Biometrics (CSE 40537/60537)

Lecture 7: Speaker recognition

Adam Czajka

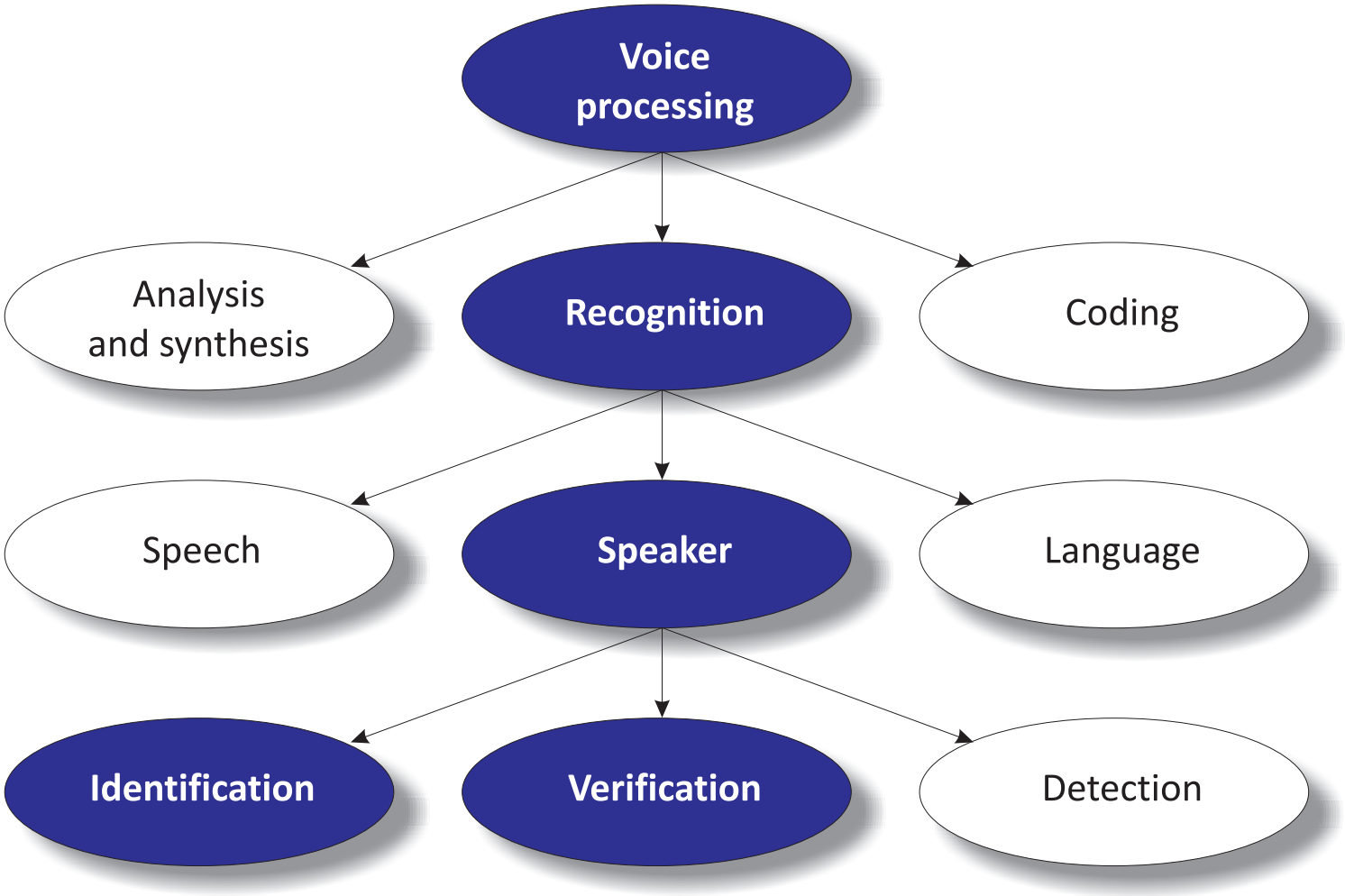
Biometrics and Machine Learning Group
Warsaw University of Technology, Poland

Fall 2014
University of Notre Dame, IN, USA

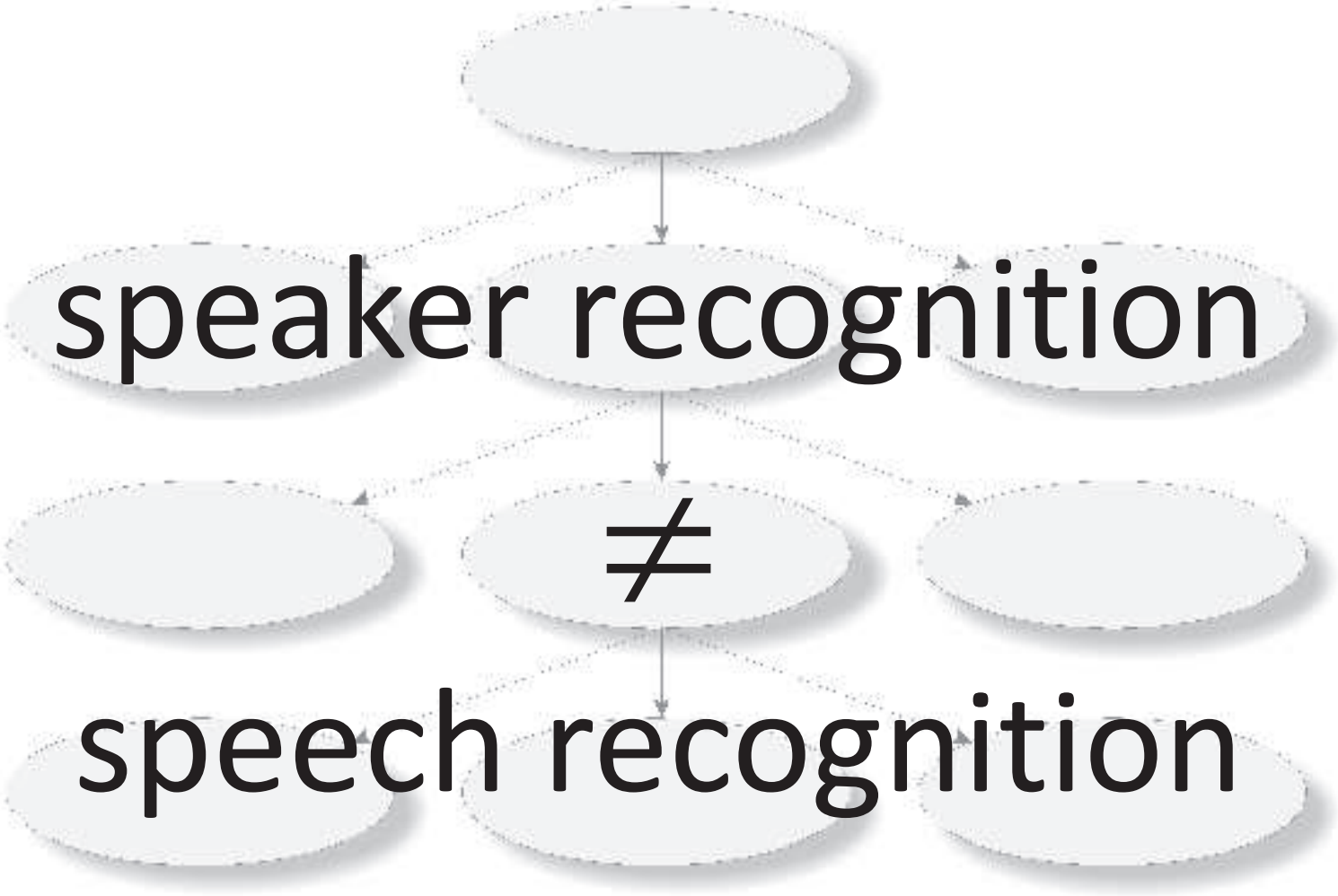
Lecture 7: Speaker recognition

- Speaker recognition as a part of voice processing
- Pre-processing of voice signals
- Extraction and classification of speaker features

Processing of voice signals



Processing of voice signals



Short history

1. 1960, Gunnar Fant, Sweden
   - first model of speech production (based on X-Ray images)
2. 1970, Joseph Perkell, MIT, USA
   - expanding the Fant's models: use of X-Ray movies
3. 1967–1985, IBM Research Laboratory, San Jose, USA
   - TASS-II/III (1961-67/1967–70) double sounds
   - TASS-IV (1980–1985) speech synthesis

HAL9000 from Stanley Kubrick movie *2001: A Space Odyssey*
source: wordpress.com

Short history

4. 1977, Texas Instruments, MITRE, US Air Force, USA
   - prototype of the first speaker recognition system
   - tests with 209 subjects
5. 1977, Matsimi Suzuki, Fuji Xerox, Japan
   - first patent of automatic speaker recognition method

Speaker recognition variants

1. Fixed-text
   - enrollment and authentication realized with the same, fixed text
   - the text may be simultaneously a password
   - easy to be forged by presenting a recorded phrase
2. Text-dependent
   - authentication based on *vocabulary* used to generate *authentication phrases*, e.g. 0-9 digits (the vocabulary) used to pronounce the numbers (the authentication phrase)
   - possible to be forged recording a vocabulary elements and presenting them in the expected order

Speaker recognition variants

3. Text-independent or Unconstrained-phrase
   - text phrase is selected by the user
   - hard to be forged since we would need a synthesizer mimicking one's voice
4. Conversational (under research)
   - hiding a message within the conversation; we need an analysis of semantics
   - merge of speaker and speech recognition
   - almost impossible to be forged since we would need a synthesizer mimicking one's voice AND the knowledge of a hidden message

1. Filtering

1. Finite Impulse Response (FIR) filters

$$s'_n = \sum_{k=0}^{N-1} a_k s_{n-k}, \quad n = 0, \ldots, N-1$$

2. Typically first order FIR is used: $a_0 = 1$, $a_1 \in \langle -1, -0.9 \rangle$, $a_n = 0$ for $n > 1$

2. Voice detection and division into segments

1. Detection of *silence-voice-silence* moments
2. Division into L segments, each of the length of K points; segments may overlap

$$s''_{k;l} = s'_{k+Ml}, \quad k = 0, \ldots, K-1, \quad l = 0, \ldots, L-1$$

where $M = K$ when segments do not overlap, or $M \neq K$ otherwise.

3. Minimizing the discontinuities

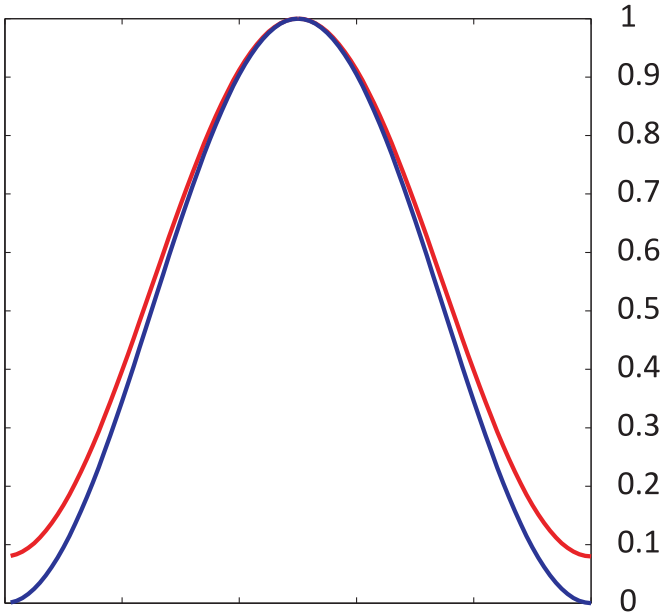
Use of weighting windows:

$$s''_{k;l} = s''_{k;l} w_k, \quad l = 0, \ldots, L-1$$

where

$$w_k = \alpha - (1-\alpha)\cos\left(\frac{2\pi k}{K}\right)$$

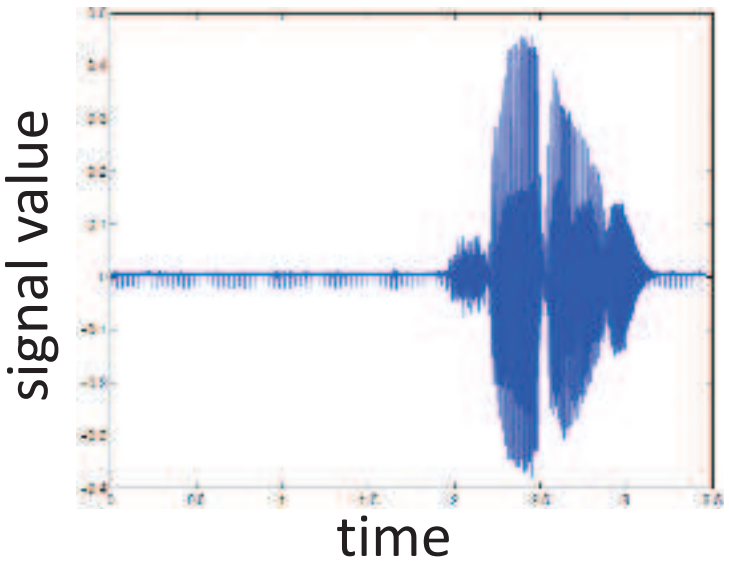
is a family of window functions and $\alpha \in (0, 1)$

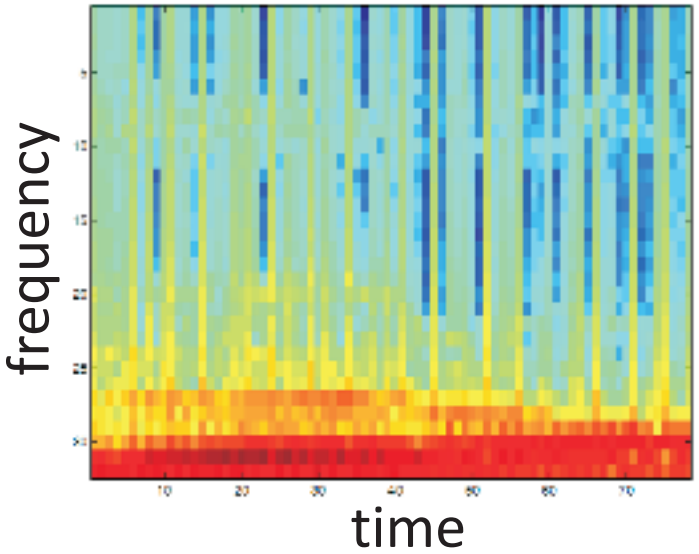


$\alpha = 0.5$: Hanning window
$\alpha = 0.54$: Hamming window

4. Representation of voice signals

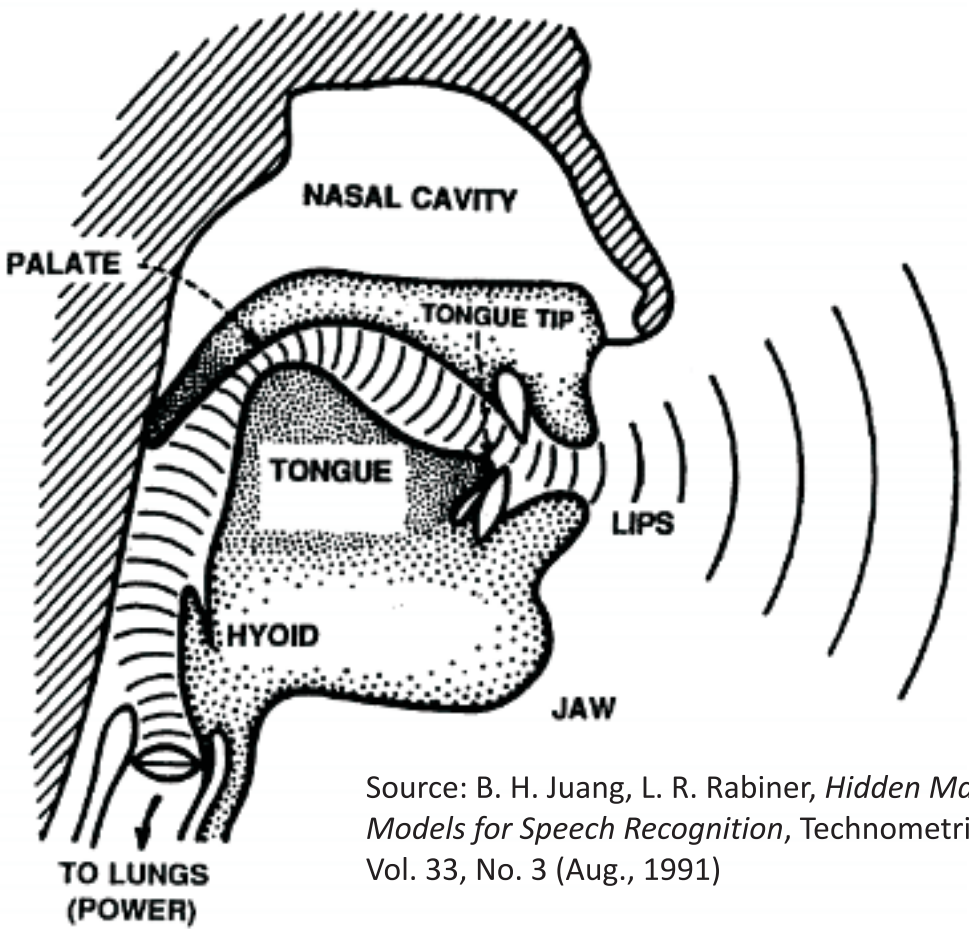
1. Time domain (graph of the signal's value, or signal's energy)
2. Frequency domain (frequency spectrum)
3. Time-frequency domain (spectrogram, a.k.a. voiceprint, voicegram, spectral waterfall, etc.)





Recording of the word 'ZERO'. Source: W. Kasprzak, *et al.*, 'Applying Independent Component Analysis for Speech Feature Detection', IWSSIP, 2004

Speaker features



Source: B. H. Juang, L. R. Rabiner, *Hidden Markov Models for Speech Recognition*, Technometrics, Vol. 33, No. 3 (Aug., 1991)

Simplified outline of the human vocal organ

Formants: characteristic quantities of the sound spectrum

Interesting experiment

Alvin Lucier, "I am sitting in a room", 1970

I am sitting in a room different from the one you are in now. I am recording the sound of my speaking voice and I am going to play it back into the room again and again until the resonant frequencies of the room reinforce themselves so that any semblance of my speech, with perhaps the exception of rhythm, is destroyed.

What you will hear, then, are the natural resonant frequencies of the room articulated by speech. I regard this activity not so much as a demonstration of a physical fact, but more as a way to smooth out any irregularities my speech might have.

Feature extraction in time domain

1. Auto-regressive models
   - Building a model: expressing the present voice sample as a linear combination of past samples

$$\hat{s}_n = \sum_{m=1}^{M} a_m s_{n-m} + e_n$$

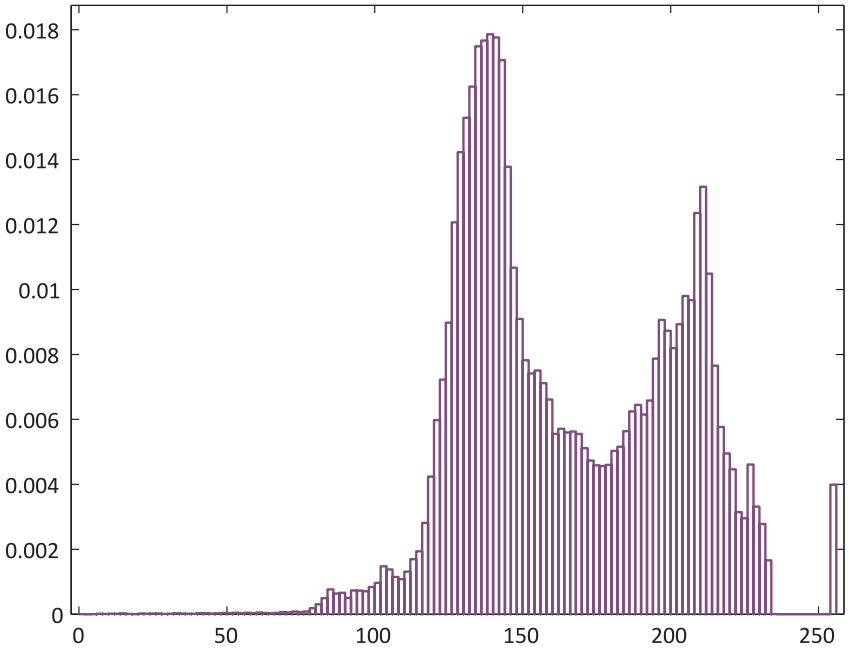
     where M – model order, e – noise with constant variance (here: representing the stimulation signal, i.e. the sound of vocal cords)
   - Features: a_m coefficients
   - Feature extraction (i.e. model identification): minimizing the mean squared error between s and $\hat{s}$

Feature extraction in time domain

2. Independent Component Analysis (ICA)
   - Assumption: voice is a linear superposition of statistically independent sound 'sources'
   - Task: find 'sources' and the way how they are superposed

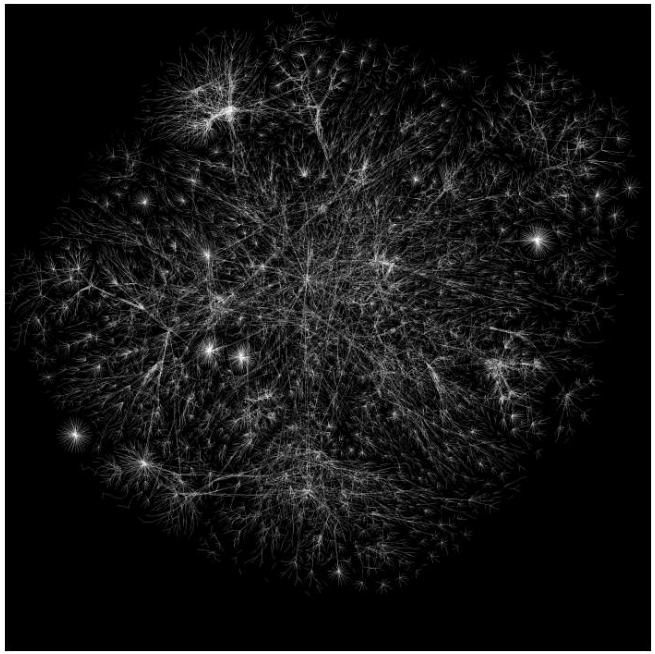
Independent Component Analysis

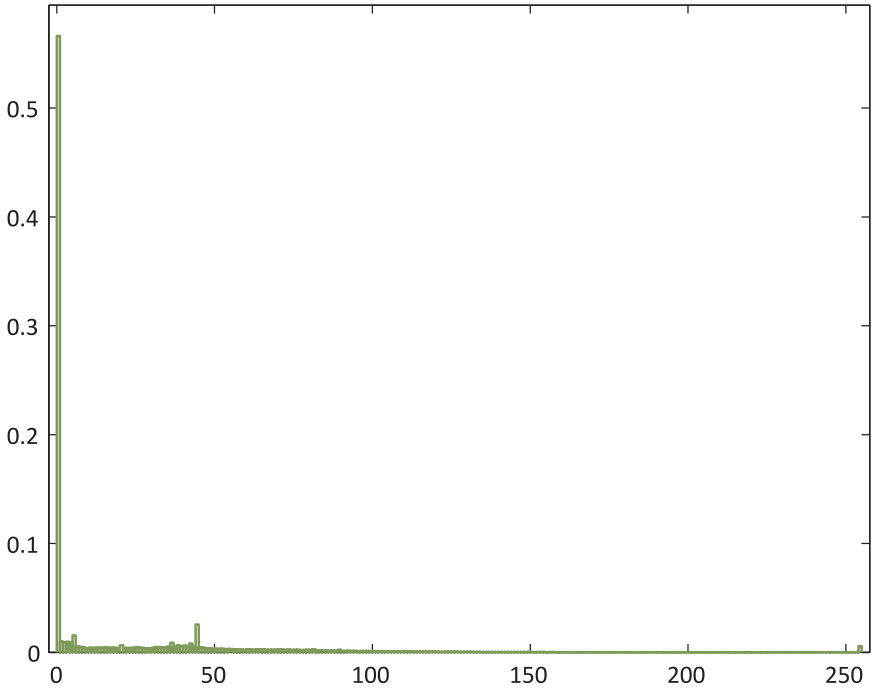
Example: superposition of images



Independent Component Analysis

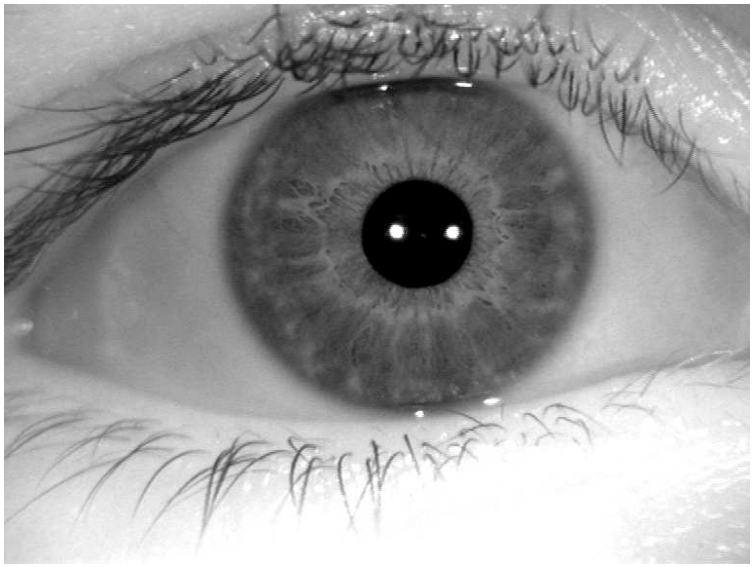
Example: superposition of images

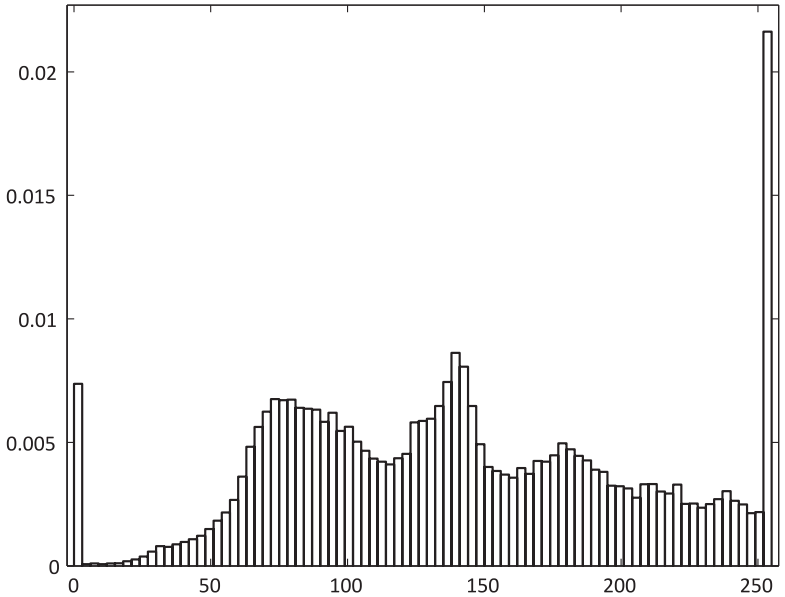




Independent Component Analysis

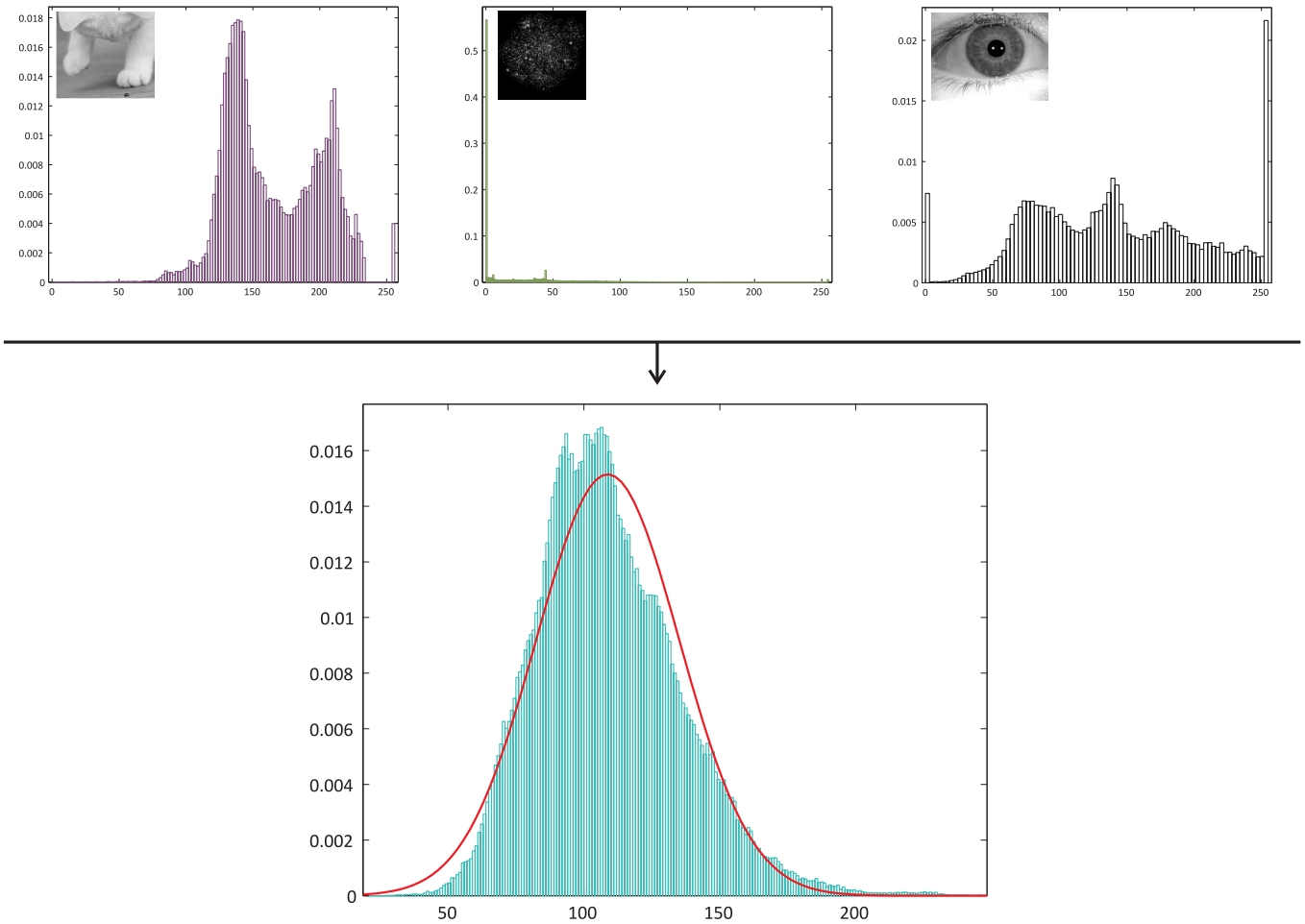
Example: superposition of images





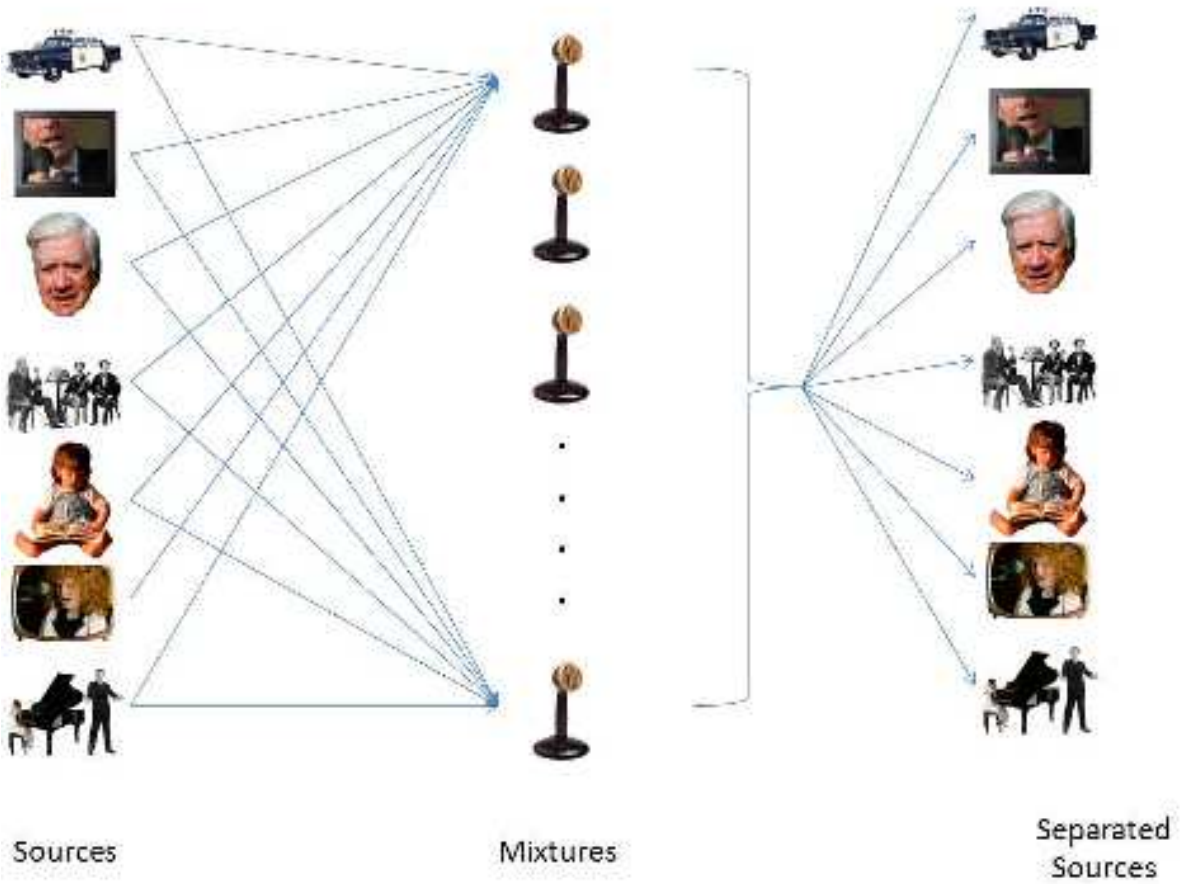
Independent Component Analysis

Example: superposition of images



Independent Component Analysis

Example: cocktail party problem



Source of the example: Helsinki Technical University, Finland, `http://research.ics.tkk.fi`

Independent Component Analysis

Mathematical background

- ICA modeling

$$\mathbf{y}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{e}(t)$$

where

$$\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$$

is a vector of m statistically independent and unknown sources, and

$$\mathbf{y}(t) = [y_1(t), \ldots, y_n(t)]^T$$

is a vector of n observations (linear superposition of sources s), and $\mathbf{e}(t)$ is a Gaussian noise

Independent Component Analysis

Mathematical background

- Definition of our task:
  - find s (and A) given (only) y and m
- Assumptions:
  - non-Gaussian sources, or one Gaussian source in the simplified model (with no $\mathbf{e}$(t) in the model)
  - typically $n \geqslant m$ (number of independent observations not less than the number of independent sources)
  - known variance of the sources, e.g. equal to 1; this resolves ambiguity since we are looking for s and $\mathbf{A}$ simultaneously

Independent Component Analysis

Interpretation in speaker recognition

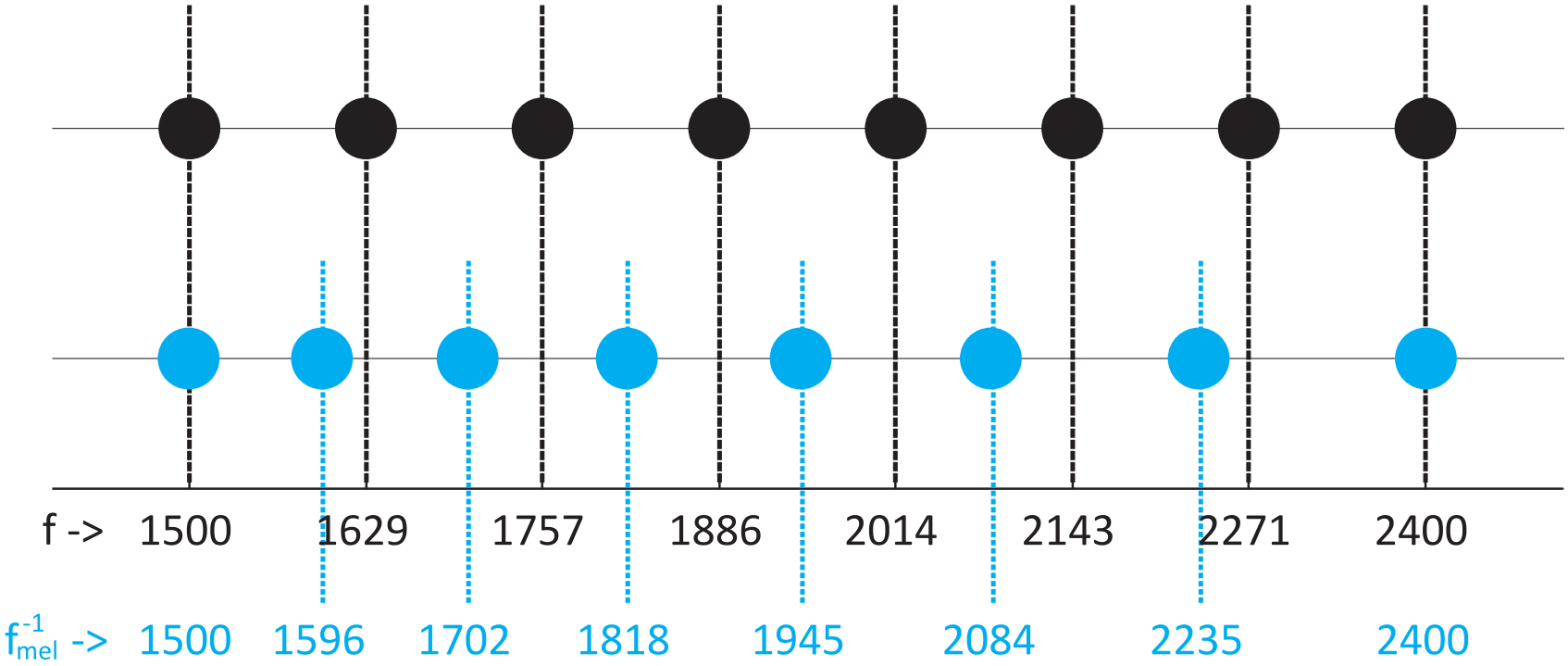
1. Sources are 'virtual' and dispersed in the vocal tract (imagine a number of small loudspeakers spread in the vocal tract)
2. We have to assume the number of 'sources'
3. Observations = different biometric samples of the same person
4. Biometric features: the linear superposition coefficients ($\mathbf{A}$ matrix)

Feature extraction in frequency domain

1. Selected Fourier coefficients (use of triangle windows to select coefficients)
2. Power coefficients based on Fourier transform
3. Fourier coefficients (selected or all) expressed in mel-scale

Mel-scale: a simple experiment

Mel-scale: a simple experiment

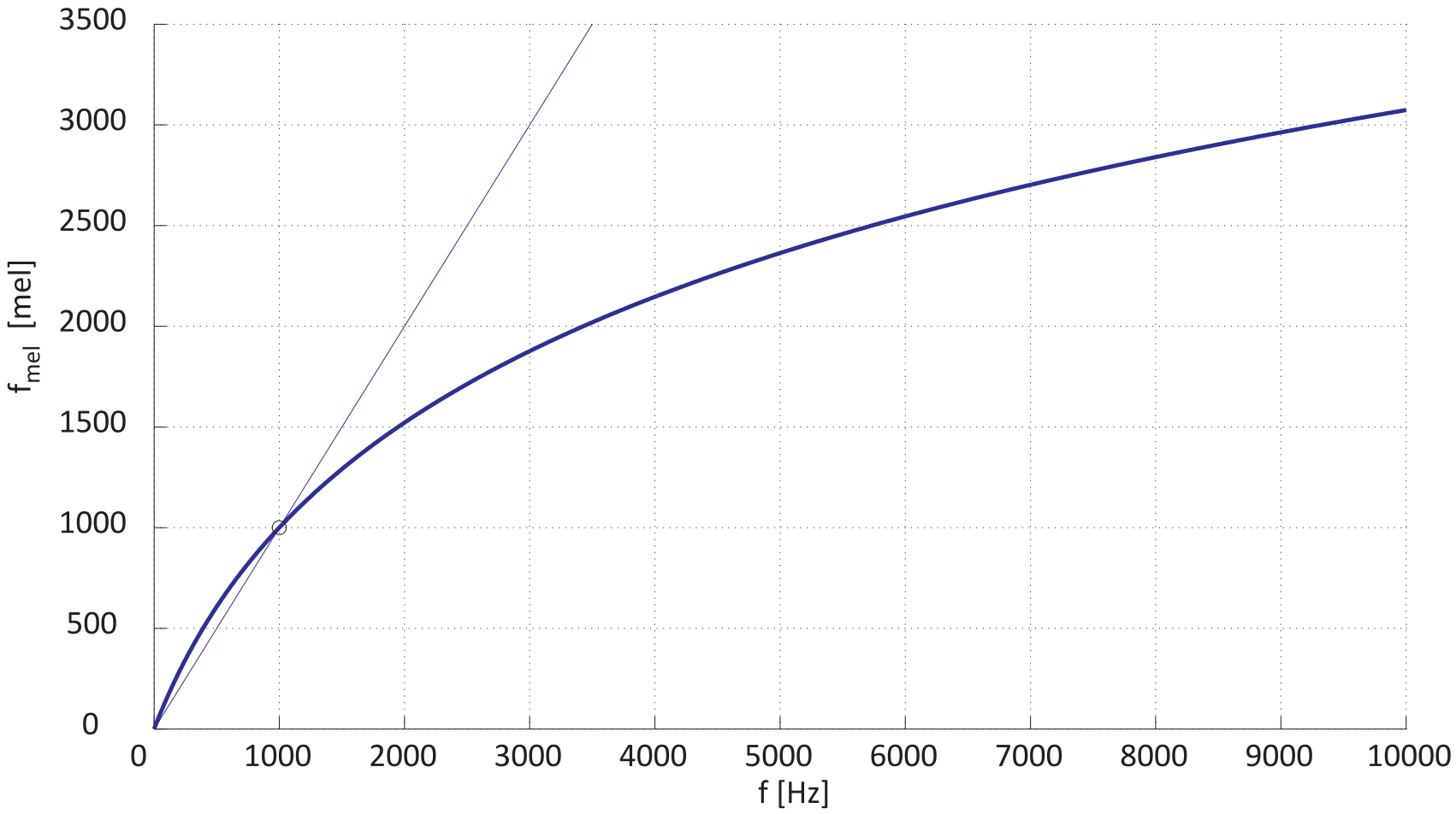


Mel-scale

1. Based on subjective estimation of distances between sound ('mel' like 'melody')
2. Nonlinear change of frequency related to the properties of human hearing organ, for instance: $f_{\mathrm{mel}} = 2595 \log(1 + f/700)$
3. It is commonly believed that mel-scale (not linear scale) is used by our brain to interpret sounds

Mel-scale

Example relation between mel-scale and linear scale



Feature extraction in cepstral domain

Homomorphic deconvolution

1. Assume that the observed voice signal y is a convolution of stimulus x (vocal cords) and impulse response h of the vocal track:

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k)$$

or, in frequency domain:

$$Y(\omega) = X(\omega)H(\omega)$$

Feature extraction in cepstral domain

Homomorphic deconvolution

2. Make a logarithm of both sides of the equation and calculate inverse Fourier transform, namely:

$$\mathrm{DFT}^{-1}\Big(\log_{10}\left(Y(\omega)\right)\Big) =$$

$$\mathrm{DFT}^{-1}\Big(\log_{10}\left(X(\omega)H(\omega)\right)\Big) =$$

$$\mathrm{DFT}^{-1}\Big(\log_{10}\left(X(\omega)\right) + \log_{10}\left(H(\omega)\right)\Big)$$

$$\mathrm{DFT}^{-1}\Big(\log_{10}\left(X(\omega)\right)\Big) + \mathrm{DFT}^{-1}\Big(\log_{10}\left(H(\omega)\right)\Big)$$

Feature extraction in cepstral domain

Homomorphic deconvolution

3. Cepstrum (reversing the first four letters in 'spectrum')
   - real cepstrum (derived from power spectrum)

$$RC(y) = \mathrm{DFT}^{-1}\Big(\log_{10}\big|\mathrm{DFT}(y)\big|\Big)$$

   - complex cepstrum

$$CC(y) = \mathrm{DFT}^{-1}\Big(\log_{10}\big(\mathrm{DFT}(y)\big)\Big)$$

   - complex mel-cepstrum

$$MFCC(y) = \mathrm{DFT}^{-1}\Big(\log_{10}\Big(\mathrm{MF}\big(\mathrm{DFT}(y)\big)\Big)\Big)$$

where MF denotes the transformation to mel-frequency scale

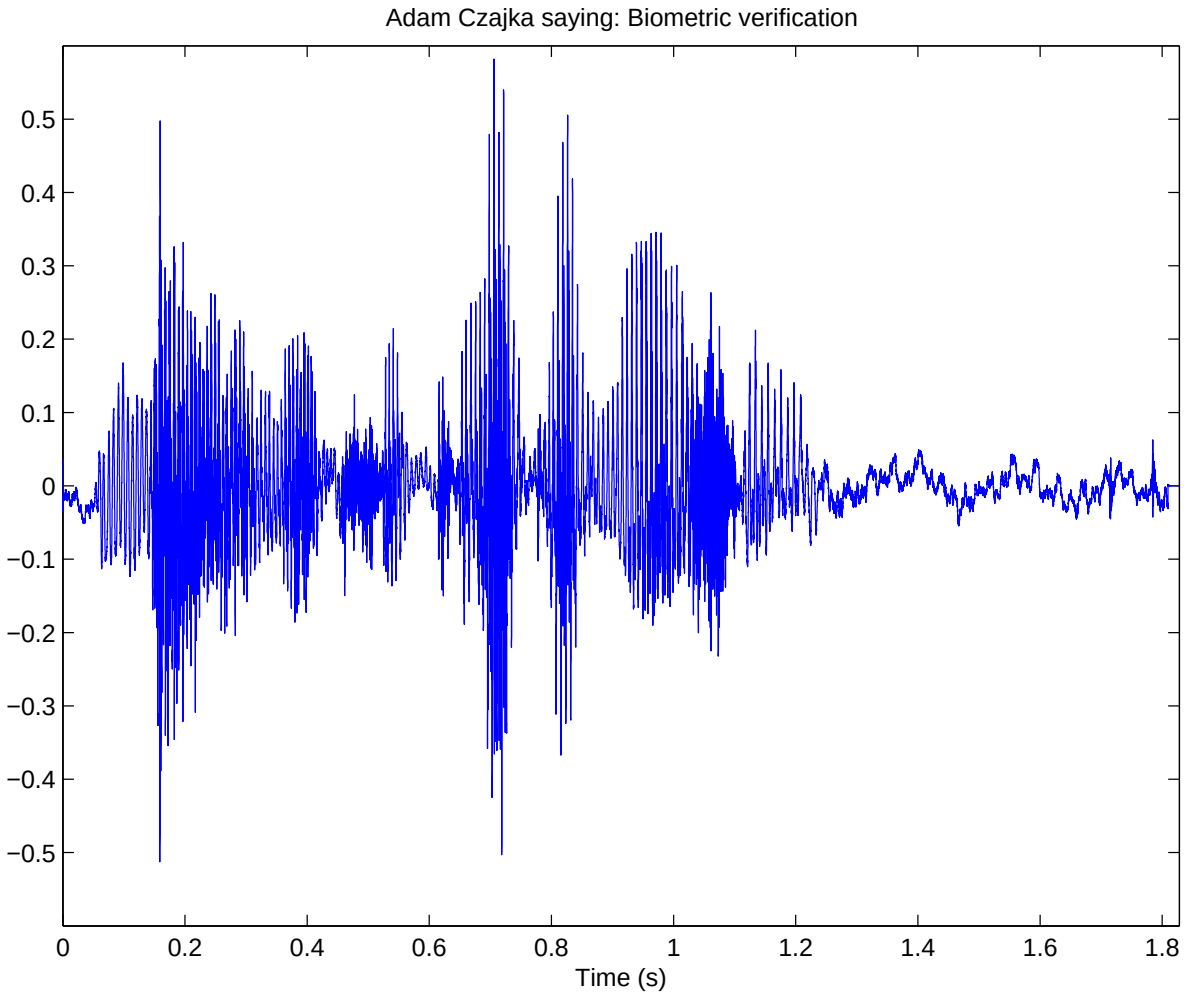
Feature extraction in cepstral domain

Features and comparison

4. Feature extraction
   - information about the vocal tract is concentrated at the beginning of the cepstrum, hence we can make windowing of the cepstral representation (e.g. using triangular windows) to separate interesting signals
   - typical features: *Cepstral Coefficients* (CC) or *Mel-Frequency Cepstral Coefficients* (MFCC)
5. Comparison of cepstral features: weighted Euclidean distance
6. Optional follow-up with cepstral representation:
   - calculation of the spectrum (e.g. DFT) and exponentiation (inverting the logarithm) separately for the stimulus CC and vocal track CC $\Rightarrow$ stimulus spectrum + vocal track spectrum

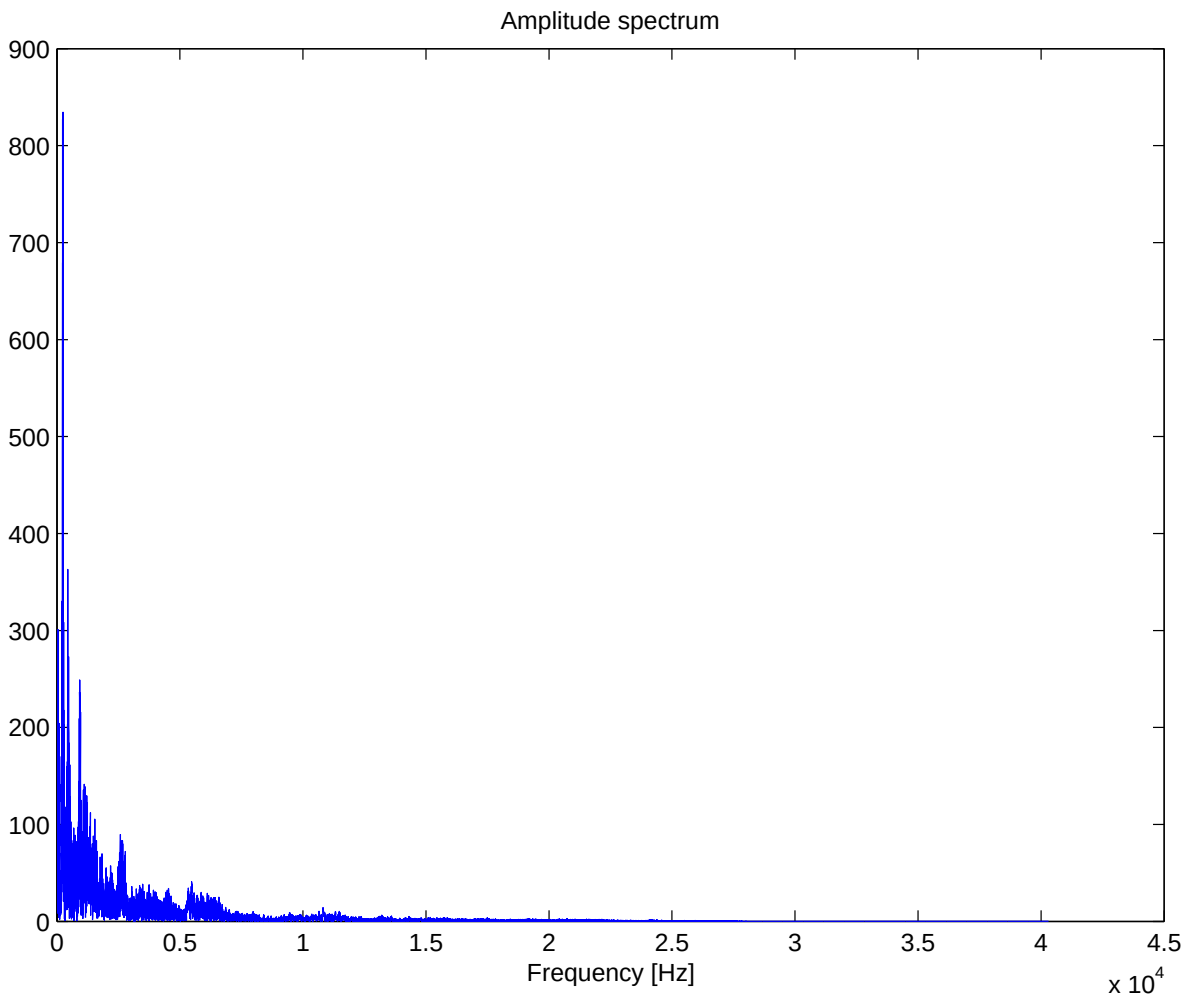
Feature extraction in cepstral domain

Example: 'Biometric verification' said by Adam Czajka



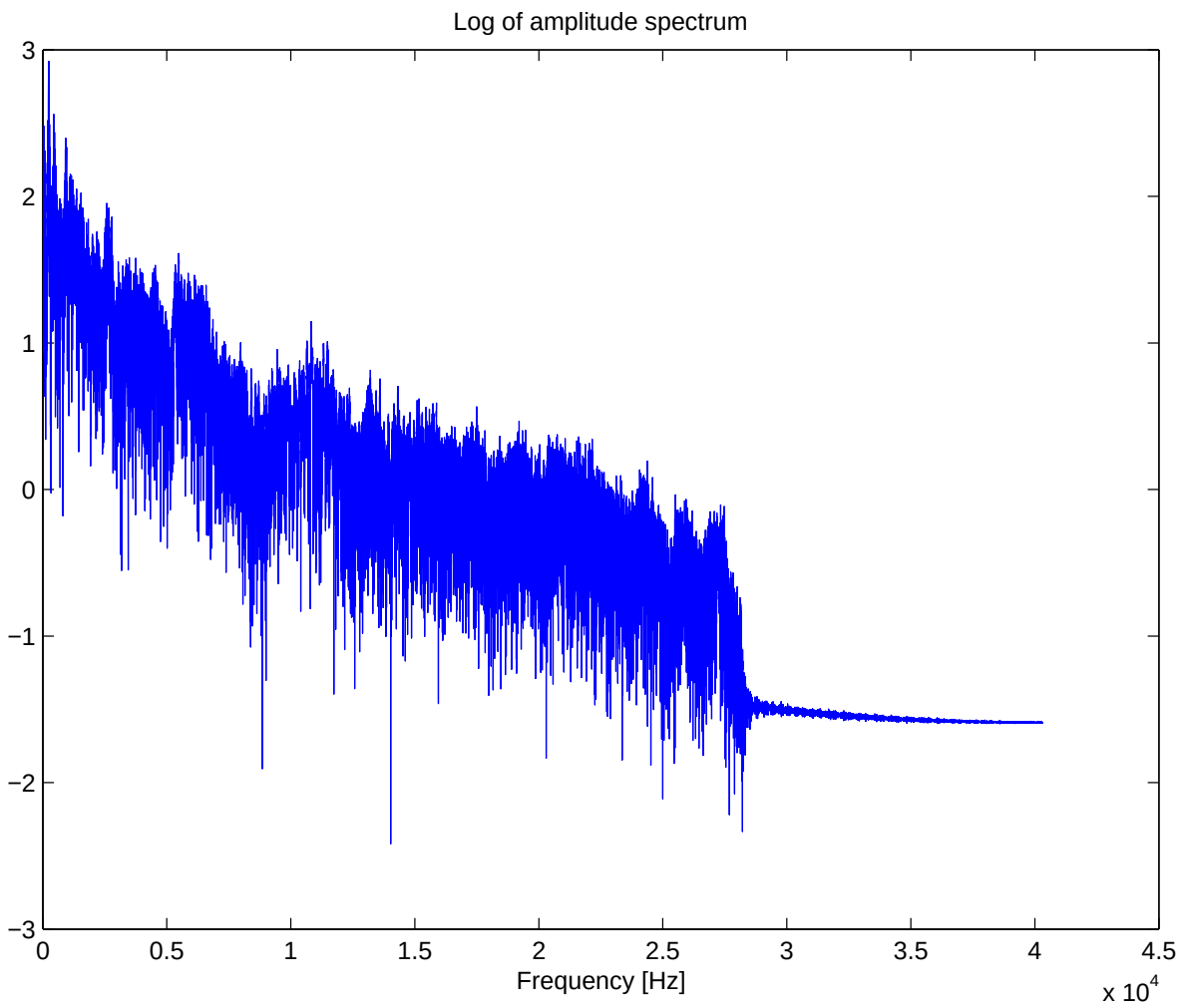
Feature extraction in cepstral domain

Example: 'Biometric verification' said by Adam Czajka



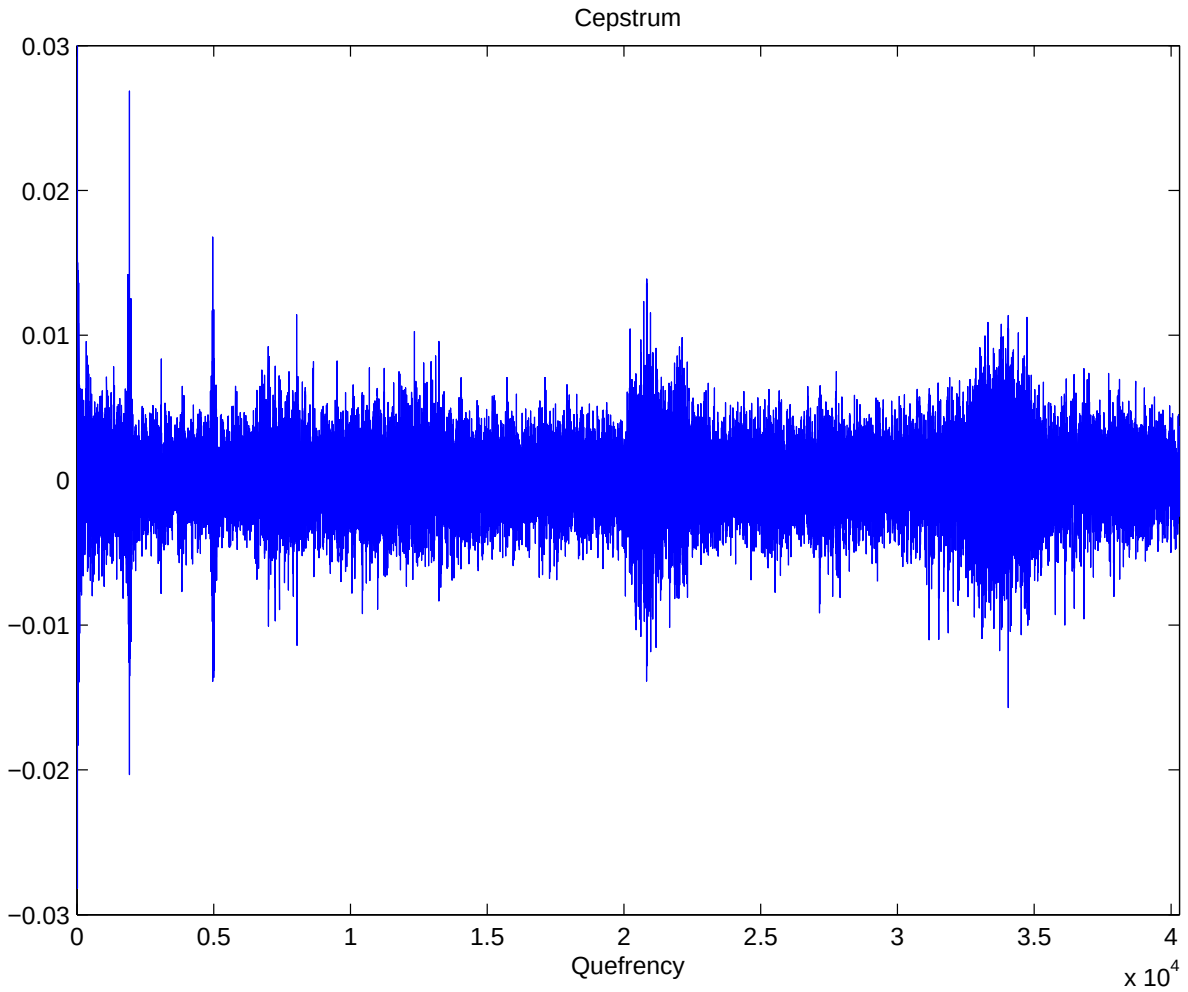
Feature extraction in cepstral domain

Example: 'Biometric verification' said by Adam Czajka



Feature extraction in cepstral domain

Example: 'Biometric verification' said by Adam Czajka



Feature extraction in cepstral domain

Summary of all steps