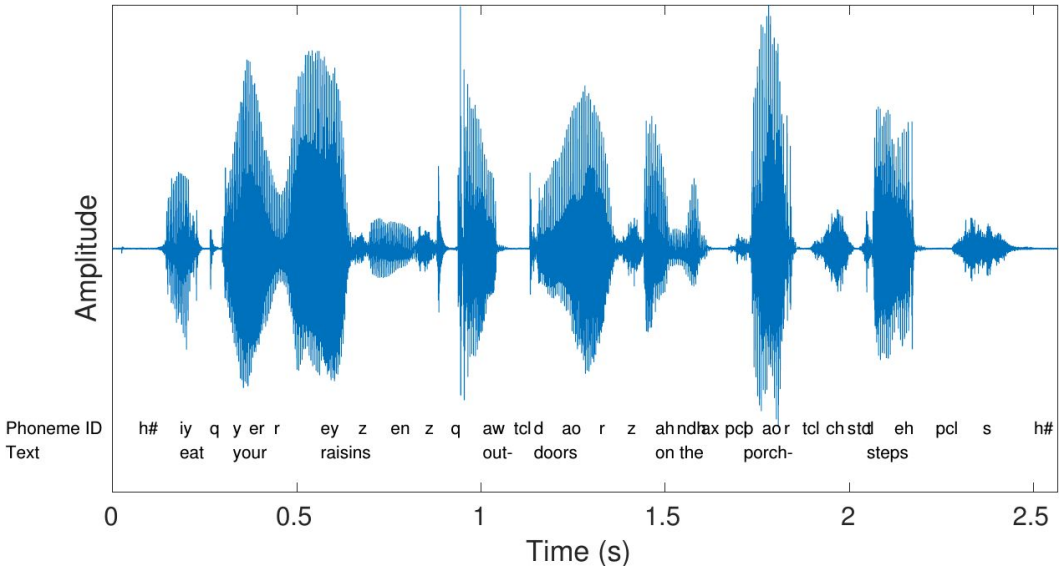
ASR Features

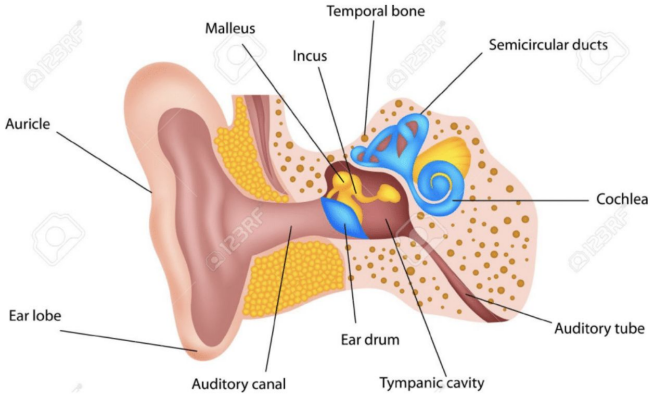
Task example:

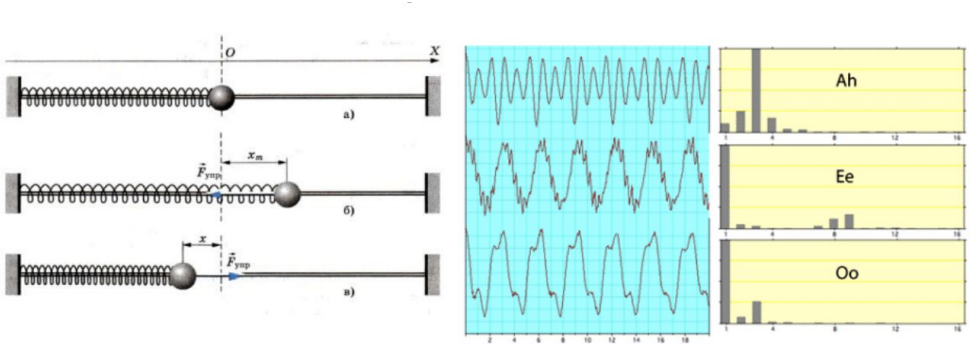
We want to spot some fixed phrase or word.

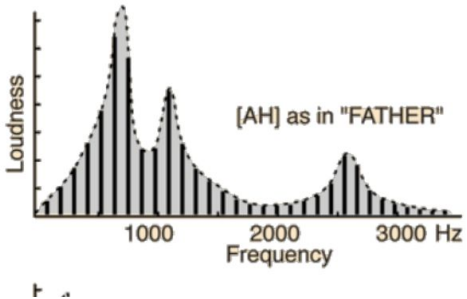
Input is a simple wav file.

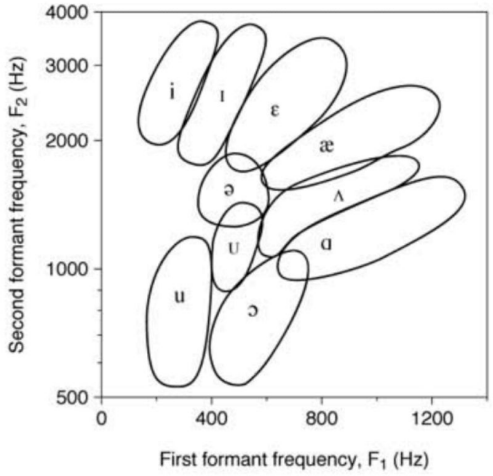


How do human do it?









Credits to: @buriy

Discrete fourier transform

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{i 2\pi k n / N} \qquad \textbf{(Eq.3)}$$

Definition [edit]

The *discrete Fourier transform* transforms a sequence of *N* complex numbers $\{\mathbf{x_n}\} := x_0, x_1, \ldots, x_{N-1}$ into another sequence of complex numbers, $\{\mathbf{X_k}\} := X_0, X_1, \ldots, X_{N-1}$, which is defined by

$$\begin{aligned} X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i 2\pi}{N} k n} \\ &= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N} k n\right) - i \cdot \sin\left(\frac{2\pi}{N} k n\right) \right], \end{aligned} \qquad \textbf{(Eq.1)}$$

Framing

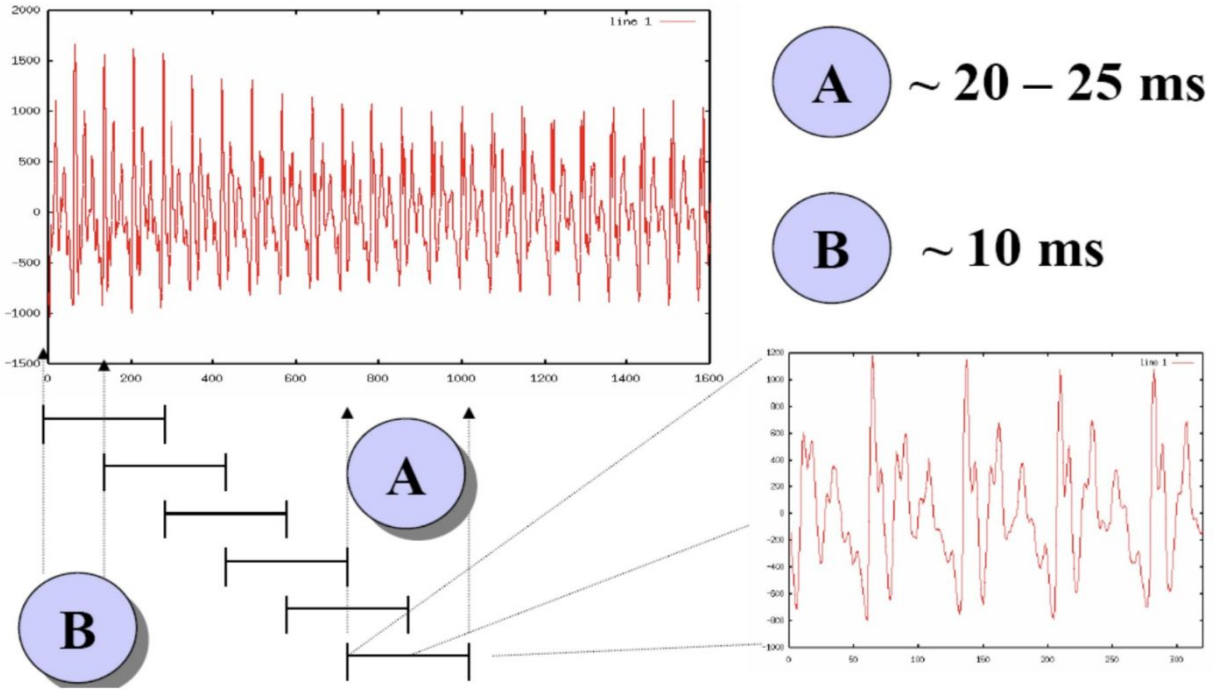
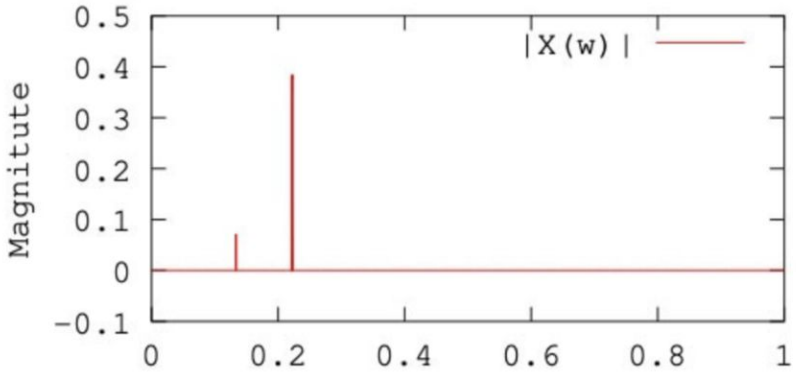
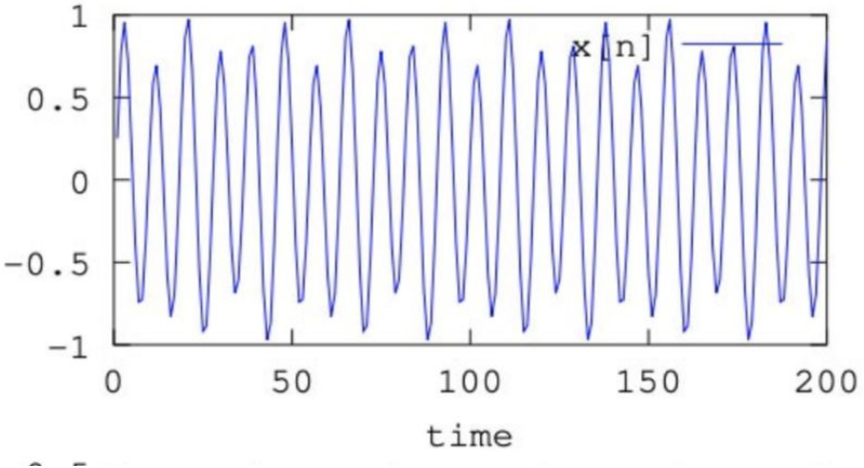


Image from Bryan Pellom

https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

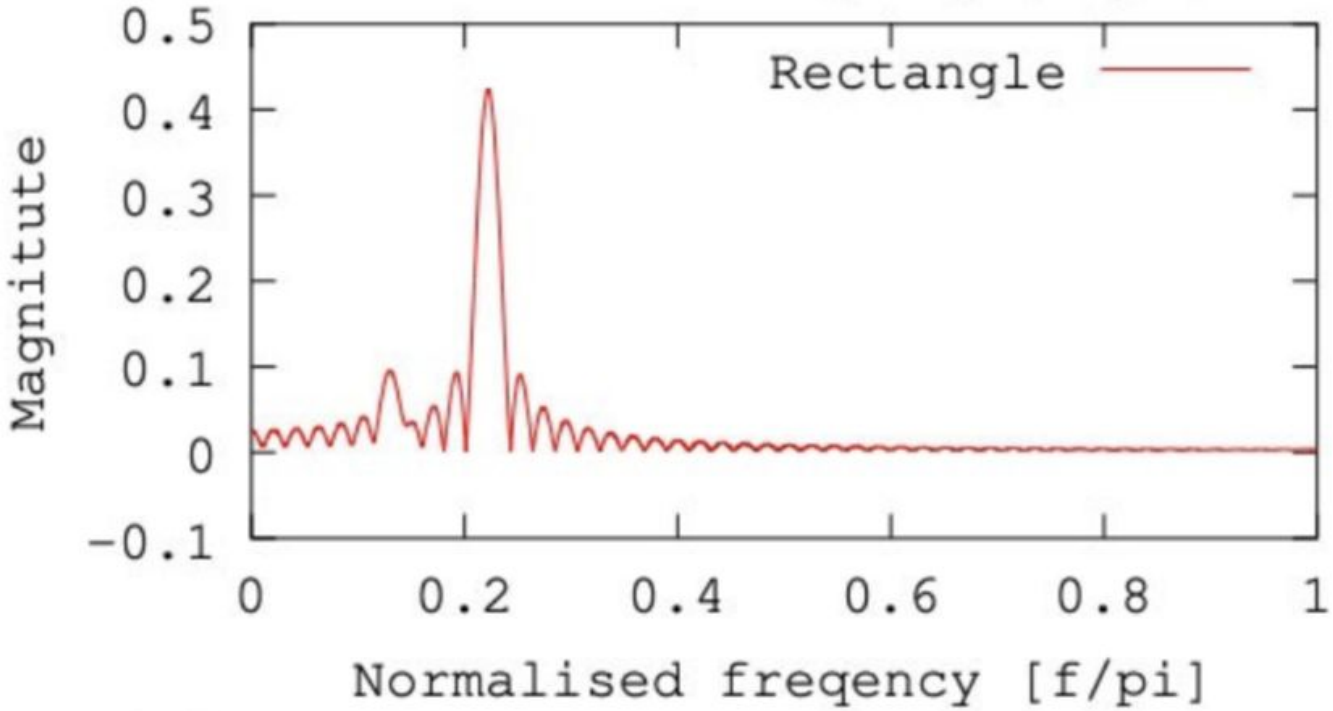
Windowing example (source)





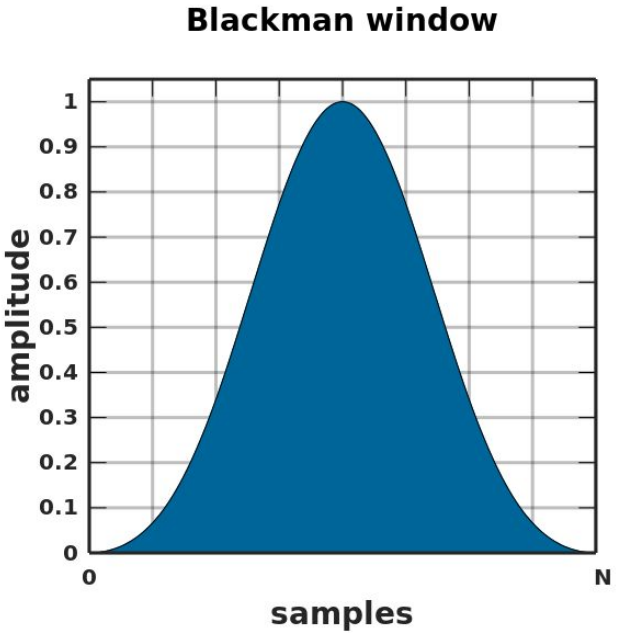
https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

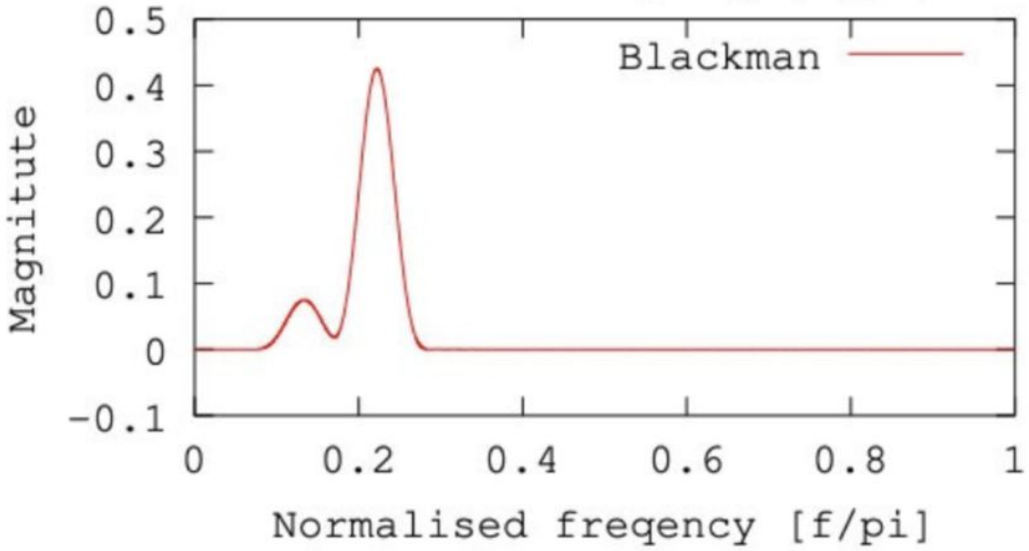
Windowing (naive, rectangular result)



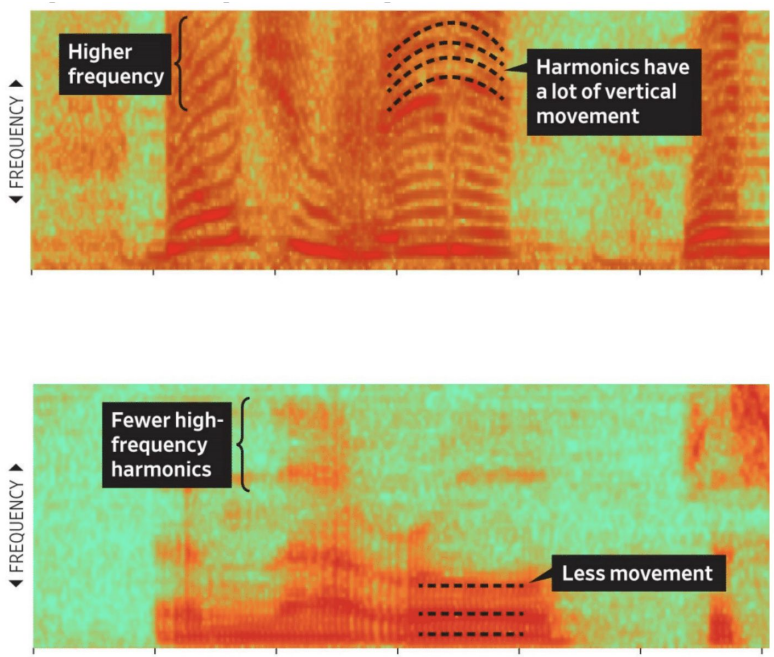
https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

Windowing (with smoothing)





And that way WAV turns into a...

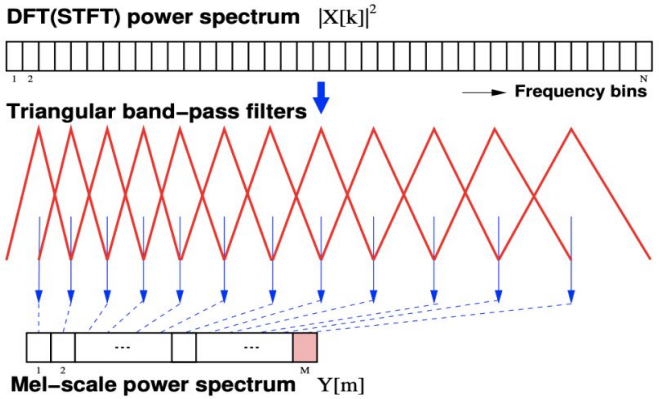


Specific augmentation is usually here too

Credits to: @buriy

Also you can add:

- Pre emphasis (x[t]=x[t] - αx[t-1])
- Delta + Delta-Delta
- Normalization

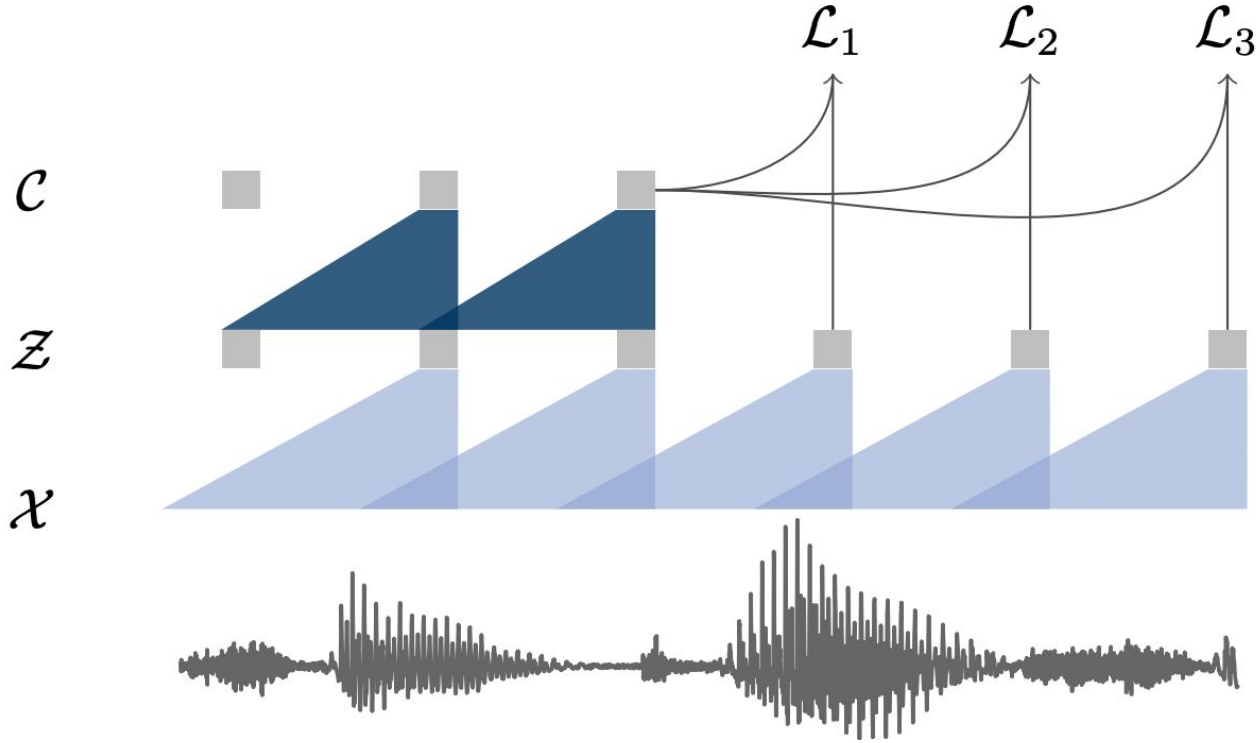


- Mel Scale

$$Y_t[m] = \sum_{k=1}^{N} W_m[k]\,|X_t[k]|^2$$

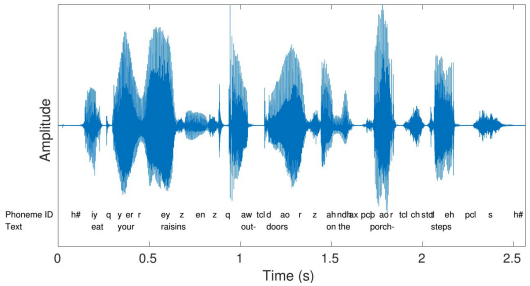
- Log
- Apply more functions… (IDFT for MFCC etc.)

Wav2Vec



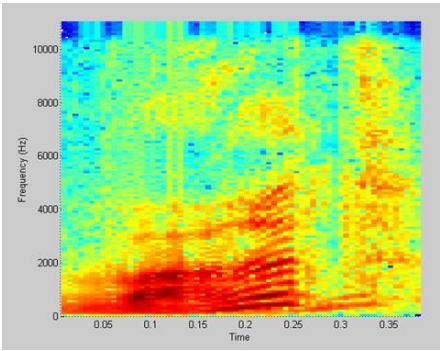
https://arxiv.org/pdf/1904.05862.pdf

TL; DR



Sound

MAGIC BLACK BOX



Image

Other “small” tasks to be mentioned:

- VAD (Voice Activity Detector)
- EOU (End of Utterance)
- Spotter
  - Phrase spotter (OK, Google! etc.)
  - Action spotter (a gunshot, a car passing by, etc.)
- Biometry-похожие:
  - Identification (which one of the preset N speakers was that?)
  - Verification (is speaker from the “permitted” set or not?)
  - Speaker features extraction (M-F, child vs adult, age, etc.)
- Acoustic Scene Classification