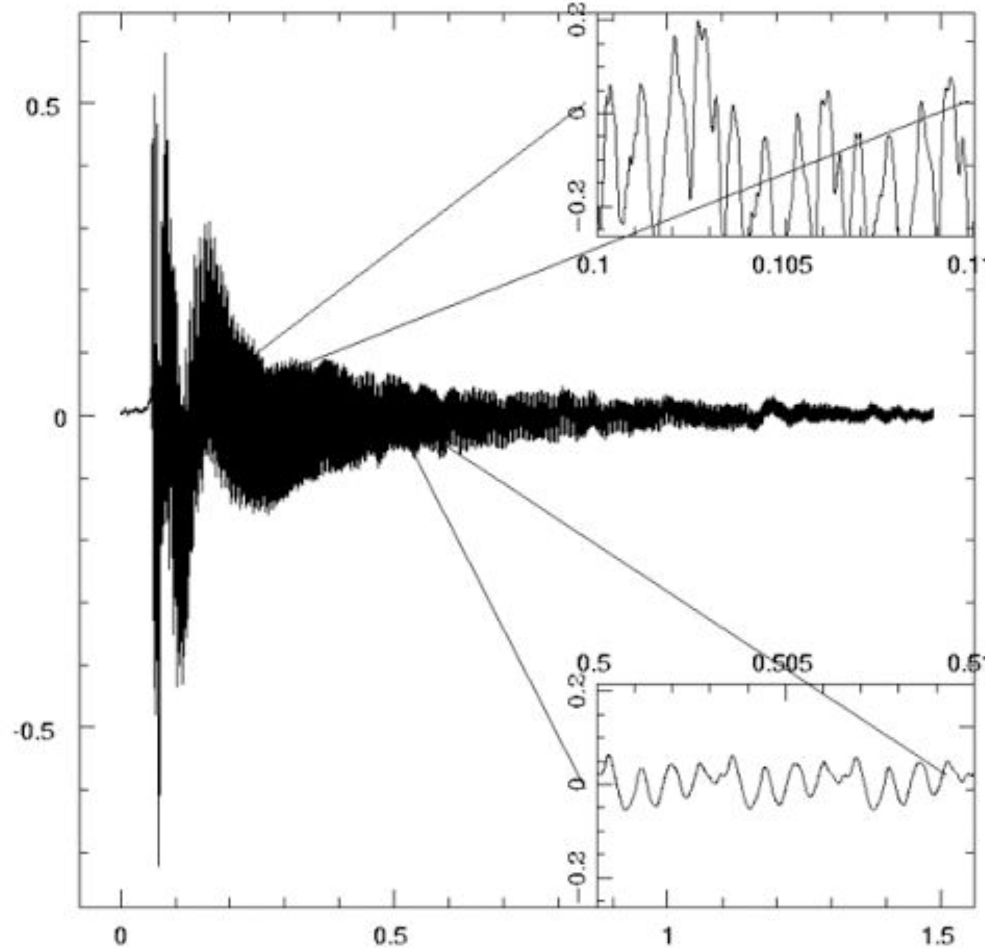
$$k = 0, 1, \ldots, N-1$$

$$\text{DFT}: \quad X_{T,N}(k) = X_T\left(\frac{k}{NT}\right) = \sum_{n=0}^{M} f(nT) w\left(\frac{n}{M}\right) e^{-2\pi i \frac{kn}{N}}$$

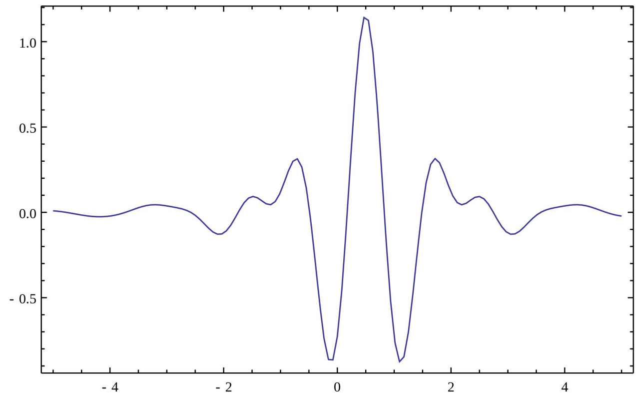
Fourier transform

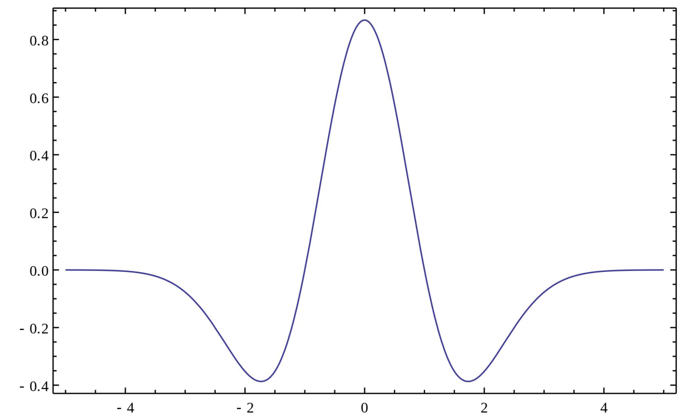


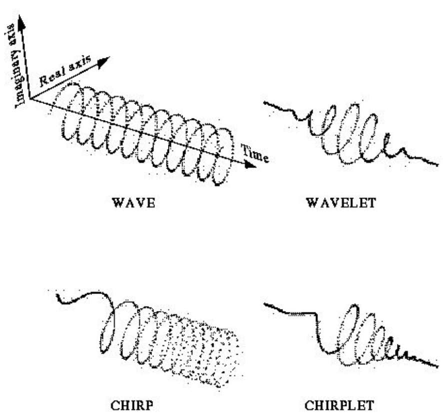
Guitar pitch

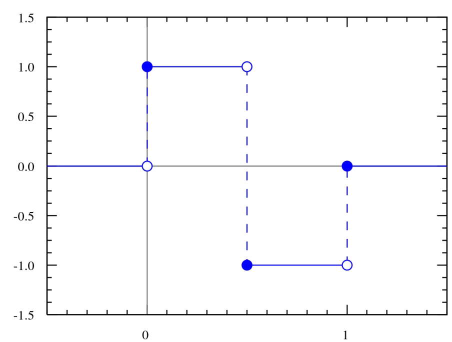


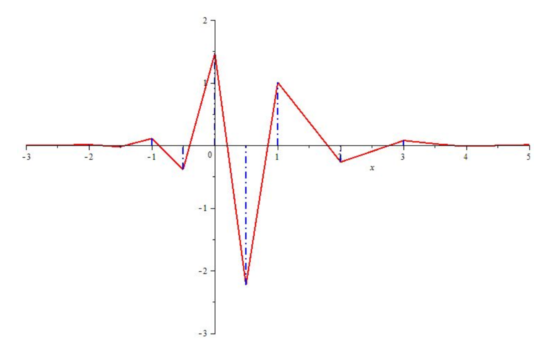
Wavelets





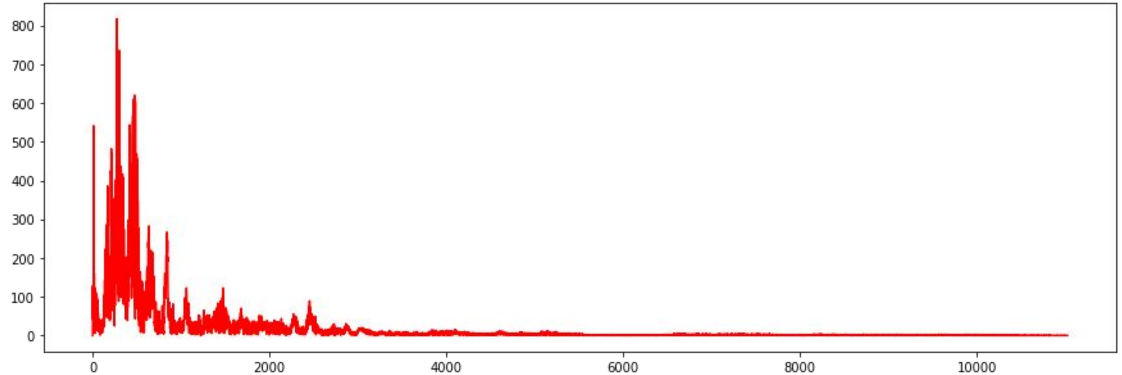






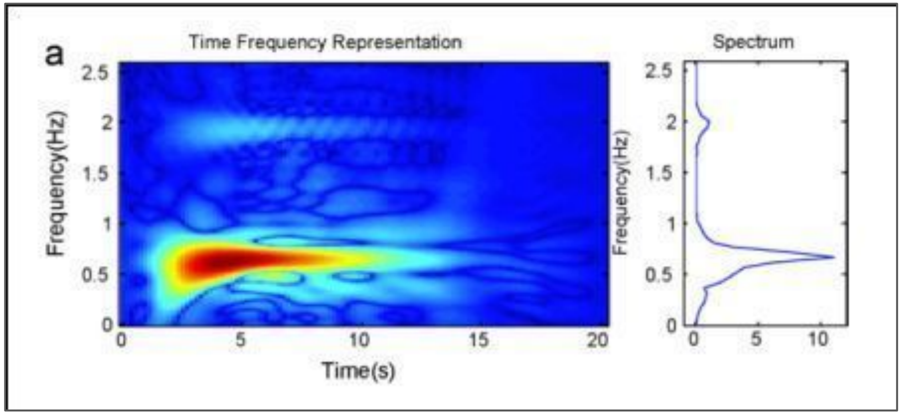
What is the sound for human?

We percept sound using **frequency** receptors. Each moment looks like this:



Also important — we perceive sounds in **log scale**

Timeline is like this:

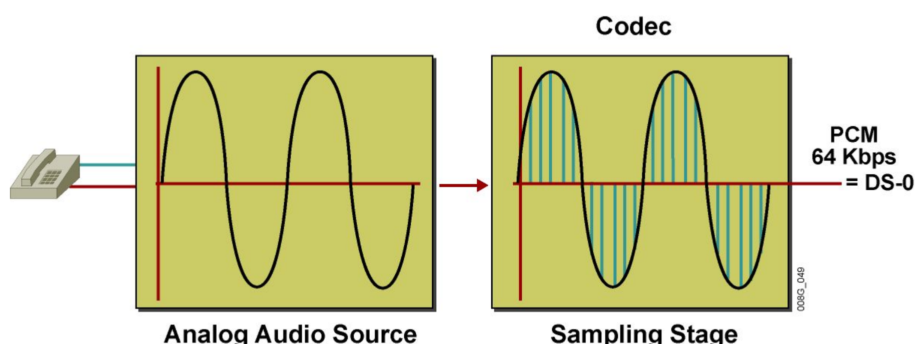


Sound *recording* and *playback*

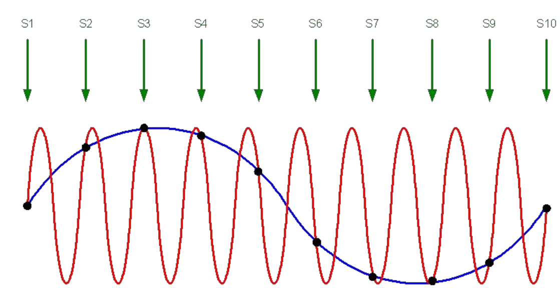
- Digital uncompressed sound consists of regular measurements of signal.
- Measurement frequency is managed using RATE parameter
  - 22050 means 22050 measurements per second (**discretization**)
- How accurate we measure in managed is tuned with format (**quantization**)
  - How many different amplitude values can be encoded
- Channels — number of inputs/outputs (stereo=2, mono=1)
- BPS = RATE * CHANNELS * FORMAT
- Together this is **PCM — pulse code modulation**

Nyquist-Shannon (Kotelnikov) theorem

If a function **x(t)** contains no frequencies higher than ***B*** hertz, it is **completely determined** by giving its values at a series of points spaced **1/(2B)** seconds apart.



V$\{\sin(k\,x) = \sin(n\,x),\ n < k\}$sing. ***n(k)***?



- `sin(a)+sin(b) = 2·sin(½(a+b))·cos(½(a-b))`

Takeaway:

Ok, machine can represent sound wave in human-like form with no information loss

Music fingerprinting

Takeaway for exact search:

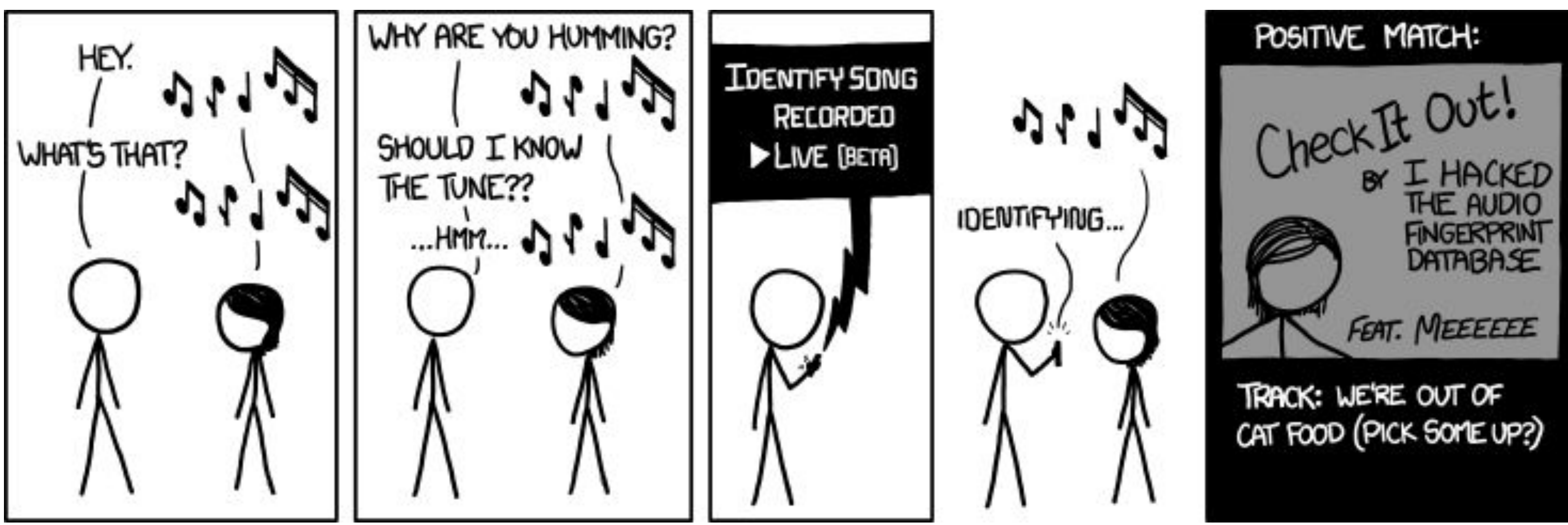
1. Find peaks on spectrogram
2. Use their relative positions in freq-time space as descriptors
3. Maximize descriptors intersection for the query and candidates

Why?

- I like this song, I want to buy it
- Forensic (when was this song playing)
- **Copyrights** (see youtube or instagram policy)

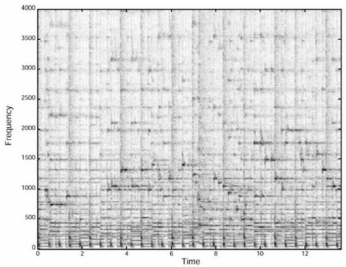
How to form a query

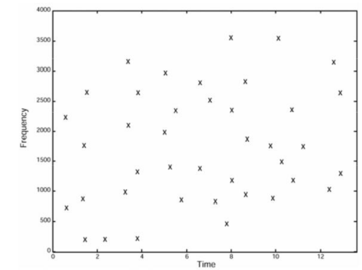
1. Exact sample
2. Humming



Shazam exact match algorithm (1)

1. Robust spectrogram. (Log-scale bins of frequencies)





2. Build **pairs for hashing (32bit)**: anchor point + other point from target zone.

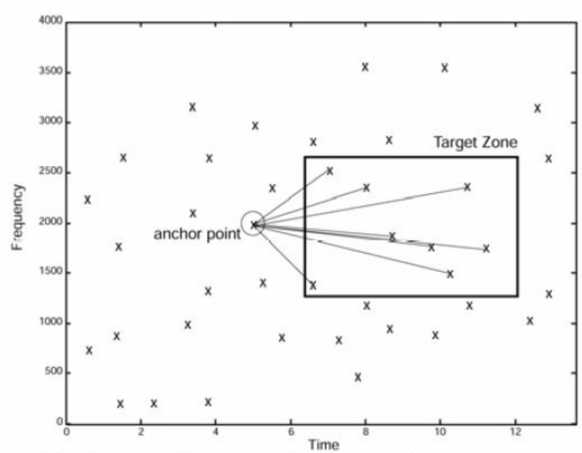


Fig. 1C - Combinatorial Hash Generation

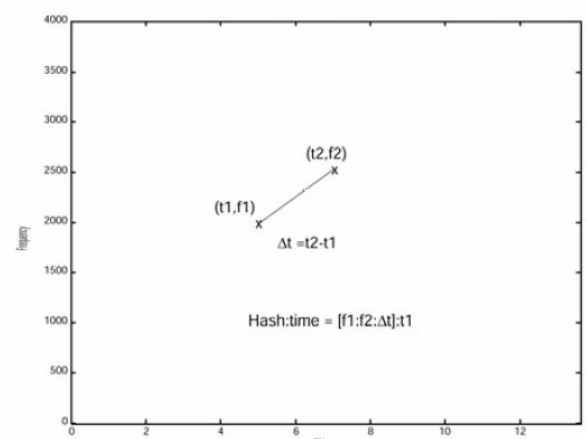


Fig. 1D - Hash details

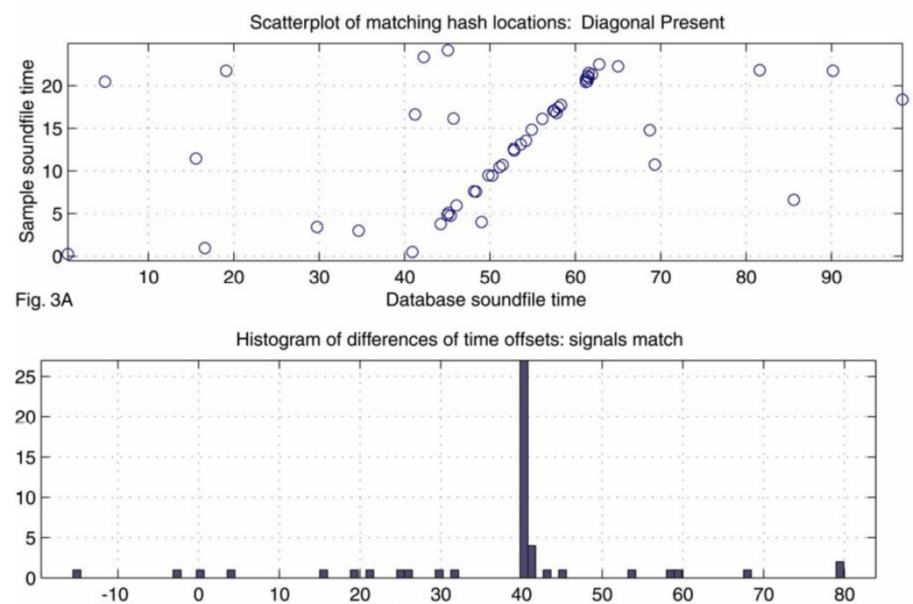
Shazam algorithm (2)

3. Put those points to a **hashmap**.

Memory: `[4B (hash) + 4B (val)] * peaks.`

4. Query for songs with a processed sample.

5. Plot a histogram of offsets (get times from HT, put them in bins), identify real offset



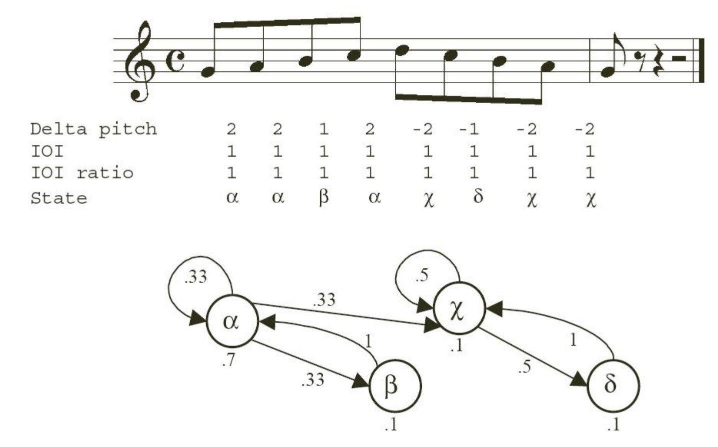
Other fingerprinting approaches

Exact frequency and tempo are not always important:

- Zero-crossing rate
- Spectrum
- Envelope (spectral flatness, frequency band)
- ...

Query by Humming (QbH)

- Detect **coarse melodic contour**, retrieve by string search
  - S=same note, U=up, D=down
  - E.g., Beethoven's 5th: – **S S D U S S D**
  - OR U/D/S – but with five contour levels
- Add **rhythm information**
- Use **beat information**
- Use **HMMs** to represent song database
- ***Dynamic Time Warping*** (DTW) based algorithm, match waveform directly



Dynamic Time Warping

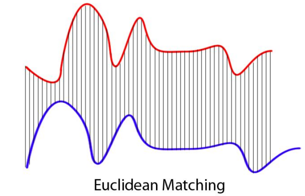
- Take to sequences of lengths N and M
- Build `N*M` matrix ***d*** of distances (diffs)
- Build a matrix of deformation,

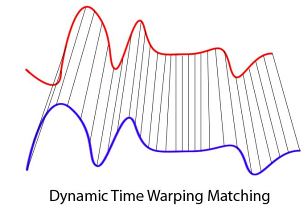
$$D_{i\ j} = d_{i\ j} + min(D_{i-1\ j}, D_{i-1\ j-1}, D_{i\ j-1}). \qquad (3)$$

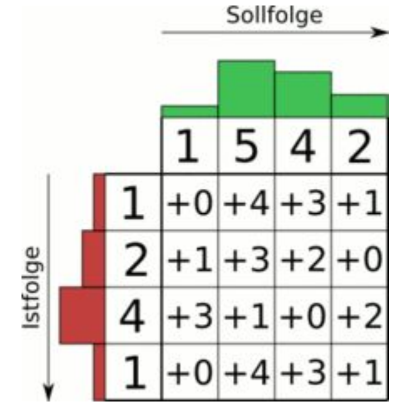
- Search for a path `(1,1)-(N,M)` with minimal average value weight.

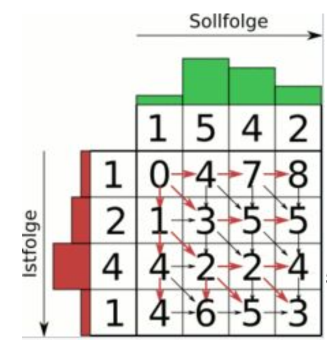
$$DTW(Q,C) = min \left\{ \frac{\sum_{k=1}^{K} d(w_k)}{K} \right\}. \qquad (4)$$

Can give false positives

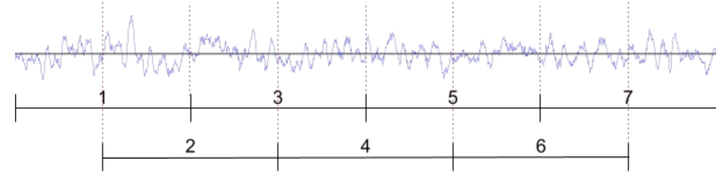








Google Hum to search

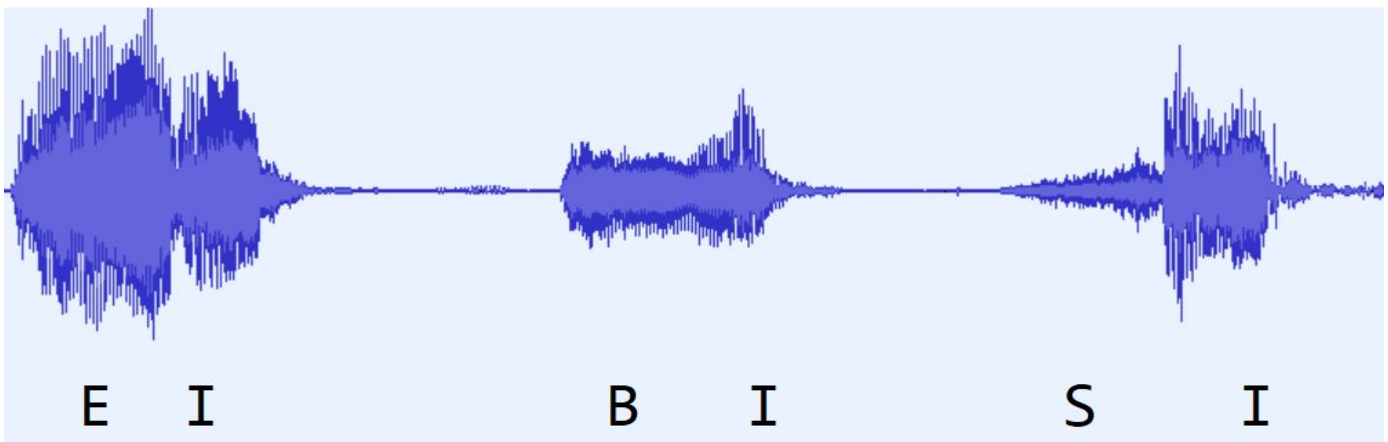


- Convolutional net to build a fingerprint-based from 8 seconds with 0.5 sec. step
- Inaccurate vector search with space partitioning and vector quantization
- Retrieve all candidate’s fingerprints
  - Accurate match on the whole set of candidates’ embeddings

Speech (to text) processing

Acoustic model

As text consist of letters, speech consists of phonemes.



AM: spectrum → phoneme

Language model (in recognition)

Probabilistic model that predicts probability of a word given a sequence of phonemes.

Similar model is used to model sentences of words.

Speech generation

1) Text preprocessing
    a) Number to text
    b) Abbreviations to text
    c) Typo fix
2) Split text into phrases (punctuation, constructions)
3) Phonetic construction (language model)
    a) queue - [kju]
    b) **Арбалетчиков**
        i) `a0 r b a0 lj e`**`1`**` t ch i0 k o0 v`

Speech generation

1) **Accents** are set
   a) Using a dictionary
   b) Using rules
   c) Using statistics (speaker examples)
2) **Reversed acoustic model** is used to consider surrounding
3) **Timbre** is generation with **vocoder**
   a) or RNNs