Speech Recognition
Speech signal descriptors

Krzysztof Ślot

Instytut Informatyki Stosowanej,

Politechnika Łódzka

Politechnika Łódzka

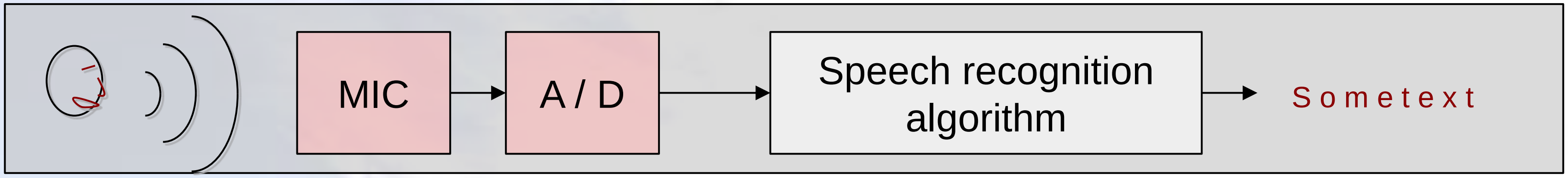
Instytut Informatyki Stosowanej

Course information

- **Schedule**
  - Meeting time: Wed 2pm-4pm, E110
  - Lectures/labs interleave – actual schedule tba
- **Course slides (supplementary material)**
  - Accessible from kslot.iis.p.lodz.pl under menu: Teaching - Courses in English – Speech Recognition
- **Assessment**
  - Test
  - Lab projects
- **Contact information**
  - kslot@p.lodz.pl
  - Office hours: Wednesdays, 3pm-4pm, building C7, room 6

Course outline

- **Subject**
  - Speech recognition by computers
    - Input: a sequence of samples representing sounds (electroacoustic conversion, A/D conversion will not be covered)
    - Output: phonetic transcription / words



- **Topics**
  - Speech signal: production and characterization (phone classes)
  - Speech signal representation (LPC, cepstrum, MFCC)
  - Sequence matching (Dynamic Time Warping)
  - Hidden Markov Models
  - Language modeling

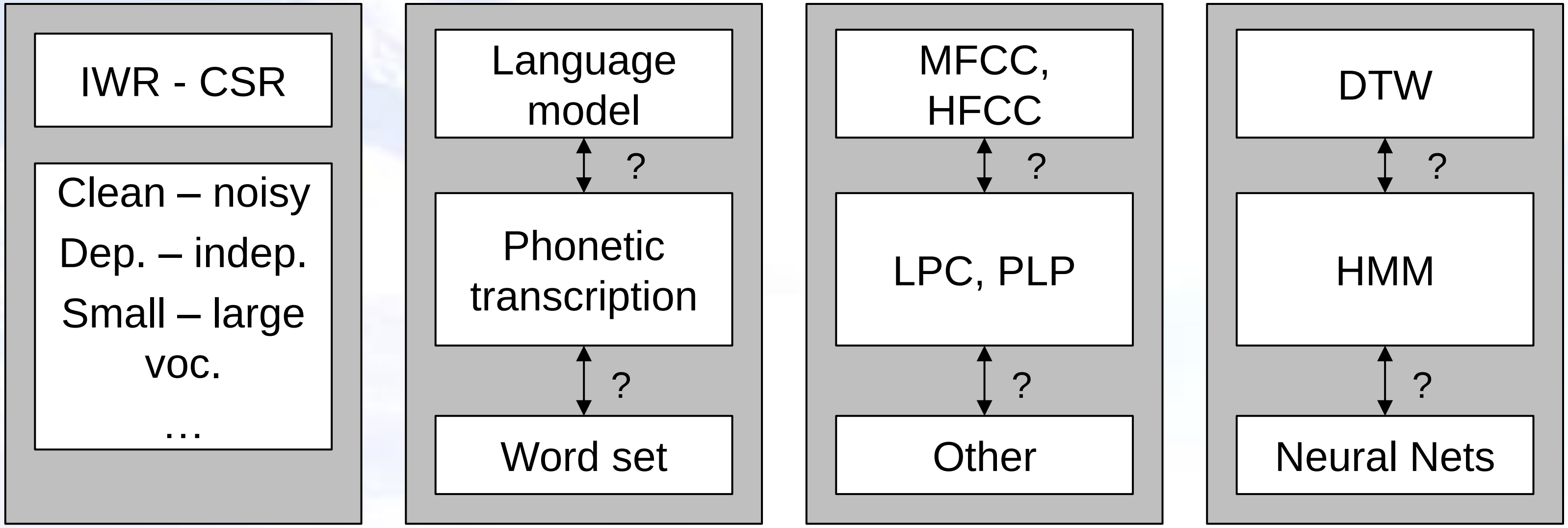
ASR preliminaries

- **Automatic Speech Recognition (ASR)**
  - Difficult task: patterns, variability, noise and distortions
  - Huge demand: natural human-computer interface, data analysis for retreivial, identification, understanding
  - Noticeable progress: operational continuous speech recognition (CSR) systems (for English), widespread isolated word recognition (IWR) applications
  - Data-driven problem solution (training)

- **ASR contexts**
  - Vocabulary size: limited-large (modeling and recognition strategies)
  - Universality: speaker-dependent vs.speaker independent systems
  - Target domain: domain-specific or unconstrained
  - Task definition: isolated word recognition or continuous speech
  - Input quality: noise and reverberations vs. clean speech
  - Modifying factors: emotions, ilness, age, gender
  - Language dependence

ASR preliminaries

- **General framework**
  - Define the context and objectives, get training data
  - Determine target categories and decide on class-modeling strategy
  - Derive appropriate quantitative representation of speech signal
  - Decide on recognition strategy
  - Build models for considered categories and train their parameters
  - Execute some adopted recognition scheme



ASR preliminaries

- Basic concepts
  - Sounds and their representation
  - Speech signal

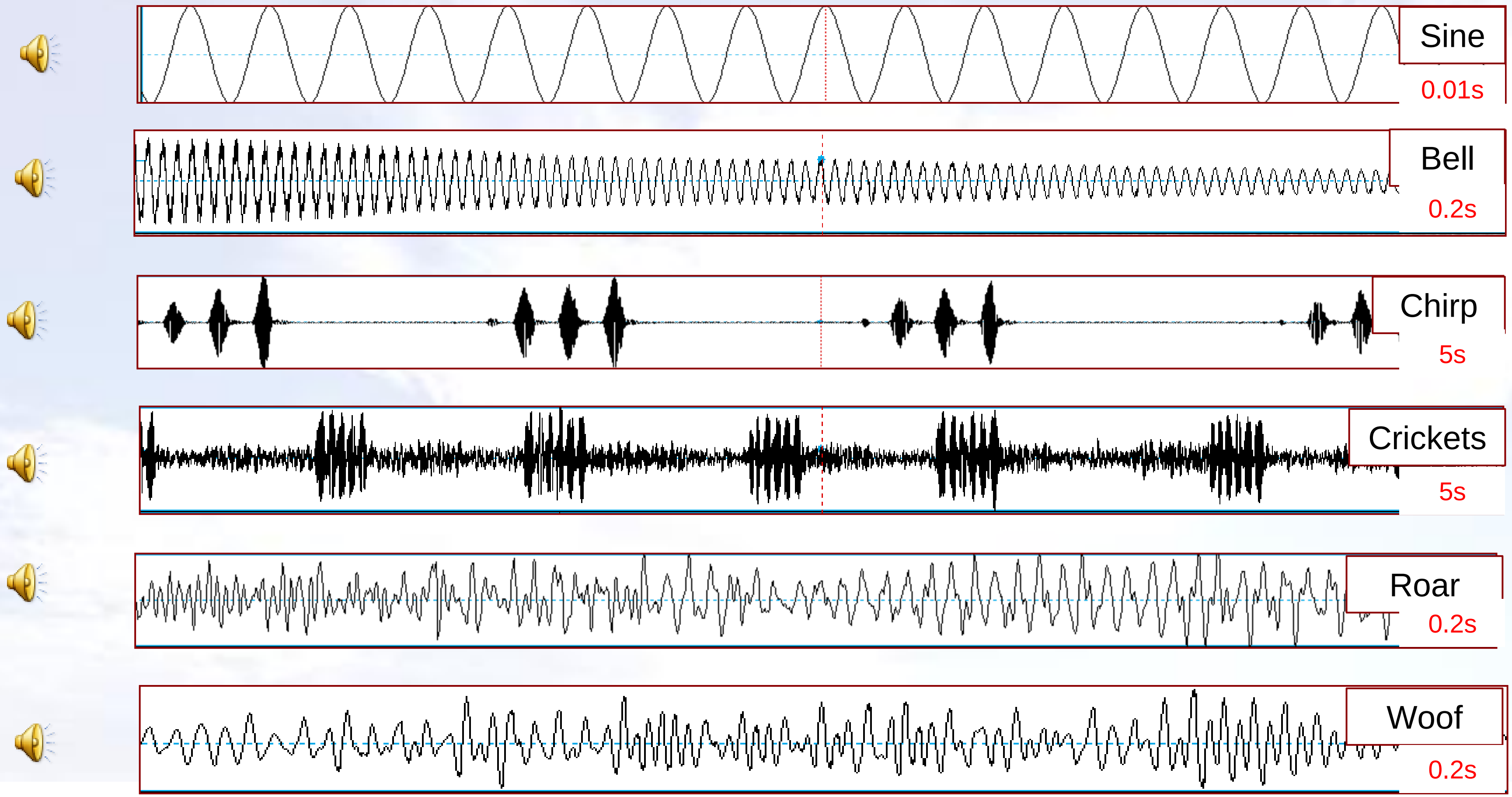
- Acknowledgements: examples were created using the following software
  - Praat
  - Htk
  - Matlab

Sounds

Instytut Informatyki Stosowanej

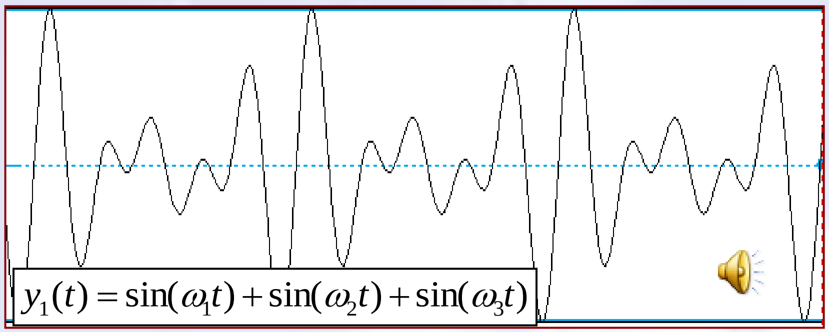
- **Sound perception and recognition**
  - Attributes: loudness (intensity), timbre (composition), variability, …
  - Recognition: requires quantitative description

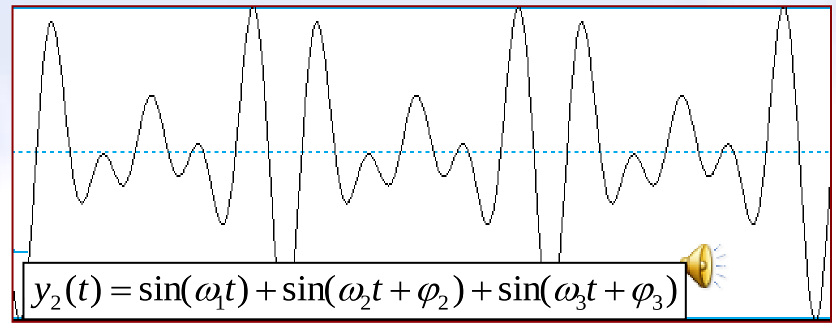
Temporal waveforms



Sound perception and representation

- **Representation as a function of time**
  - Phase-insenitivity: waveforms cannot be the right representation

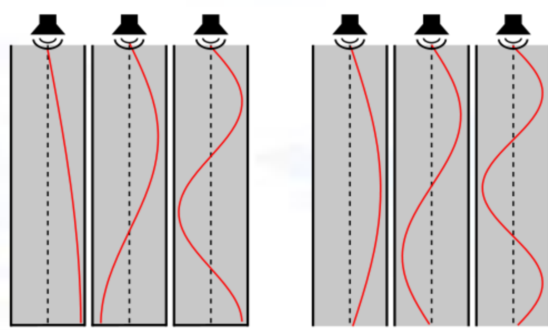




Different waveforms – the same perception

- **Representation in terms of harmonic components**
  - Physically-motivated representation of sounds: acoustic resonance

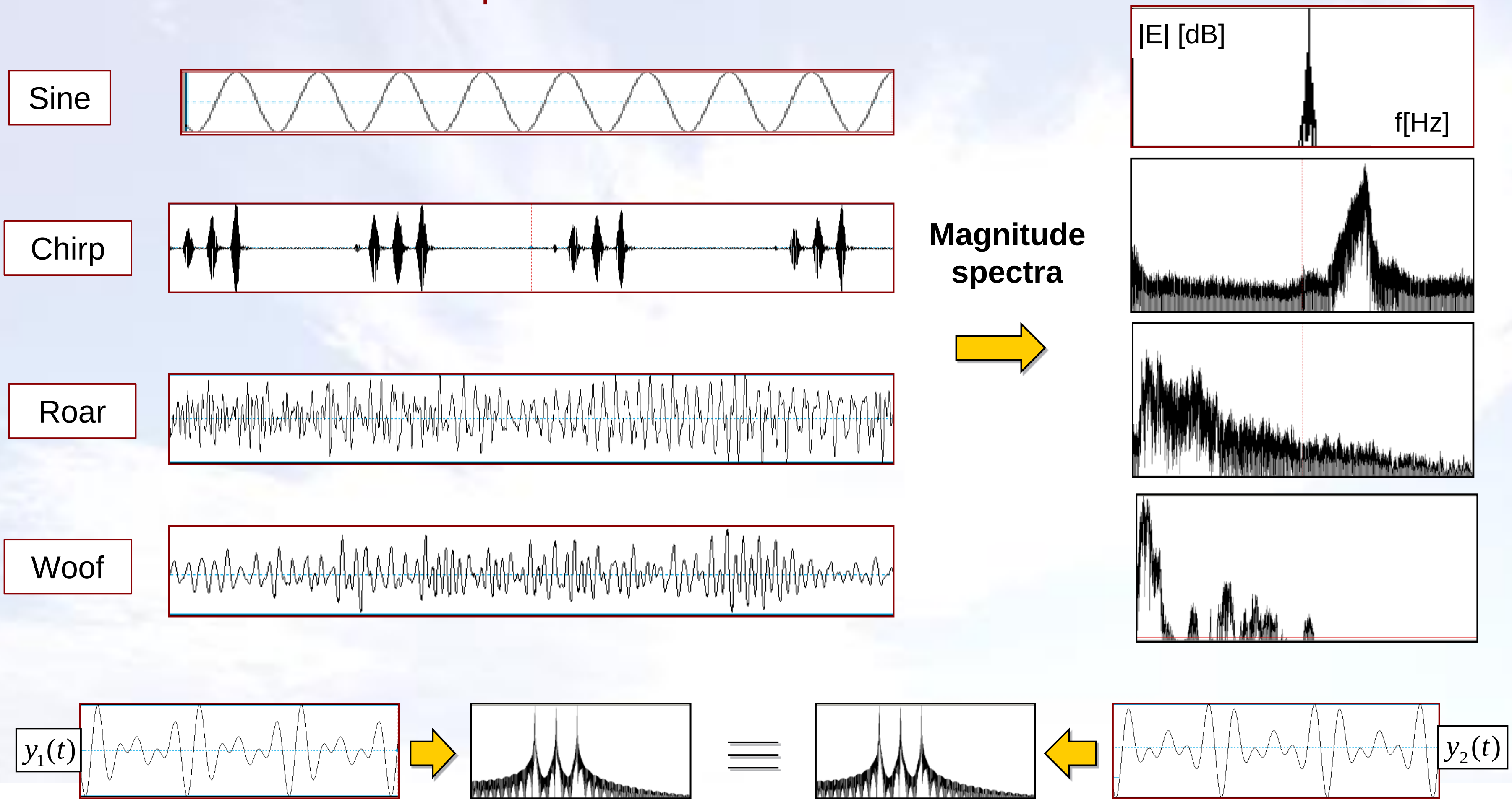
Some source is producing a collection of sine waves – only selected are emphasized and make a timbre of a sound



http://www.quora.com

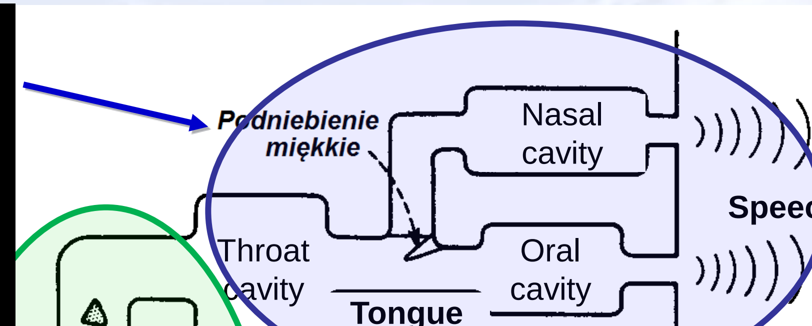
Sound perception and representation

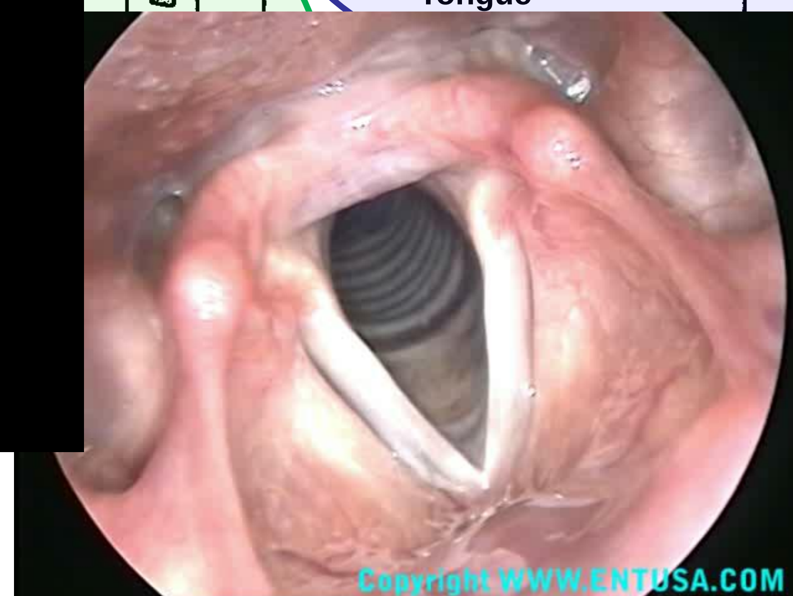
- **Representation: a collection of harmonic components**
  - Fourier Transform and spectra (details – later): decomposition of a signal into harmonic components



Speech production

- **By humans (and many other species)**
  - Generation of acosutic stimuli (phonation)
  - Forming the stimuli into differentiable sounds (articulation)

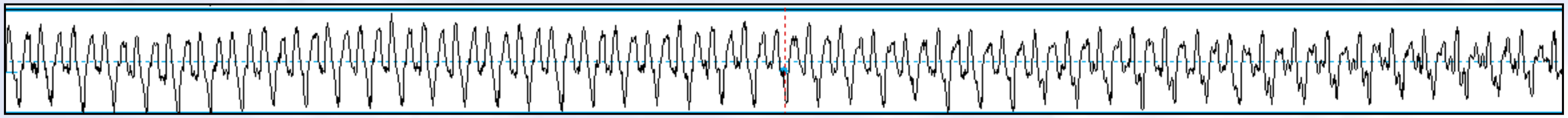




Speech production

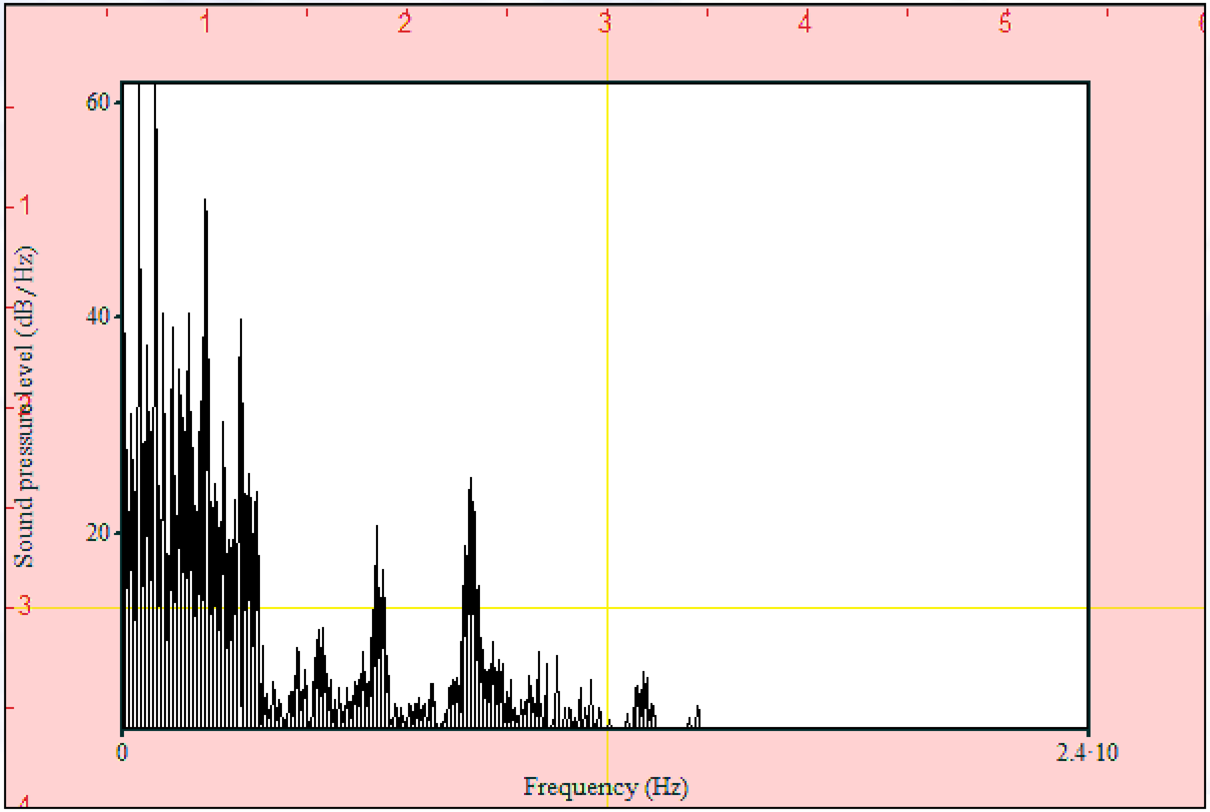
Instytut
Informatyki
Stosowanej

- **Phonation**
  - Vocal folds: vibrate (voiced) do not vibrate (voiceless)



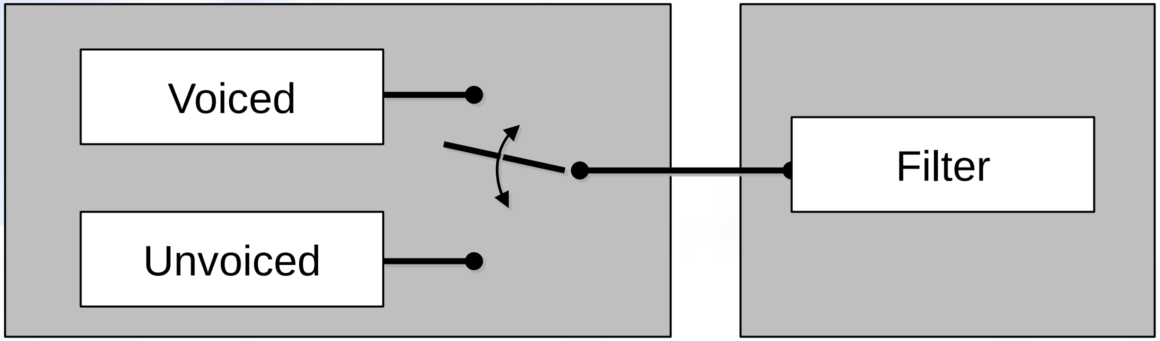
- **Voiced phonation**
  - Multi-harmonic
  - Fundamental frequency (pitch)
  - **Voiced speech**

- **Unvoiced phonation**
  - Vocal folds do not vibrate
  - Turbulent flow
  - Unvoiced speech



Speech production: source-filter model

- **Source**
  - Voiced excitation: vocal folds vibrate
  - Unvoiced excitation
- **Filter**
  - Resonant cavities of dynamically changing structure



Speech signal: acoustics

- **Phones and phonemes**
  - Phoneme: the smallest structural unit that distinguishes meaning in a language
  - Phone: a physical instance of phoneme

- **Phone types**
  - Vowels – flow of air is not impeded: a e o i u
  - Consonants – impeded or stopped
    - Nasals – formed by opening nasal cavity (m,n)
    - Fricatives – formed by impeding air-flow: may be voiced (v,z, δ as in **th**is) or voiceless (f,s,sh, θ as in **th**ought, h)
    - Stops – complete stopping then releasing air: voiced (b,d,g), voiceless (p,t,k)
    - Liquids – r,l
  - Semi-vowels: w as in wet, y as in yard

Phonetic transcription

- **Highlights**
  - Textual representation of sounds – of what we hear
  - Several standards, the most common: IPA (International Phonetic Alphabet)
  - There are more symbols than letters
  - **This is to be recognized in CSR**

ðɪs ɪz ðə fərst lɛktʃər

This is the first lecture

http://www.antimoon.com/how/pronunctransdemo.htm

IPA

vowels

IPA	examples
ʌ	cup, luck
ɑ:	arm, father
æ	cat, black
e	met, bed
ə	away, cinema
ɜ:ʳ	turn, learn
ɪ	hit, sitting
i:	see, heat
ɒ	hot, rock
ɔ:	call, four
ʊ	put, could
u:	blue, food
aɪ	five, eye
aʊ	now, out
eɪ	say, eight
oʊ	go, home
ɔɪ	boy, join
eəʳ	where, air
ɪəʳ	near, here
ʊəʳ	pure, tourist

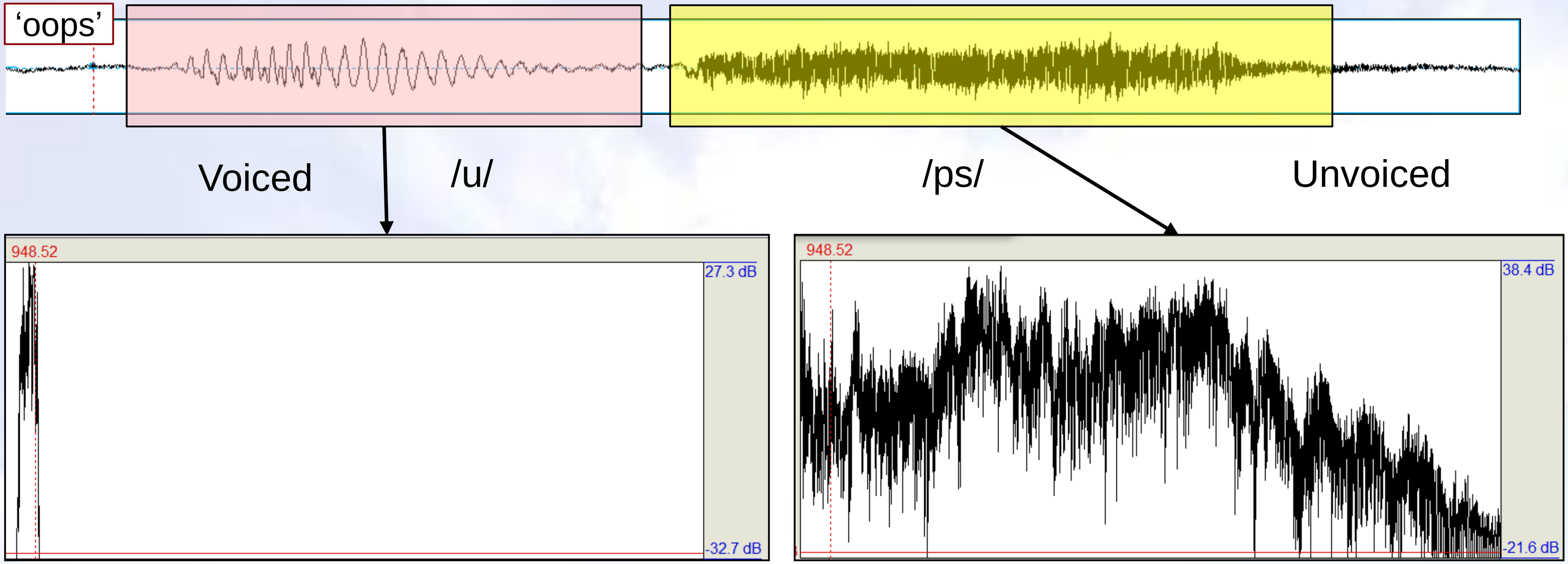
consonants

IPA	examples
b	bad, lab
d	did, lady
f	find, if
g	give, flag
h	how, hello
j	yes, yellow
k	cat, back
l	leg, little
m	man, lemon
n	no, ten
ŋ	sing, finger
p	pet, map
r	red, try
s	sun, miss
ʃ	she, crash
t	tea, getting
tʃ	check, church
θ	think, both
ð	this, mother
v	voice, five
w	wet, window
z	zoo, lazy
ʒ	pleasure, vision
dʒ	just, large

http://www.antimoon.com/how/pronunc-trans.htm

Voiced / unvoiced speech

- **Phone discrimination: phonation**
  - Voiced speech has substantially different spectrum than unvoiced one

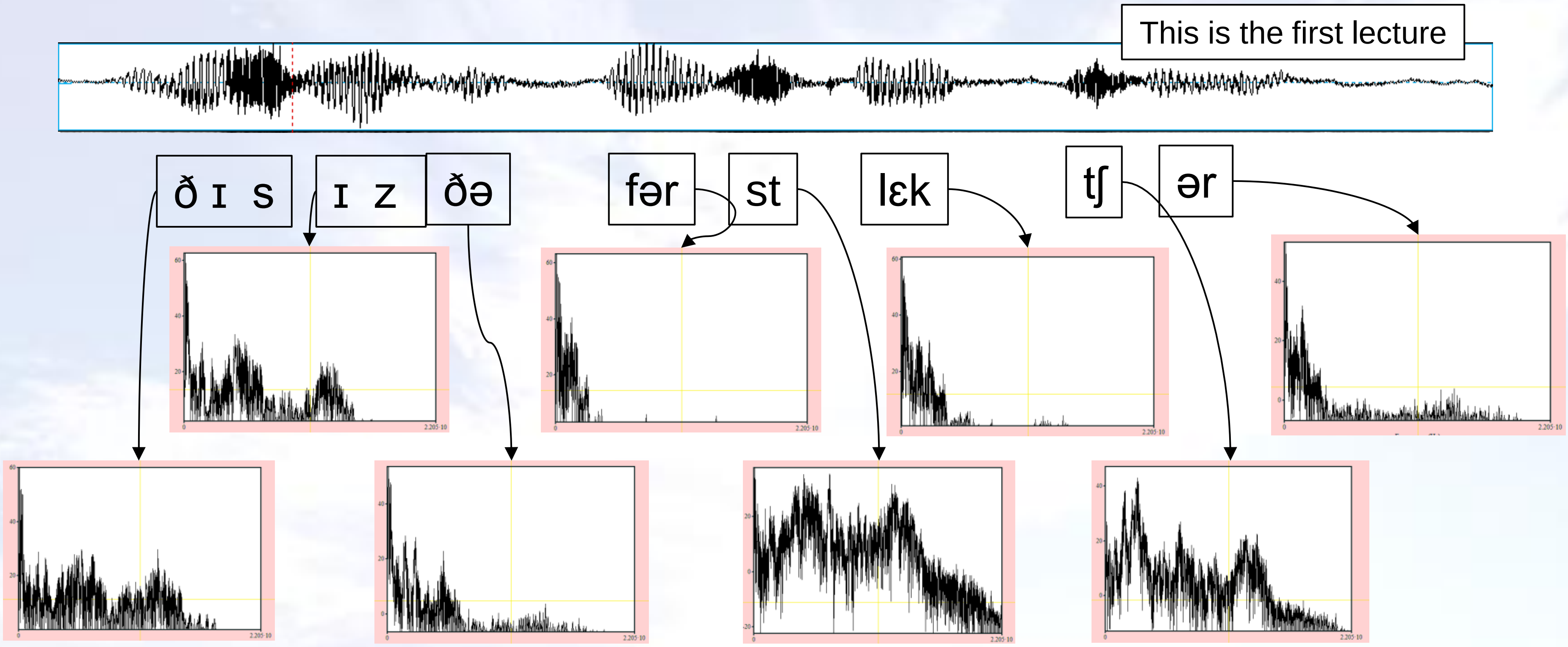


- **Clear differences**
  - Voiced: Few dominant frequencies
  - Unvoiced: wide-band spectrum

Speech signal properties

Instytut
Informatyki
Stosowanej

- **Phone discrimination: articulation**
  - Vowels, nasals, fricatives, stops, liquids etc: articulation differences are reflected in signal properties



Speech representation

- **Corollary**
  - Spectral representation seems adequate
  - Information contents: very rich
  - Is spectrum an appropriate basis for recognition?

- **Spectrum**
  - Decomposition of a function w.r.t. some set of bases
  - Many possible basis functions
  - Periodic basis functions: also several candidates
  - Discrete functions – discrete bases

- **Discrete Fourier Transform - DFT**
  - Harmonic components
  - Fast computational algorithms

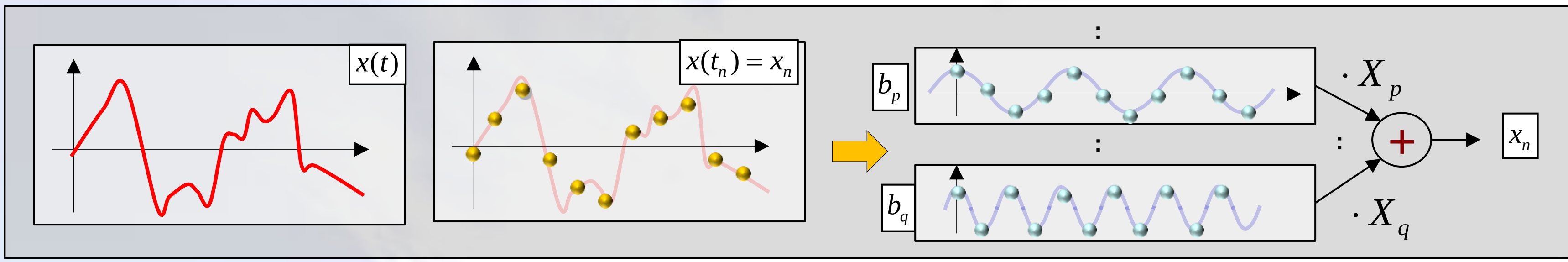
Discrete Fourier Transform: a review

- **Highlights**
  - Decomposition of a (periodic) signal into harmonic components

$x(t_n)=x_n$ expressed as a **composition of bases**

$$x_n = c\sum_{k=0}^{N-1} X_k \bar{b}_{k,n}$$

X_k : k-th weight $\quad b_{k,n}$ k-th basis function $\quad c$:a constant



- DFT bases: complex exponentials (sampled at discrete points)
- For a given signal x: determine the weights of the mixture X_k

$$b_{k,n} = e^{-j\omega_k t_n}$$

t_n : a discrete time instant

DFT

$$X_k = \sum_{n=0}^{N-1} x(t_n)e^{-j\omega_k t_n} = \sum_{n=0}^{N-1} x(t_n)(\cos\omega_k t_n - j\sin\omega_k t_n)$$

$$\omega_k = 2\pi f_k = 2\pi f_s \frac{k}{N} = \frac{2\pi k}{N}$$

f_s Sampling frequency $\quad N$ Number of samples

Discrete Fourier Transform: a review

DFT	$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}kn}$

Signal decomposition

Inverse DFT	$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn}$

Signal reconstruction

- **Comments**
  - Weights regularly spaced in frequency domain
  - Sampling frequency matters: Nyquist condition
  - Number of samples in t and f is the same

$$\boxed{f_k = \frac{k}{N} f_s}$$

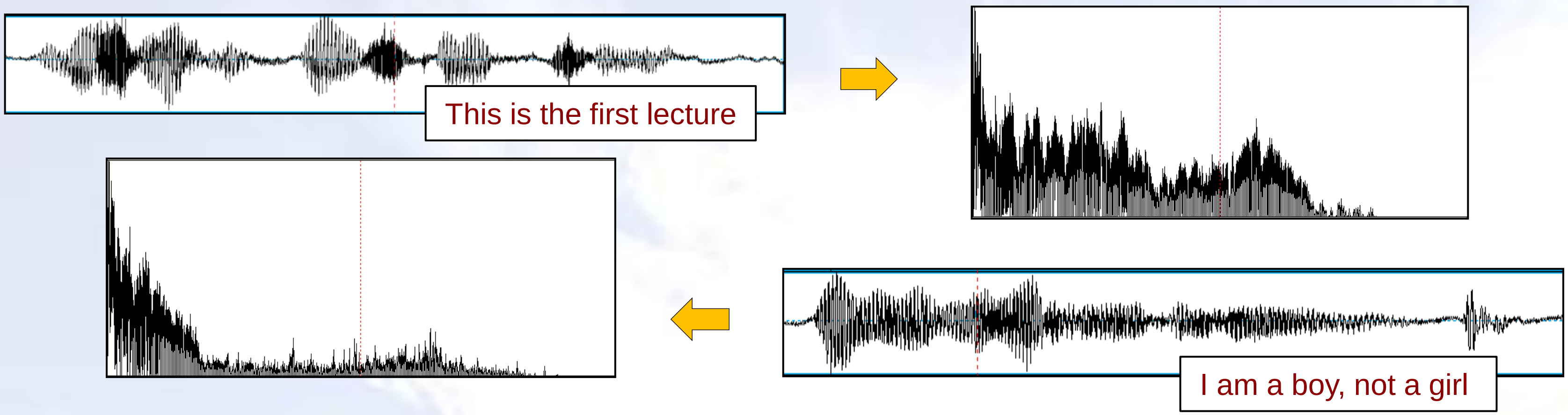
The more points, the higher resolution

- **A meaning of the adopted decomposition**
  - Complex exponential evolving in time: a circle in Im-Re space
  - Projections on a complex exponential: match to cos, match to sin rotated by -90 deg
  - This matches arbitrary phase shift

$$\boxed{x(t)\sin(\omega t + \varphi) = x(t)A_1\cos(\omega t) + x(t)A_2\sin(\omega t)}$$

Signal pectrum

- **Spectrum: a set of decomposition coefficients**
  – Is magnitude spectrum an appropriate basis for speech recognition?

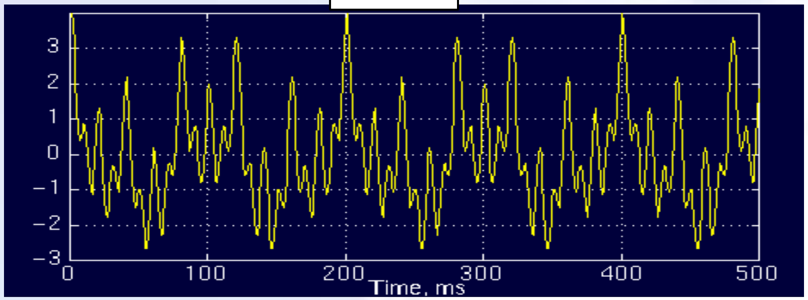


- **The answer**
  – Of course, not
  – I am a boy, not a girl = I am a girl, not a boy
  – Spectrum lacks a temporal resolution

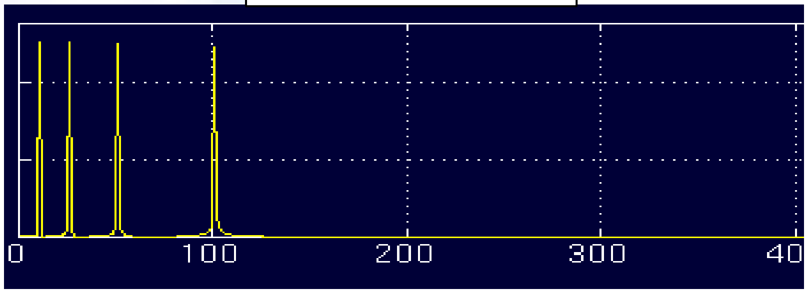
Spectrum and signal recognition

- **Spectrum lacks temporal clues**
  - Unrelated time signals can produce same / similar spectra

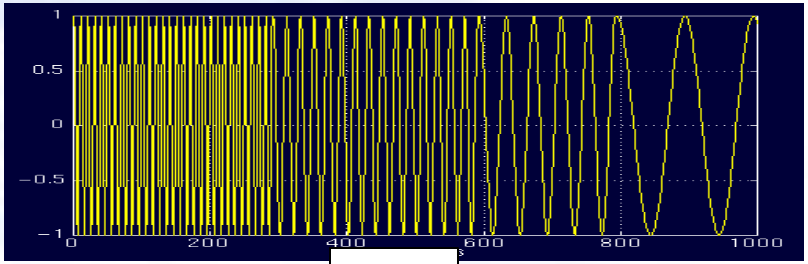
Signal



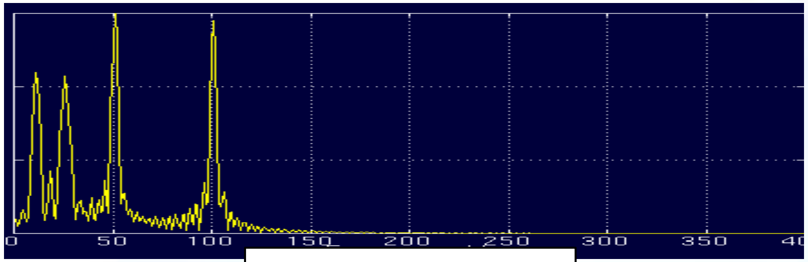
Magnitude spectrum



http://www.public.iastate.edu/~rpolikar/WAVELETS/WTpart1.html



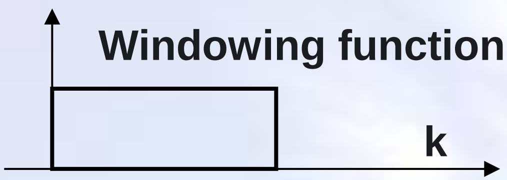
Signal



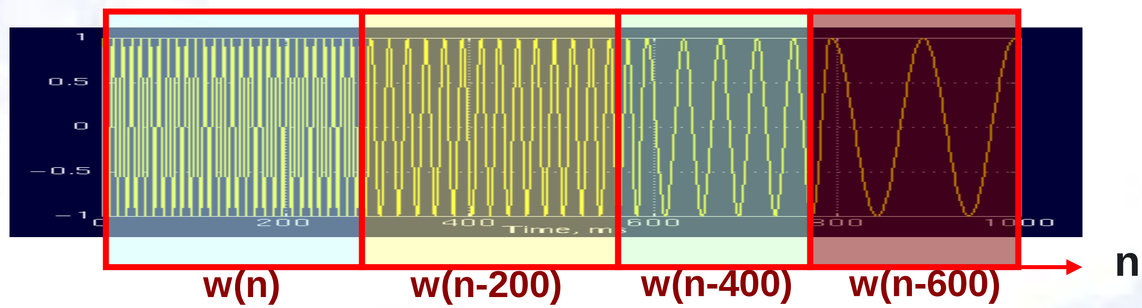
Magnitude spectrum

Short-time Fourier Transform: STFT

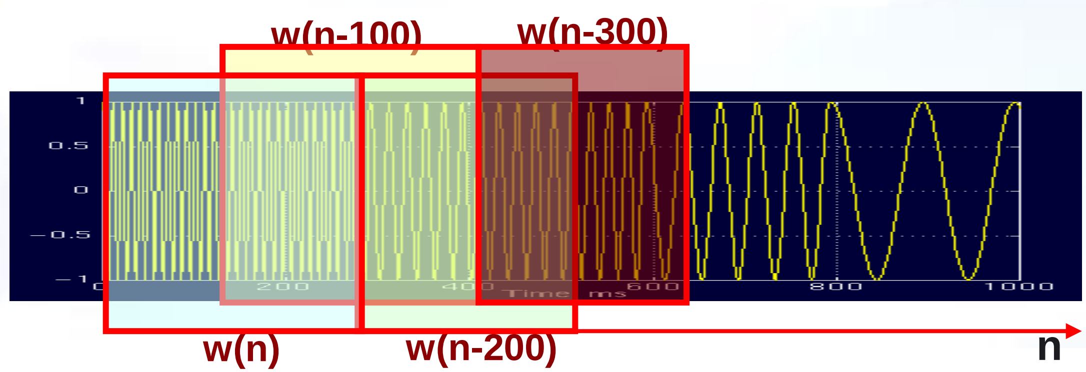
- **An objective**
  - To preserve information on temporal evolution (provide time resolution)
- **Temporal windows**
  - Extract parts of a signal – define domains of subsequent DFT analyses



$$w(n) = \begin{cases} 1 & n < \vartheta \\ 0 & \end{cases}$$

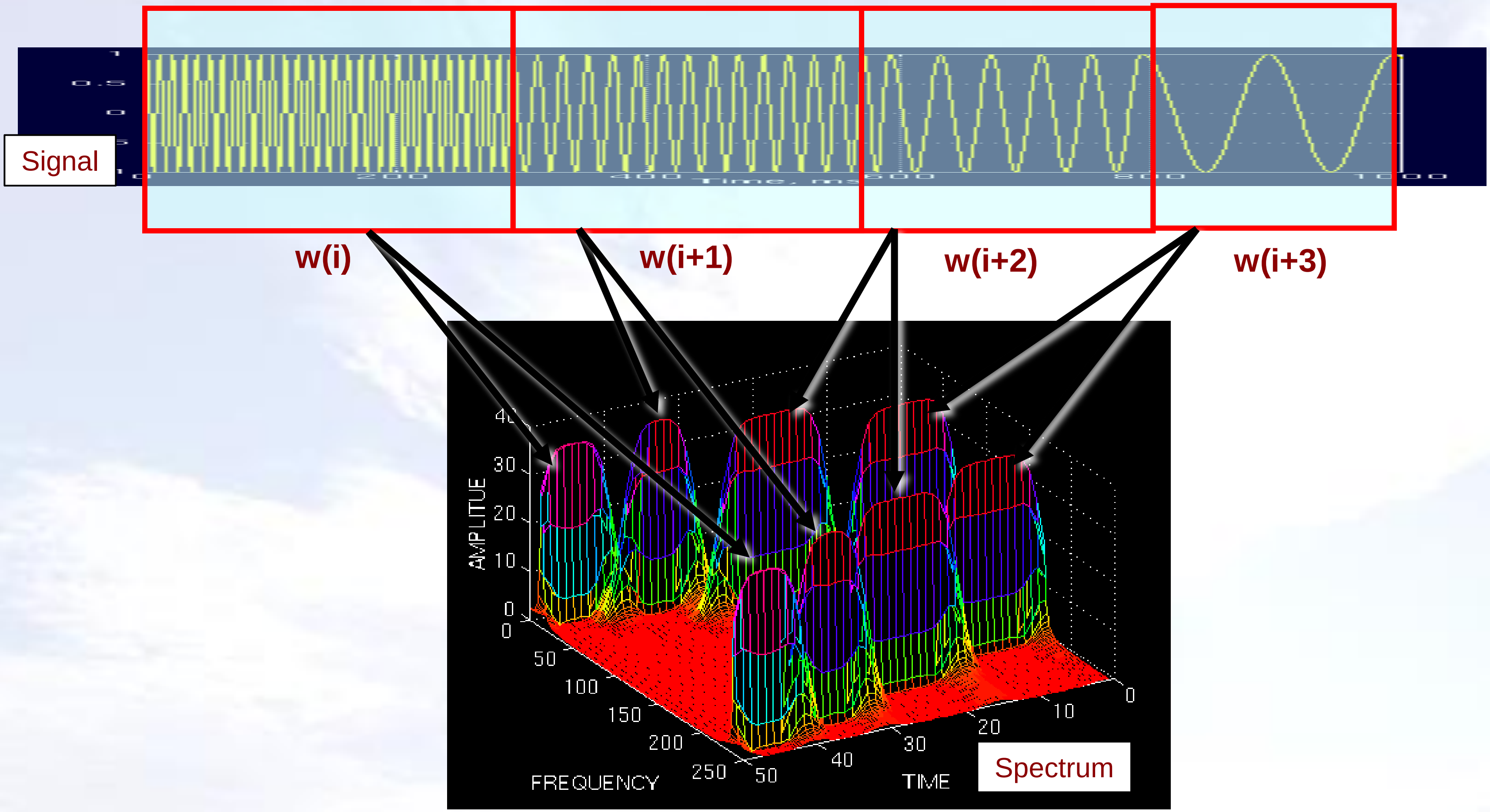


Overlapping windows



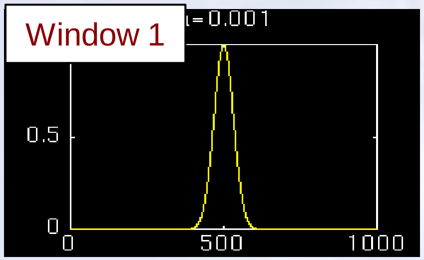
Short-time Fourier Transform: STFT

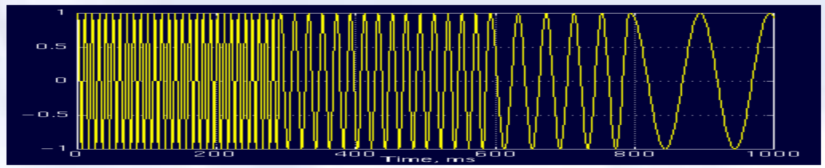
- **Signal representation**
  - Spectra computed for consecutive time 'frames'



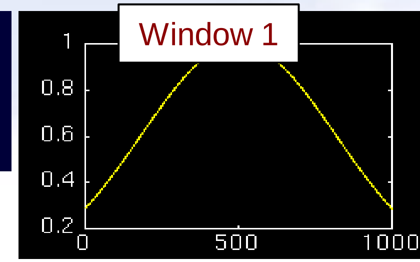
Short-time Fourier Transform: STFT

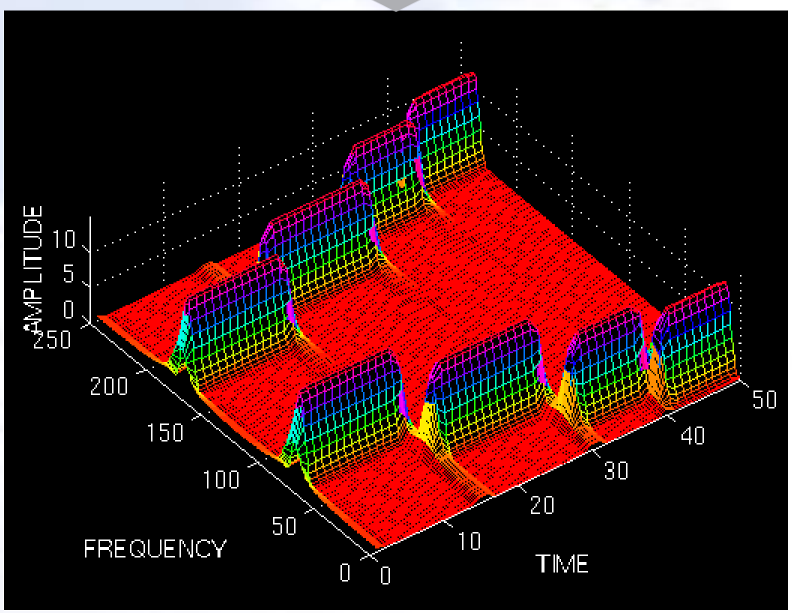
- Trade-off between temporal and frequency resolution (note: overlap)
  - Good time localization – limited frequency

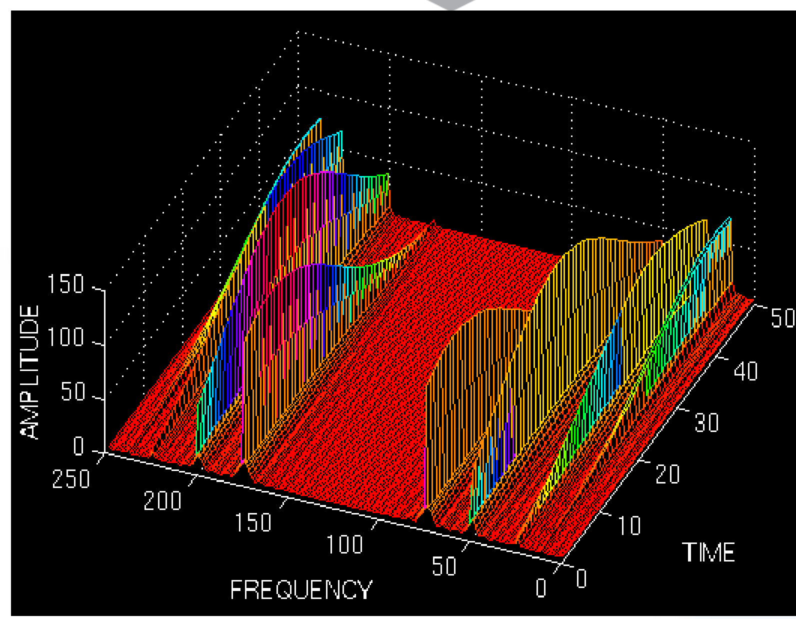




$$w(k) = e^{-[\frac{k-k_0}{a}]^2}$$

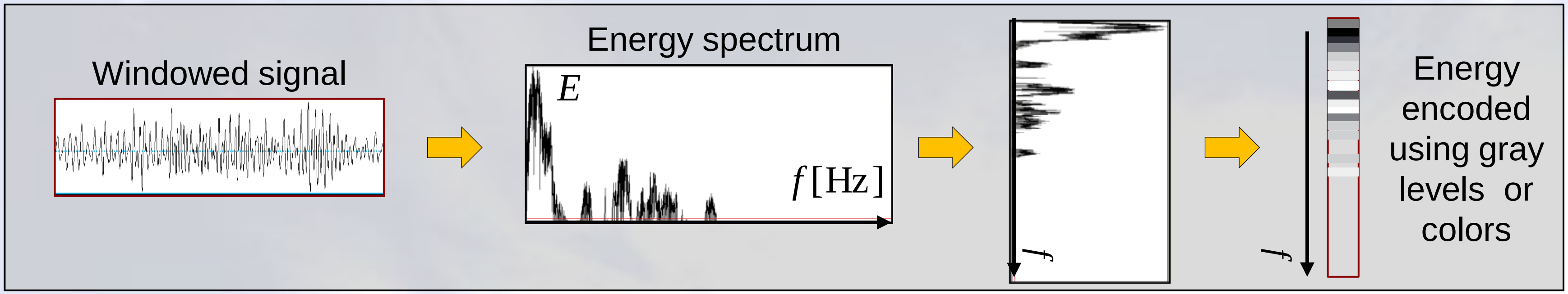


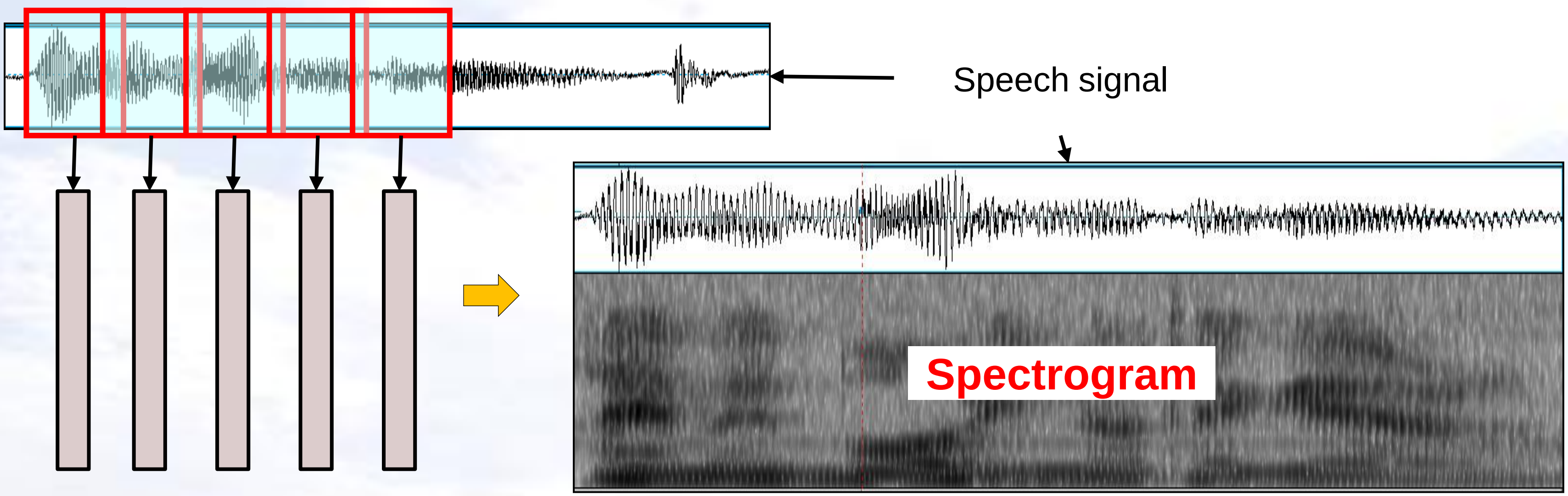




Spectrograms

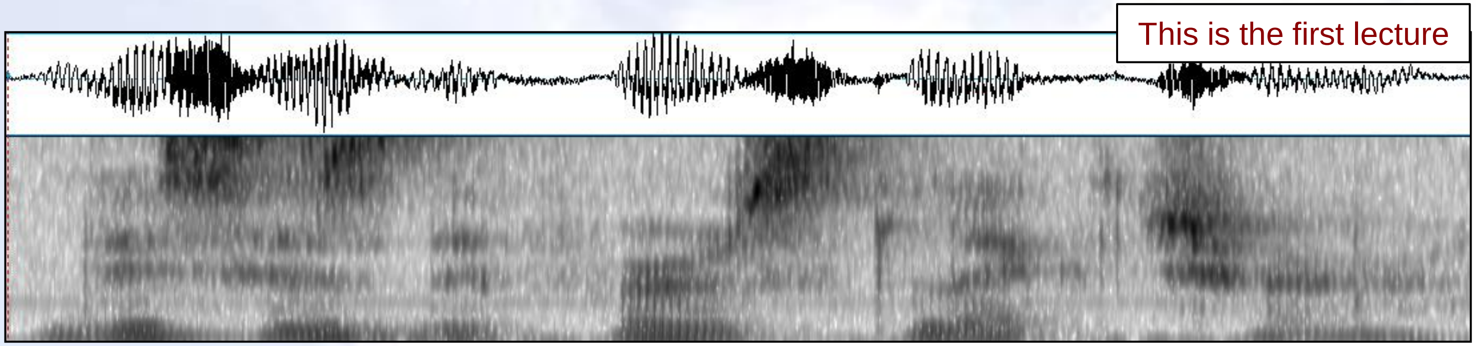
- **Time-frequency representation: spectrogram**
  - Spectra computed for windows extracted at regular time intervals

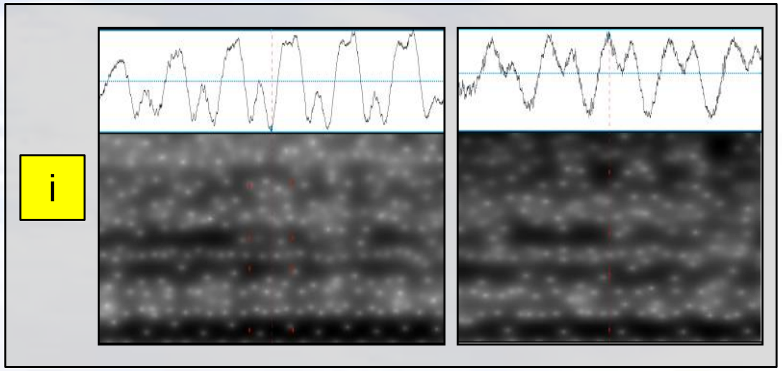




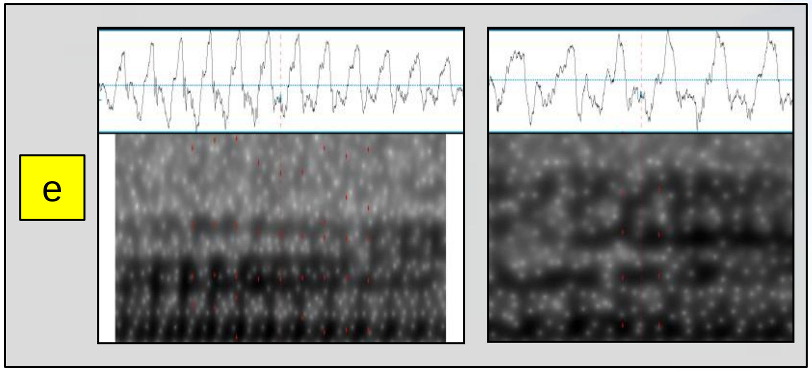
Spectrograms

- **Spectrogram contents**
  - Information on temporal dynamics of speech (sound) production
  - Resonant cavities characteristic for phones are reflected by spectral contents of spectrogram





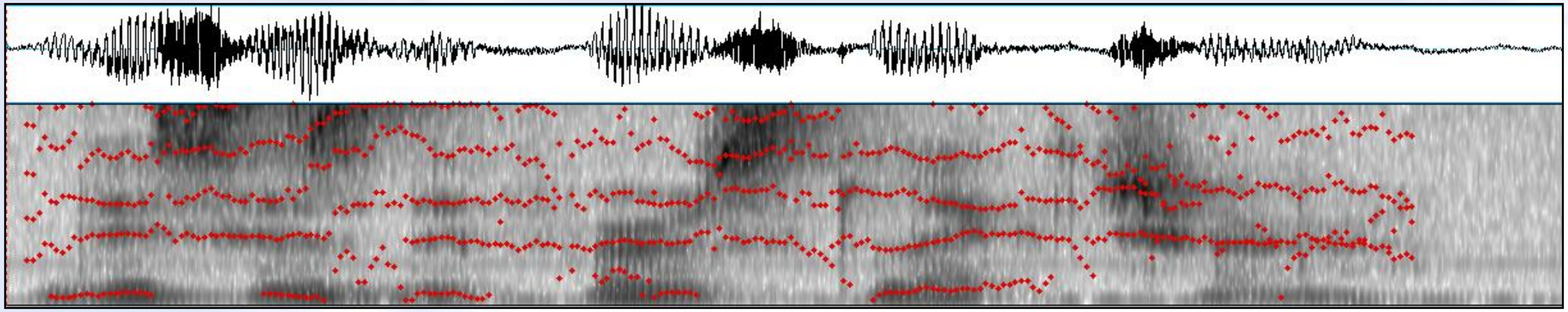
Spectrogram structure for ‘e’



Spectrogram structure for ‘i’

Speech representation

- **Spectrogram features**
  - Bands of high energies: **formants**



- **Formants**
  - Produced in resonant cavities: local energy maxima in various frequency bands
  - Vary in time as different phones get articulated
  - Form patterns characteristic to phones
  - Exist for voiced speech only , typically, up to four (can be hard to extract)
  - Can be used e.g. for vowelclassification, are insufficient for speech representation – richer representation is required

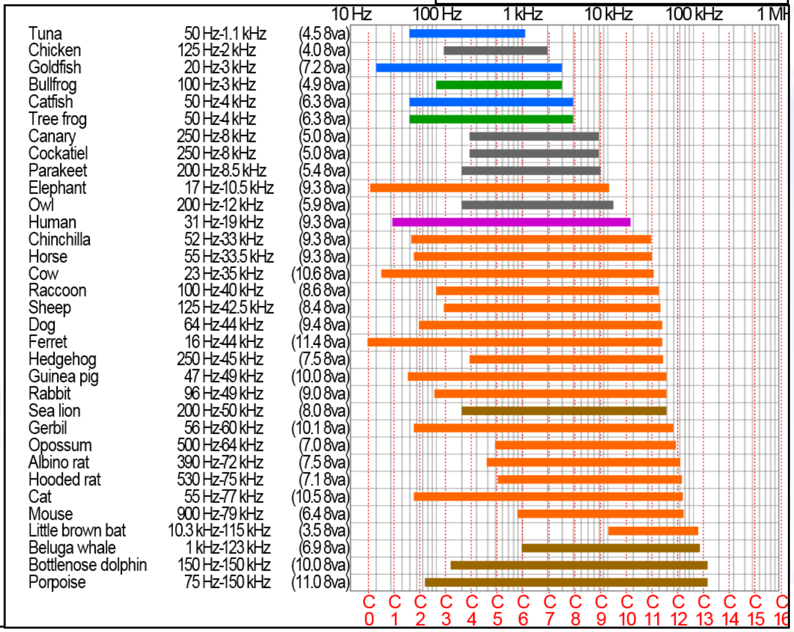
Spectrogram computation considerations

Instytut
Informatyki
Stosowanej

- **Spectrogram**
  - Provides information on temporal evolution and frequency composition of signal in windows
  - Should carry all relevant information
    - Q1 - what frequencies should be covered?
    - Q2 - what should be duration of a window?

- **Frequency range for speech representation**
  - Tip: knowledge on auditory perception - 20 Hz to 20 kHz (sampling frequency for digital music is 44 kHz)
  - Telephony: channel bandwitdh: 300 Hz – 3.4 kHz (sampled at 8 kHz)

http://en.wikipedia.org



Window length

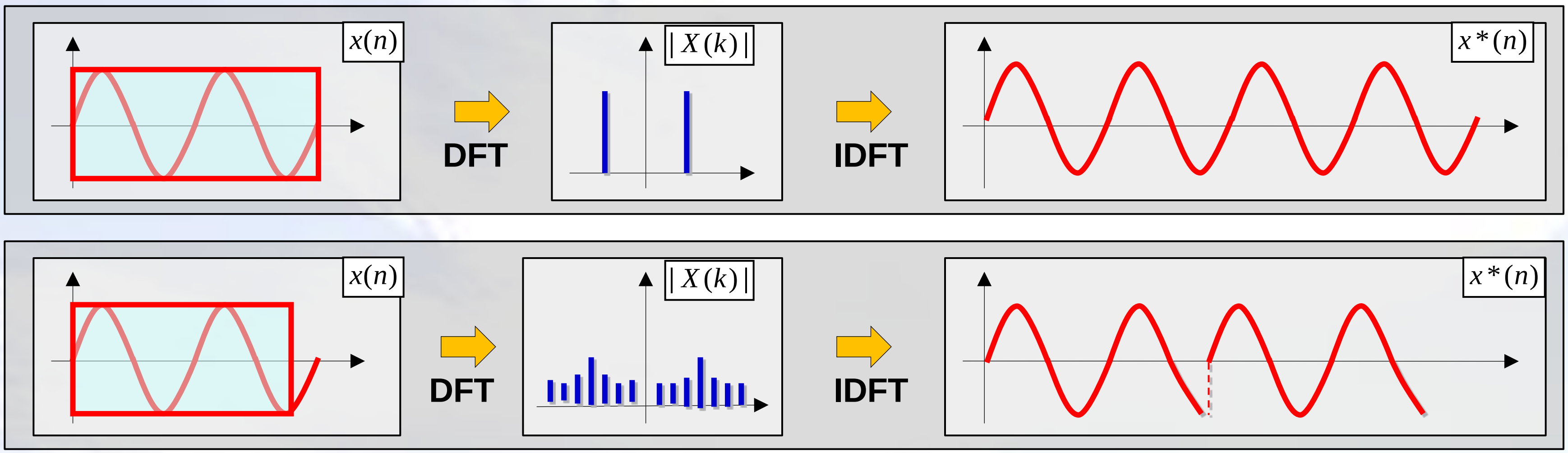
- **Window length for speech representation**
  - The window cannot be neither too long nor too short
- **Upper bound for window length**
  - Stationarity of the underlying process: a filter (articulation cavities) should not change its parameters (otherwise, frequency information becomes useless for phone identification)
- **Lower bound for window length**
  - Too short window – unable to represent low frequencies (to fit full period of 100 Hz wave, 10 ms window needs to be used)
- **Typical window size**
  - Within a range 10-30 ms
  - Exact duration often adjusted to satisfy DFT requirements (to have a number of samples equal to some integer power of 2):
    - 8kHz sampling (T_s =0.125 ms) 256 samples give T=32 ms, which means that max freq < 4kHz, min freq approx. 30 Hz

Windowing functions

$$y(n) = w(n)x(n)$$

$$w(n) = \begin{cases} 1 \\ 0 \end{cases}$$

- **A role of the window**
  - To preserve information on temporal order of events
- **Rectangular window**
  - The rectangular window ensures extraction of consecutive parts of a signal
  - Is there any problem with a rectangular window?

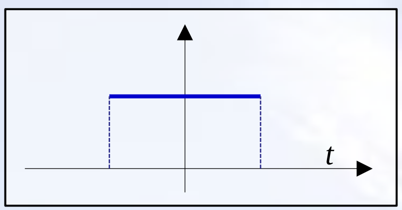


- **The problem**
  - DFT attempts to represent a discontinuous function (periodic)
  - For rectangular window inevitable! distortions Spectrum ‘leaks’ over neigboring frequencies

Windowing functions

- **Problems caused by spectral leakage**
  - Introduction of ‘false’ signal components
  - Potential hiding of lower, actual components

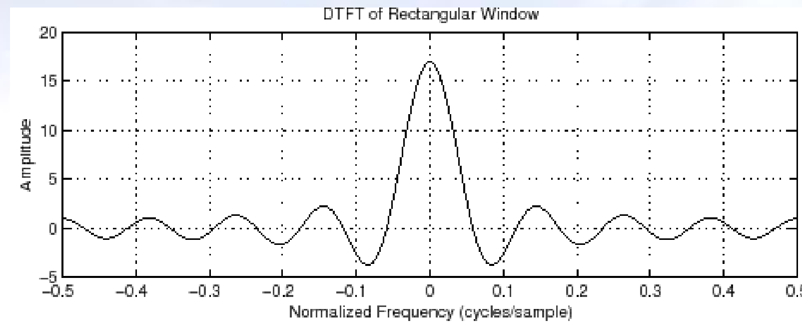
Rectangular window



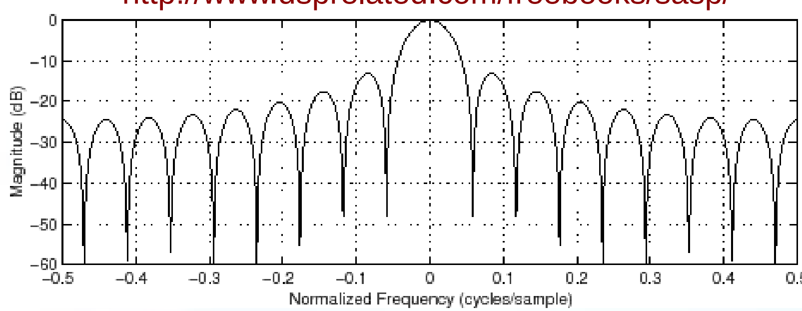
Time domain

DFT

Frequency domain



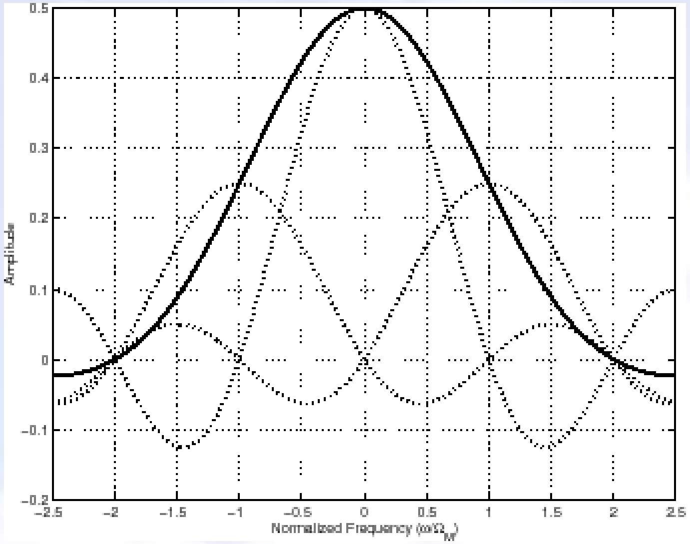
http://www.dsprelated.com/freebooks/sasp/



- **A window should reduce the disconinuity effect**
  - Time domain: reduce magnitudes at both signal ends;
  - Frequency domain; reduce side lobes

Windowing functions

- **A gallery of potential candidates for windows**
  - Hamming, Hanning (cosine combined with a rectangular window)



http://www.dsprelated.com/freebooks/sasp/

  - Three sinc functions in frequency domain, scaled in proportions *a*, and two *b* and appropriately shifted
  - Results in eliminating a side lobe (at the expense of widenning the central one)

$$W(n) = W_R(n)\left[a + 2b\cos\left(\frac{2\pi}{N}n\right)\right]$$

- **Hanning window**
  - a=0.5, b=0.25

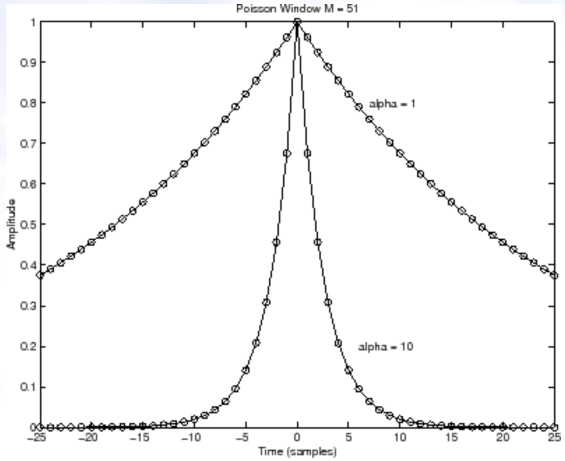
$$\Rightarrow \quad W_{HN}(n) = W_R(n)\cos^2\left(\frac{\pi}{N}n\right)$$

- **Hamming window**
  - a=0.54, b=0.23 (chosen to completely cancel the largest side lobe)

Windowing functions

Instytut
Informatyki
Stosowanej

- **Other candidates**
  - Blackman-Harris window family: similar concept as for Hanning/Hamming, but more sinc functions are used
  - Triangular - Bartlett window (convolution of two rectangular windows)
  - Poisson window



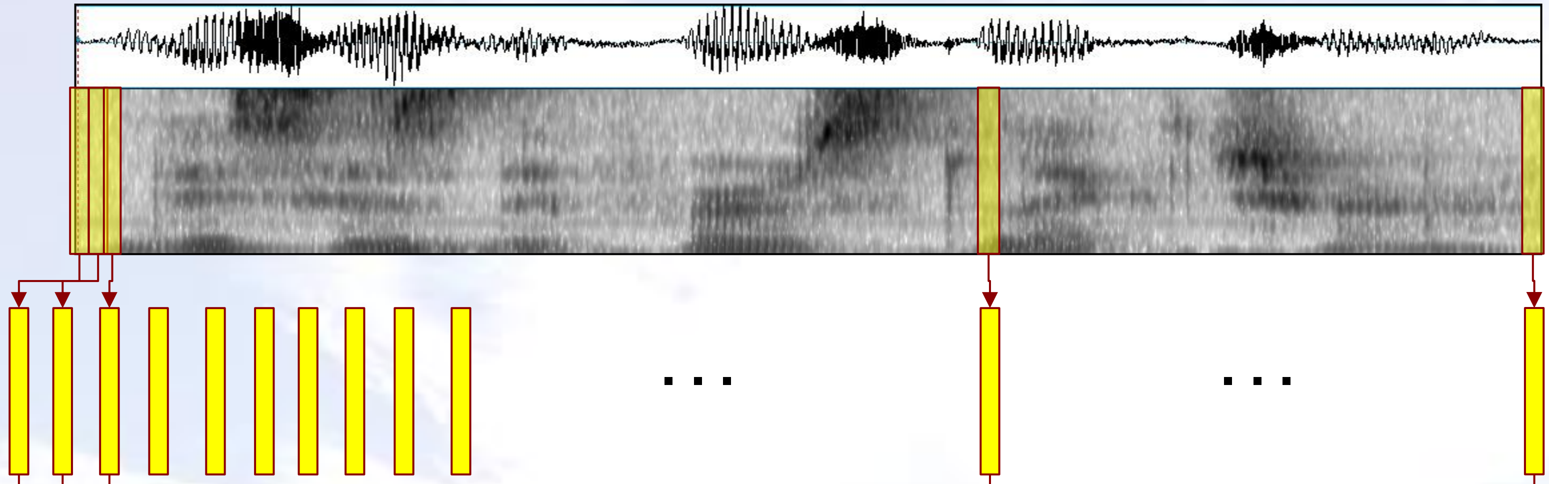
http://www.dsprelated.com/freebooks/sasp/

$$W_{HN}(n) = W_R(n)\cos^2\left(\frac{\pi}{N}n\right)$$

  - Gaussian, Kaiser …

Speech representation

- **Spectrogram**
  - A sequence of slices. A slice: a vector of DFT magnitudes for a window
  - A sequence of vectors

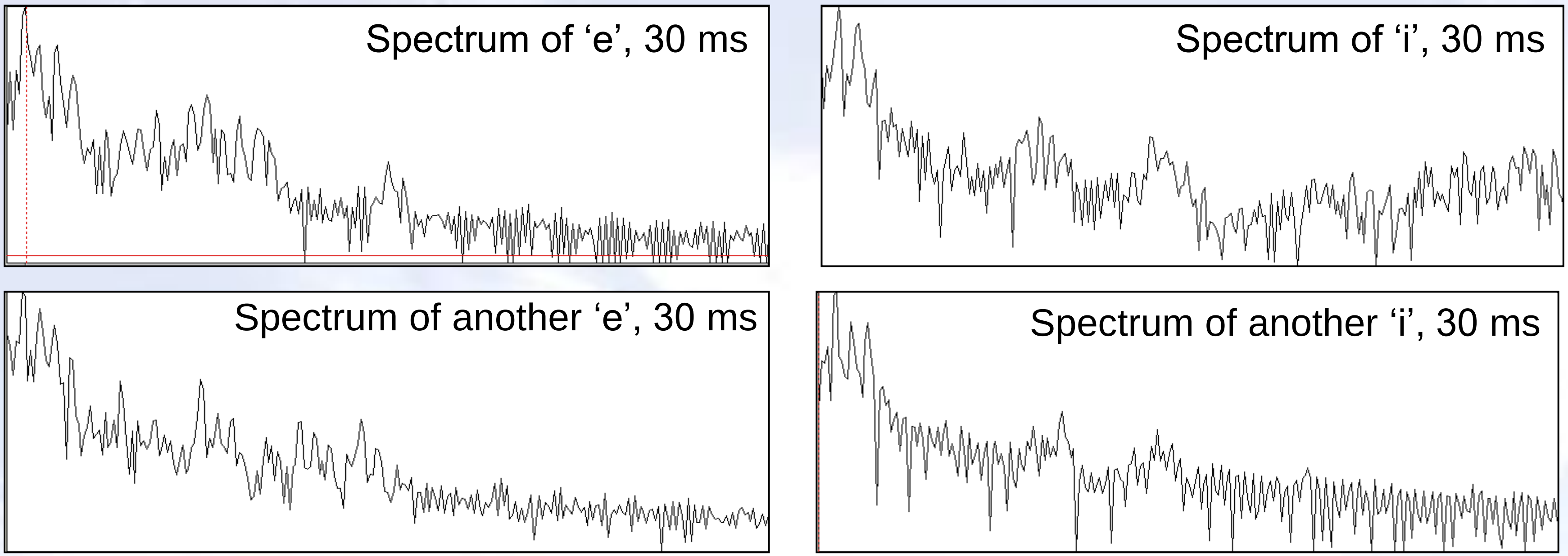


$$\mathbf{x}(t=0)=\begin{bmatrix} x_0^0 \\ x_1^0 \\ \vdots \\ x_{n-1}^0 \end{bmatrix} \quad \mathbf{x}(t=1)=\begin{bmatrix} x_0^1 \\ x_1^1 \\ \vdots \\ x_{n-1}^1 \end{bmatrix} \quad \mathbf{x}(t=2)=\begin{bmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_{n-1}^2 \end{bmatrix} \quad \cdots \quad \mathbf{x}(t=i)=\begin{bmatrix} x_0^i \\ x_1^i \\ \vdots \\ x_{n-1}^i \end{bmatrix} \quad \mathbf{x}(t=T-1)=\begin{bmatrix} x_0^{T-1} \\ x_1^{T-1} \\ \vdots \\ x_{n-1}^{T-1} \end{bmatrix}$$

Spectrogram: a sequence of vectors $\{\mathbf{x}_0, \mathbf{x}_1 \ldots \mathbf{x}_{T-1}\}$ ⇒ **Speech signal representation**

Speech representation

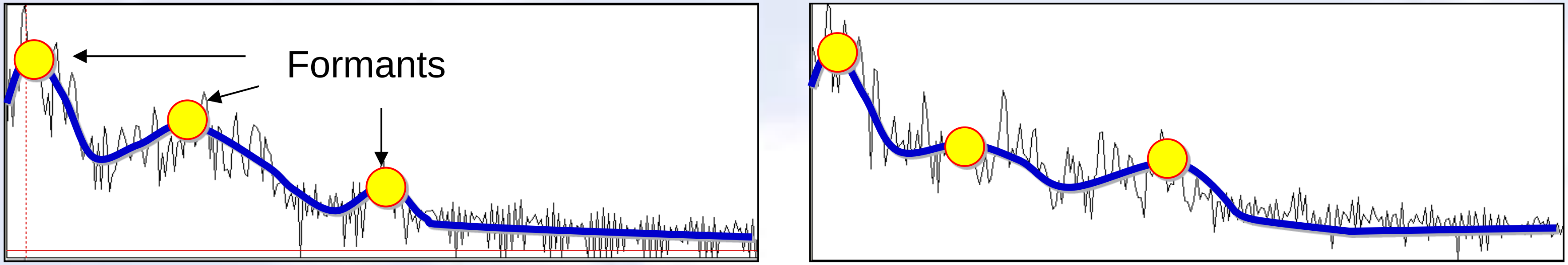
- **Feature vectors**
  - All DFT components?



- **Observation 1**
  - Huge within-class variability, unclear between-class differences: probably useless representation
  - Better representation needs to be found

Speech representation

- **Observation 2**
  - Spectrum looks like a mixture of slow-varying and fast-varying components



- ‘Slow’ (frequency axis) varying components: combination of filters corresponding to resonant cavities (articulation)
- Formant frequencies – filter maxima

- **Informative representation of speech**
  - Slow-varying frequency-domain components
  - Analogy to performing low-pass filtering of temporal signals

- **Filtering of temporal signals**
  - Based on assumption of linearity (additive ‘noise)

Speech representation

- **Temporal signal filtering**
  - Get Fourier spectrum
  - Retain selected frequencies (low for eliminatiion of noise)
  - Underlying model: linear composition of a signal

$$y(t) = x(t) + n(t) \Rightarrow Y(f) = X(f) + N(f) \Rightarrow \tilde{Y}(f) = X(f) \Rightarrow \tilde{y}(t) = x(t)$$

- **Speech signal**
  - Convolution of source (vocal fold excitation) and filter (articulation cavities)
  - Spectrum is a product of the two
  - We want to get rid of excitation (for us this is a noise)
  - We are in frequency domain: how to proceed?

$$s(t) = a(t) * p(t)$$

$s(t)$ speech signal $p(t)$ phonation

$a(t)$ articulation (impulse response)

$$\Rightarrow S(f) = A(f)P(f)$$

Speech representation

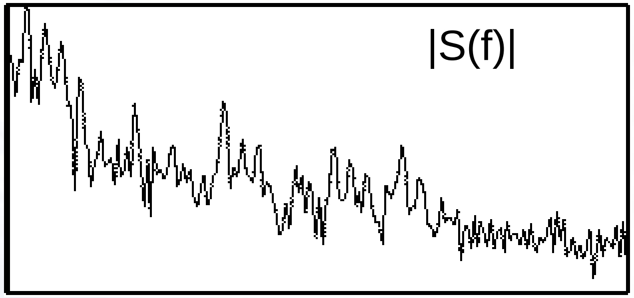
- **The problem**
  - Linear separation of components given in a product form

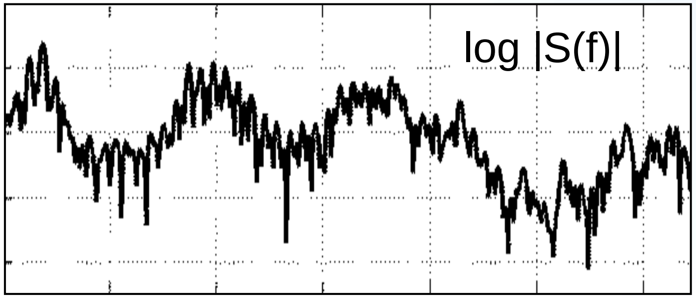
$$S(f) = A(f)P(f)$$

- **Solution**
  - Homomorphic filtering (simple trick: use logarithms)

$$\boxed{S(f) = A(f)P(f)} \Rightarrow \boxed{\log S(f) = \log A(f) + \log P(f)}$$

  - Spectrum becomes a sum of components: can be approached using the presented framework

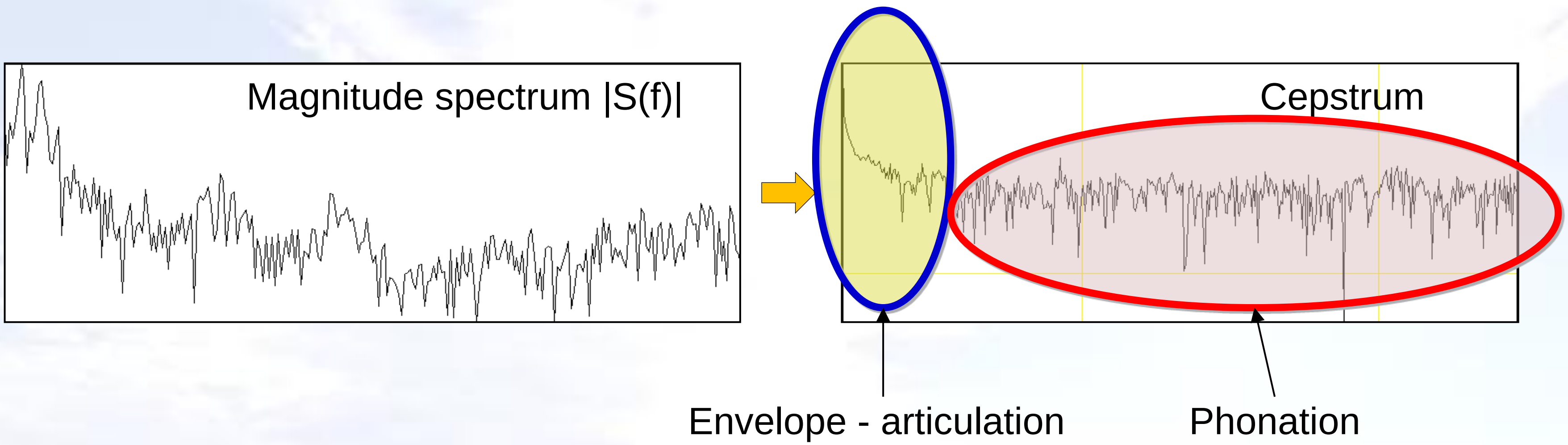




Speech representation

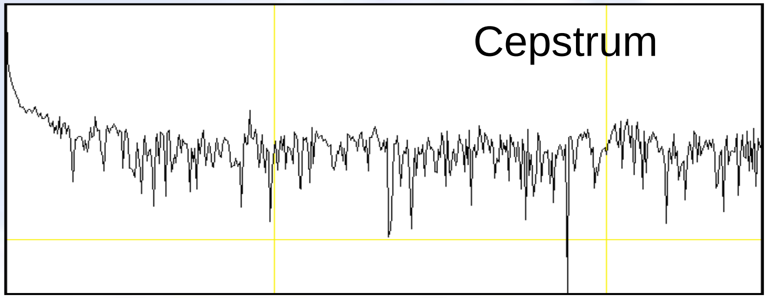
Instytut
Informatyki
Stosowanej

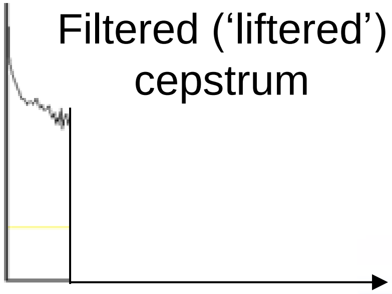
- ‘Spectral decomposition’ of log(abs(spektrum))
  - An objective: to separate slow-varying and fast-varying components
  - Apply inverse Fourier transform on log-abs spectrum
  - We go back to time axis, but this is not a reconstructed signal
  - To emphasize a difference it is called CEPSTRUM (reordering of SPECTRUM)
  - Cepstral analysis terminology: quefrency, liftering, …



Speech representation

- **Low-order cepstral coefficients**
  - Low-order cepstral coefficients correspond to spectral envelope – represent acoustic filters formed by resonant cavities of vocal tract (speech articulation)
  - Possible methodology: cepstrum truncation (10-20 leading coefficients are left)





- **The first representation of speech signal**
  - A sequence of vectors (for each window) containing low-order cepstral coefficients

Cepstral mean subtraction

- **Channel impact on speech signal properties**
  - A channel: all components that exist between speech production point and a point where electrical speech waveform is produced
  - Path variability: different laptops where ASR software is executed have different microphones so prototypes are hardware-specific, room-specific reverberations …
  - Channel effects: convolution in time (product in frequency)
  - Real channels– non-ideal frequency response: spectral distortions

- **Cepstral mean**
  - Cepstral vector: combination of speech and channel components
  - Channel component does not change
  - Mean of cepstral vectors
  - Mean cepstrum of speech is zero (as mean of original signal is zero)
  - Mean subtraction: channel elimination

$$\boxed{\mathbf{C}_i = \mathbf{C}_i^s + \mathbf{C}^{ch}}$$

$$\boxed{\frac{1}{n}\sum_{i=0}^{n-1}(\mathbf{C}_i^s + \mathbf{C}^{ch}) = \frac{1}{n}\sum_{i=0}^{n-1}\mathbf{C}_i^s + \mathbf{C}^{ch}}$$

$$\boxed{\boldsymbol{\mu} = \mathbf{C}^{ch}}$$

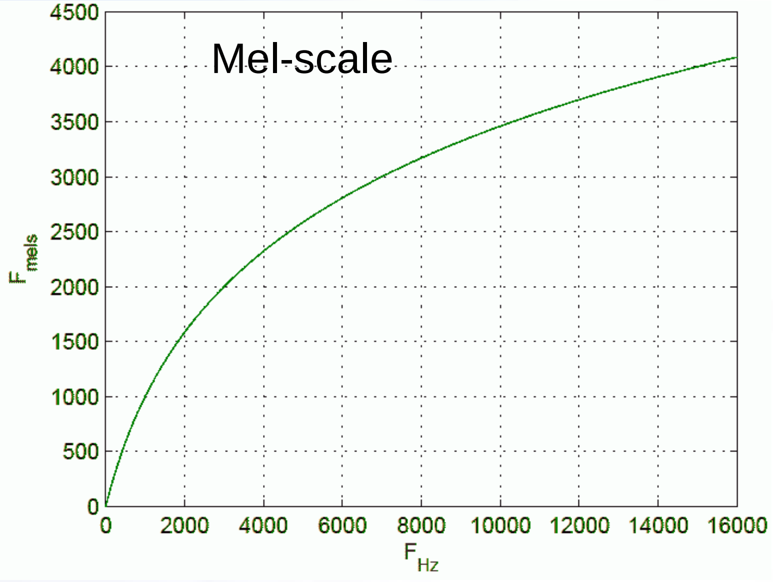
$$\boxed{\tilde{\mathbf{C}}_k = \left(\mathbf{C}_k^s + \mathbf{C}^{ch}\right) - \boldsymbol{\mu} = \mathbf{C}_k^s}$$

Mel-frequency scale and MFCC

Instytut
Informatyki
Stosowanej

- **Possible speech representation**
  - A sequence of cepstrum-mean subtracted low-order cepstral coefficients
  - Why it is not commonly used?

- **Psychology of hearing**
  - Perception of frequencies is nonlinear: linear up to 1 kHz, then logarithmic
  - Masking phenomenon and critical bands

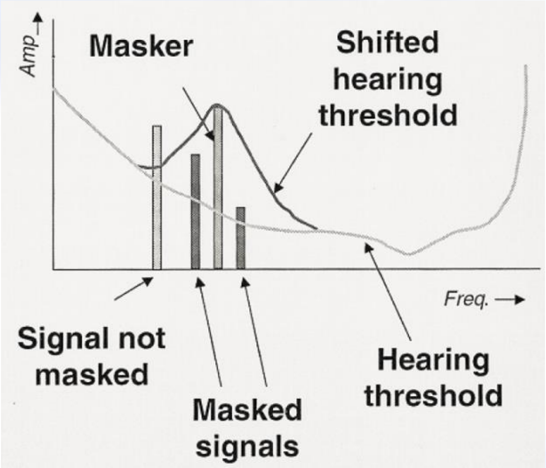


For F>1 kHz

$$F^{Mel} = 2595 \log_{10}\left(1 + \frac{F^{Hz}}{700}\right)$$

Mel-frequency scale and MFCC

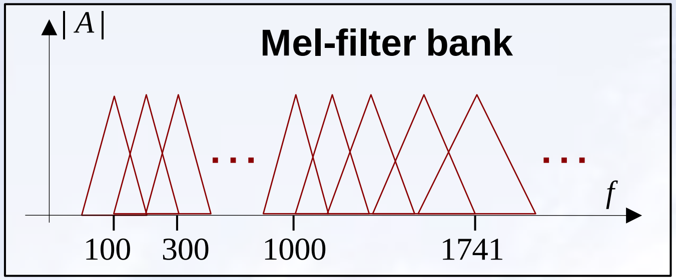
- **Masking**
  - Hiding neighboring tones by a stronger component
  - Can be explained only if we assume that we perceive sounds within bands (critical bands)
  - Ear acts as a set of filters
  - Cannot ignore it in modeling speech

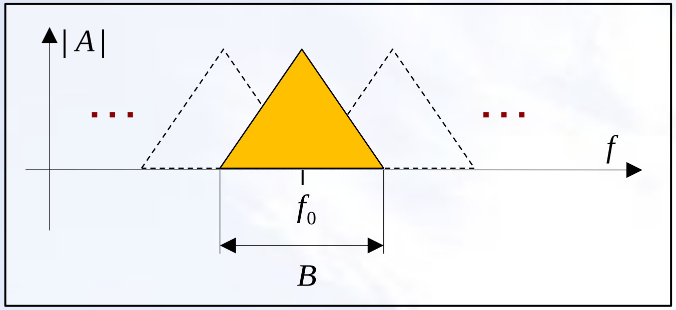


- **Integration of frequency stimulus using filters**
  - Filters distributed evenly in subjective frequency domain (Mel-scale)
  - Filters have same width in Mel-scale
  - … i.e. they are nonlinearly distributed along objective frequency axis

Mel-frequency scale and MFCC

- **Mel-filter bank**
  - 24 triangular overlaping filters
  - Centers – linear in Mel-scale
  - Bandwidth: masking range





f_0	B	f_0	B
100	100	1516	211
200	100	1741	242
300	100	2000	278
400	100	2297	320
500	100	2639	367
600	100	3031	422
700	100	3482	484
800	100	4000	556
900	100	4595	639
1000	124	5278	734
1149	160	6063	843
1320	184	6964	969

- **Mel-spectrum**
  - Speech spectrum is integrated within critical bands
  - Can be seen as spectrum downsampling

Mel-frequency scale and MFCC

- **MFCC**
  - Spectrum is accumulated in bins defined by Mel-filters
  - 24 filters – 24 intervals along frequency axis
  - Mel-Cepstrum computed from 24-element Mel-spectrum
  - Result: Mel-Frequency Cepstral Coefficients (MFCC)
  - Low-order MFCC's use to represent articulation (typically – up to 12)

Modeling articulation dynamics

- **Dynamics**
  - A rate of change: delta-MFCC: subtraction of MFCC's from consecutive frames
  - Second derivative of MFCC changes: delta-delta-MFCC

- **Common speech representation**
  - A sequence of vectors (for each window) containing low-order MFCC, delta-MFCC and delta-delta coefficients,