Audio-based Retrieval of Medical Information

(Teaching Inspired by Research)

Prof. Dr. Marcin Grzegorzek and
the Medical Data Science Team

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

Contents of the Course

Week	Lecture	Practical Exercises
1	**(05/04)** Introduction to Medical Information Retrieval (MIR)	**(05/04)** Introduction to Python
2	**(12/04)** Main Components and Classification of MIR Systems	**(12/04)** Introduction to Python
3	**(19/04)** Metadata in Medical Information Retrieval Systems	**(19/04)** CBIR in Medical Applications
4	***(26/04)*** *No Lecture due to a Business Trip*	**(26/04)** CBIR in Medical Applications
5	**(03/05)** Set Theoretic Model: Boolean Retrieval	**(03/05)** CBIR in Medical Applications
6	**(10/05)** Set Theoretic Model: Fuzzy Retrieval	**(10/05)** Flask Tutorial
7	**(17/05)** Vector Space Model: Similarity Measures	**(17/05)** Flask Tutorial

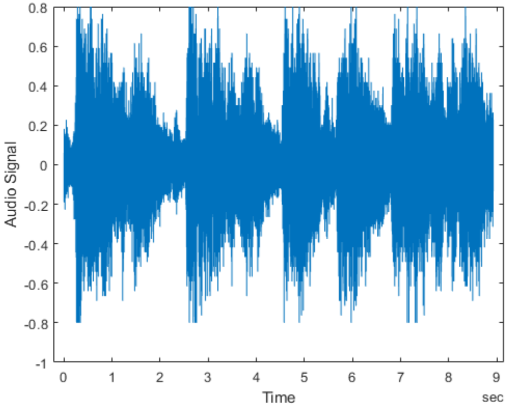
Contents of the Course

8	**(24/05)** Vector Space Model: Distance Functions	**(24/05)** HTML
9	**(31/05)** Vector Space Model: Latent Semantic Indexing	**(31/05)** HTML
10	**(07/06)** Probabilistic Model	**(07/06)** HTML
11	**(14/06)** Text-based Retrieval of Medical Information	**(14/06)** Deep Learning
12	**(21/06)** Audio-based Retrieval of Medical Information	**(21/06)** Deep Learning
13	**(28/06)** Image-based Retrieval of Medical Information	**(28/06)** Relevance Feedback
14	**(05/07)** Demonstrators from Current Research Projects	**(05/07)** Relevance Feedback
15	**(12/07)** Summary and Conclusions	**(12/07)** Evaluation

Automatic Audio Analysis – Applications

- Speech recognition systems
- Audiovisual data segmentation and indexing
- Content-based retrieval from music databases (e.g., querying by humming)
- Automatic music genre classification
- Etc.

Feature Extraction – Problem Statement



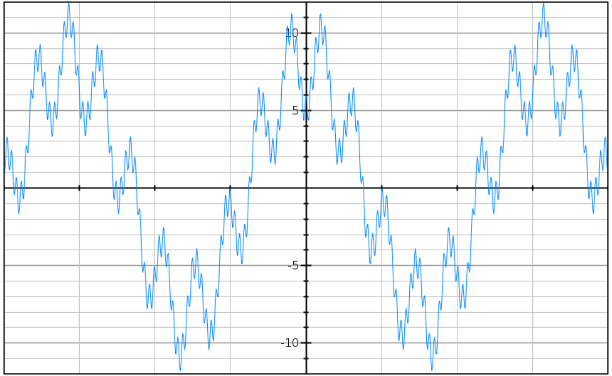
$$\Downarrow$$

$$\boldsymbol{x} = (x_1, x_2, \dots, x_l)^{\mathrm{T}}$$

Feature Extraction in Frequency Domain

$$f(t) \approx x_1 \cos(\omega_1 t) + x_2 \cos(\omega_2 t) + \ldots + x_l \cos(\omega_l t); \quad \omega_i = 2\pi f_i$$

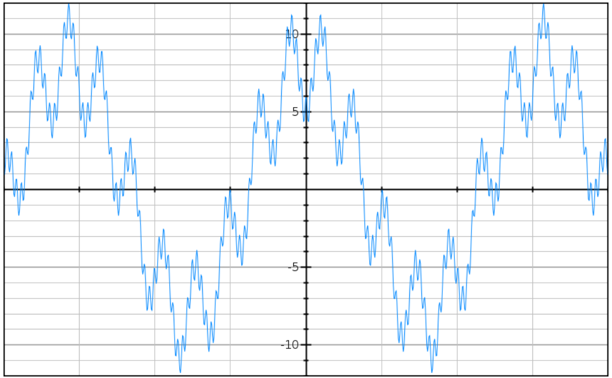
$$f(t) \longrightarrow \boldsymbol{x} = (x_1, x_2, \ldots, x_l)^{\mathrm{T}}$$



Feature Extraction in Frequency Domain

$$f(t) = 8\cos(2t) - 3\cos(15t) + \cos(100t)$$

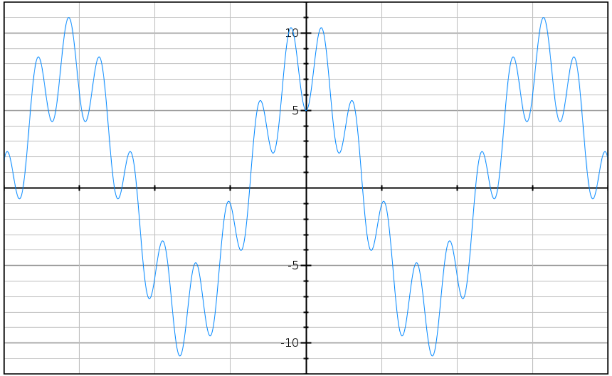
$$f(t) \longrightarrow \boldsymbol{x} = (8, -3, 1)^{\mathrm{T}}$$



Feature Extraction in Frequency Domain

$$f(t) = 8\cos(2t) - 3\cos(15t)$$

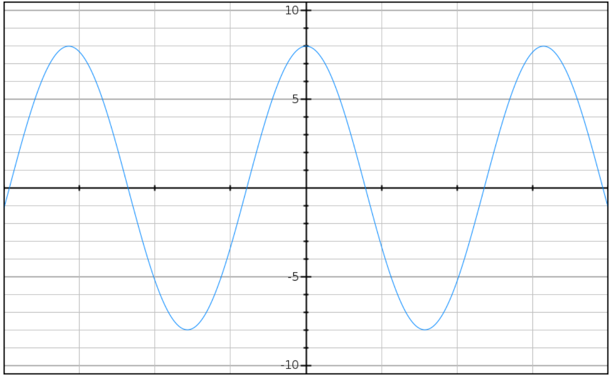
$$f(t) \longrightarrow \boldsymbol{x} = (8, -3)^{\mathrm{T}}$$



Feature Extraction in Frequency Domain

$$f(t) = 8\cos(2t)$$

$$f(t) \longrightarrow x = 8$$



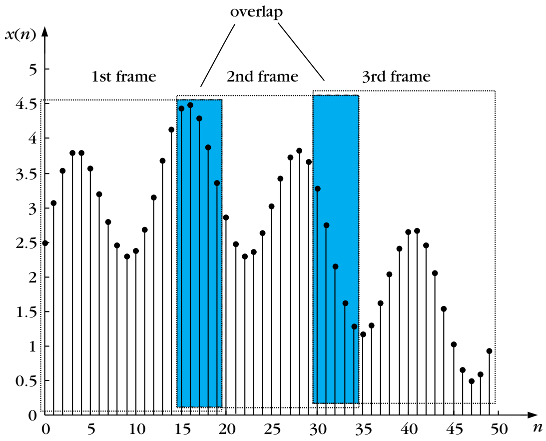
Fourier Transform – Visualisation

https://www.youtube.com/watch?v=spUNpyF58BY

Short Time Processing of Signals – Introductory Statements

- The statistical properties of the speech and audio signals vary with time (nonstationary signals).
- In order to use tools for stationary signals (e.g., Fourier transform), the signal is divided to a series of successive frames.
- Each frame consists of a finite number of N samples.
- During the time interval of a frame, the signal is assumed to be “reasonably stationary” (quasistationary, see Figure on the next slide).

Short Time Processing of Signals - Quasistationary Frames



Three successive frames, each of length $N = 20$ samples. The overlap between successive frames is 5 samples.

Short Time Processing of Signals – Choosing Parameters

- Choosing the length N is a problem-dependent task.
- On the one hand, N has to be high enough to include useful part of information. On the other hand, it has to be small for the stationary assumption.
- For speech signals sampled at a frequency of $f_s = 100$ kHz, reasonable frame sizes range from 100 to 200 samples, corresponding to 10-20 msecs duration.
- For music signals sampled at 44.1 kHz, reasonable frame sizes range from 2048 to 4096 samples, corresponding to 45-95 msecs.

Short Time Processing of Signals – Dividing into Frames

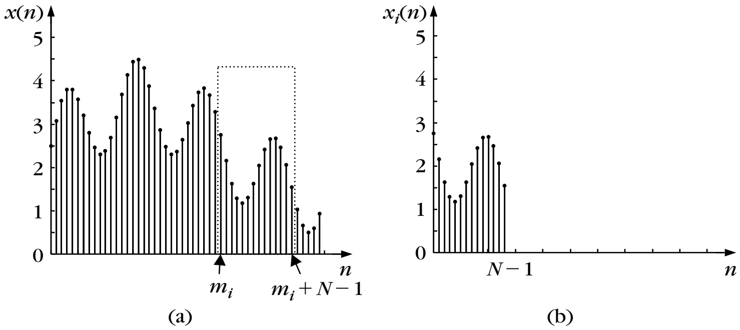
- Dividing the signal in a sequence of successive frames is equivalent to multiplying the signal segment by a window sequence $w(n)$ of a finite duration N

$$w(n) = \begin{cases} 1 & ; \quad \text{for} \quad 0 \leq n \leq N-1 \\ 0 & ; \quad \text{elsewhere} \end{cases} \quad .$$

- For different frames, the window is shifted to different points m_i on the time axis. Hence, if $x(n)$ denotes the signal sequence, the samples of the frame no. i can be written as

$$x_i(n) = x(n + m_i)w(n) \quad .$$

Short Time Processing of Signals – Dividing into Frames



A signal segment (a) and the resulting frame (b) after the application of a rectangular window sequence of duration equal to 14 samples and shifted at m_i.

Short Time Processing of Signals – Fourier Transform

- We divide a speech signal into a sequence of F frames, each of length N.

- Then, for each frame $x_i(n)$ we compute the DFT as

$$X_i(m) = \sum_{n=0}^{N-1} x_i(n) \exp\left(-j\frac{2\pi}{N}mn\right), \quad m = 0, \ldots, N-1 \quad .$$

Short Time Processing of Signals – Fourier Features

- Selecting $I \leq N$ DFT coefficients from each frame, we construct a sequence of feature vectors

$$\boldsymbol{x}_i = \begin{bmatrix} X_i(0) \\ \vdots \\ X_i(I) \end{bmatrix}, \quad i = 1, 2, \ldots, F \quad .$$

- Thus, the pattern of interest (e.g., a speech segment) is not represented by a single feature vector but by a sequence of feature vectors

$$\boldsymbol{x} \rightarrow (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_F) \quad .$$

Short Time Processing of Signals – Autocorrelation

- Another very important quantity defined for quasistationary processes is the short-time autocorrelation

$$r_i(k) = \frac{1}{N} \sum_{n=0}^{N-1-|k|} x_i(n)x(n+|k|) \quad .$$

- The limits in the sum indicate that outside the interval $[0, N-1-|k|]$ the product $x_i(n)x_i(n+|k|)$ is zero.

Spectral Features – Introduction

- Let $x_i(n), n = 0, \ldots, N-1$ be the samples of the frame no. i and $X_i(m), m = 0, \ldots, N-1$ the corresponding DFT coefficients.

- The following features are common in speech/audio recognition:
  - spectral centroid,
  - spectral roll-off,
  - spectral flux,
  - fundamental frequency.

Spectral Centroids

- Definition:

$$C(i) = \frac{\sum_{m=0}^{N-1} m|X_i(m)|}{\sum_{m=0}^{N-1} |X_i(m)|} \quad .$$

- The centroid is a measure of the spectral shape. High values of the centroid correspond to “brighter” acoustic structures with more energy in the high frequencies.

Spectral Roll-Off

- The spectral roll-off is the frequency sample $m_c^R(i)$ below which the c% of the magnitude distribution of the DFT coefficients is concentrated:

$$\sum_{m=0}^{m_c^R(i)} |X_i(m)| = \frac{c}{100} \sum_{m=0}^{N-1} |X_i(m)| \quad .$$

- This measure indicates where the most of the spectral energy is concentrated.

Spectral Flux

- Definition:

$$F(i) = \sum_{m=0}^{N-1} (N_i(m) - N_{i-1}(m))^2 \quad .$$

- $N_i(m)$ is the normalised (by its maximum value) magnitude of the respective DFT coefficient of the frame no. i

- Thus, $F(i)$ is a measure of the local spectral change between successive frames.

Final Statements

- The frequency analysis provides a great tool to represent and manipulate one-dimensional signals, e.g. audio.
- Purely signal-based techniques for audio representation are usually combined with statistical language modelling in order to close the semantic gap in speech recognition