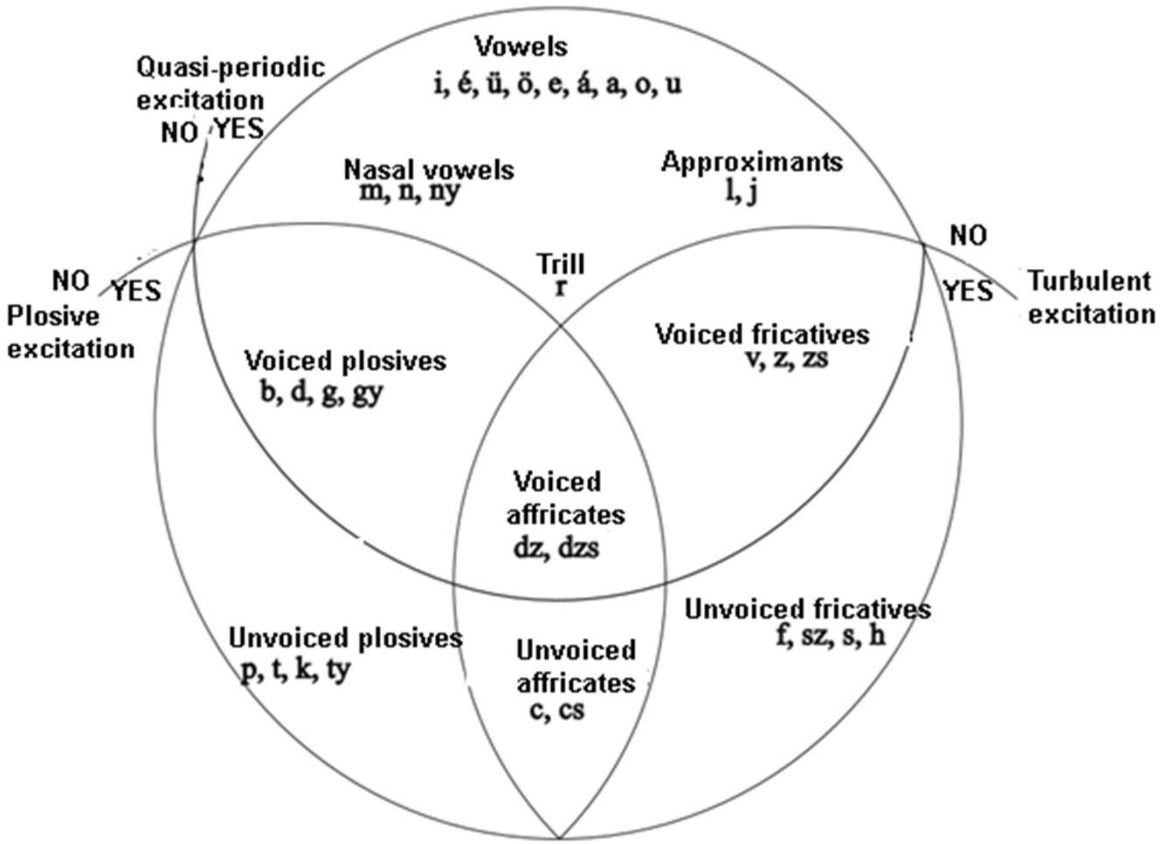
6. Classification of speech sounds

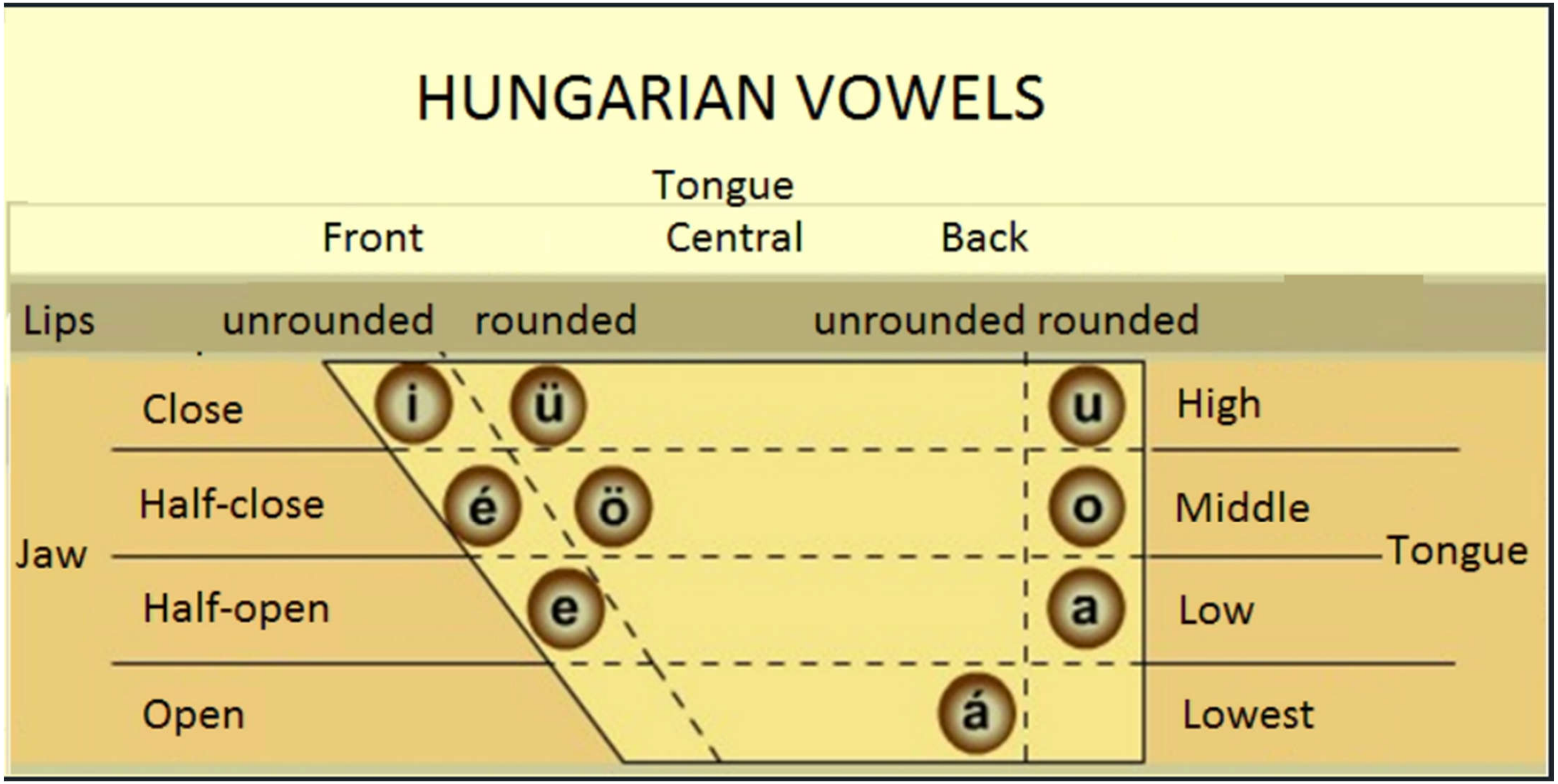
Classification categories

- *Vowel-consonant.* The vowel is a syllabic word.
- *Short-long sound.* E.g. *varr, var, öt, őt*
- *Oral-nasal.* E.g. *mama*
- *Place of generation* The sound is generated at the level of the larynx, its place of generation is constant. The noise resulting from turbulent flow is generated in the constrictions created at different points of the articulation canal. The place of generation of the latter is therefore variable.
- *Excitation form.* Purely voiced or purely noisy, or voiced-noisy excitation
- *Simple-complex sound structure.* E.g. *cat*
- *Articulatory positions*

According to the nature of excitation



Articulatory configuration of Hungarian vowels

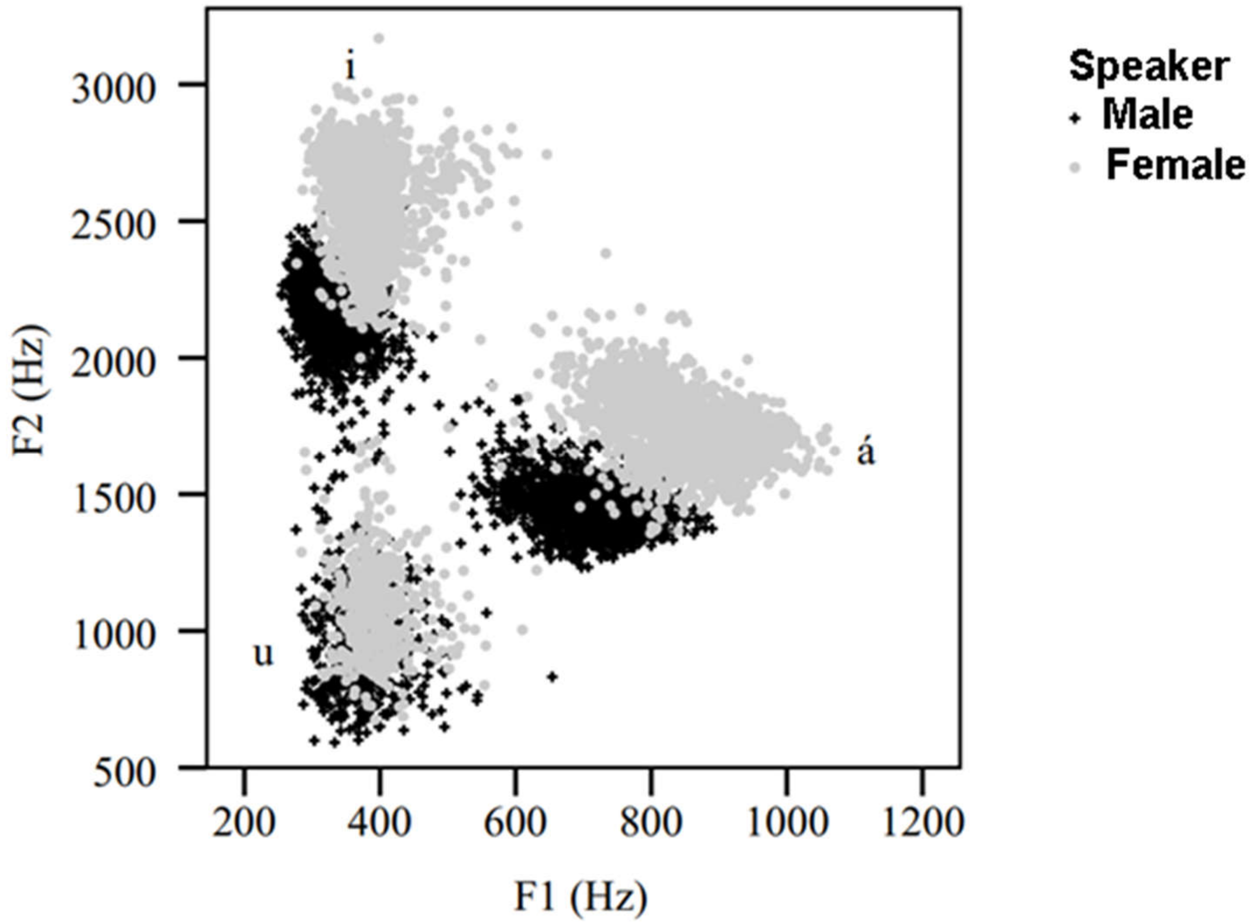


SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

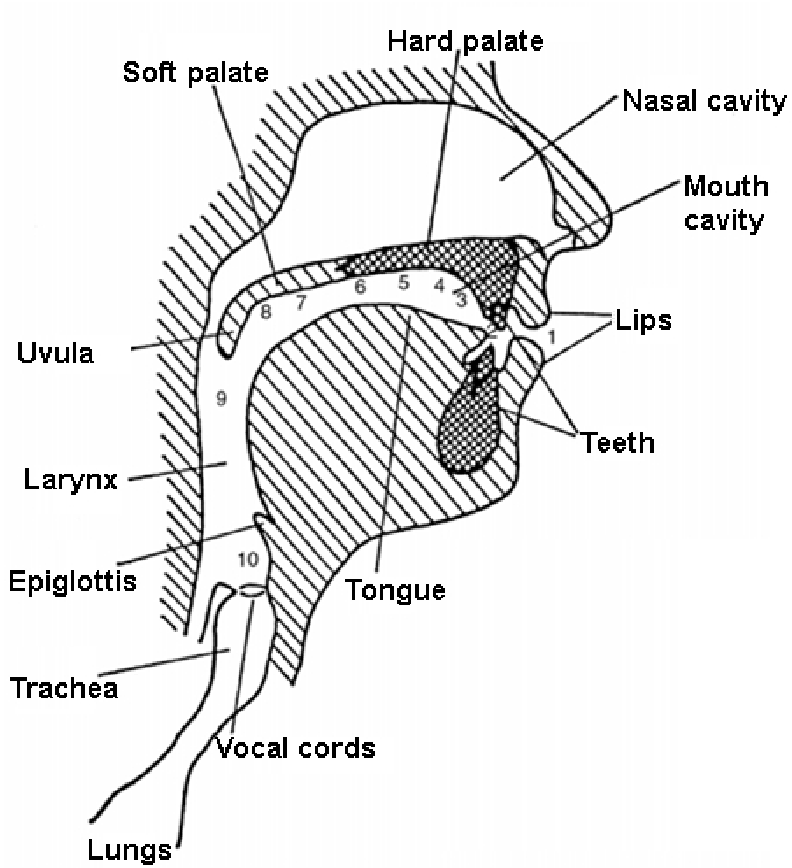
BME TMIT

GPU EDUCATION CENTER

Formant map of Hungarian vowels F1-F2



Names of articulatory positions



1. (Bi)Labial
2. Dental
3. Alveolar
4. Prepalatal
5. Palatal
6. Postpalatal
7. Velar
8. Uvular
9. Pharyngeal
10. Glottal

Articulatory configuration of Hungarian consonants

	Labial	Dental[2]	Post-alveolar	Palatal	Velar	Glottal
Nasal	m	n		ɲ		
Plosive	p b	t d		c ɟ*	k g	
Affricate		t͡s d͡z	t͡ʃ d͡ʒ			
Fricative	f v	s z	ʃ ʒ			h
Trill		r				
Approximant		l		j		

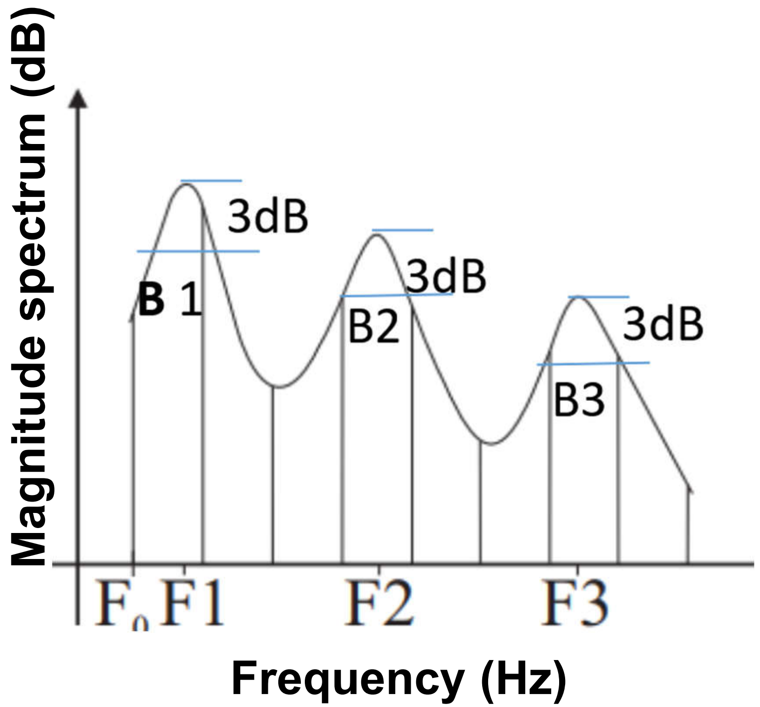
The complex classification of Hungarian consonants

Place of articulation								
Mode of production	Bilabial	Labio-dental	Denti-alveolar	Alveolar	Palatal	Velar	Pharyngeal	Source
Plosives	b							Voiced
	p							Unvoiced
			d					Voiced
			t					Unvoiced
					gy			Voiced
					ty			Unvoiced
						g		Voiced
						k		Unvoiced
Fricatives		v						Voiced
		f						Unvoiced
			z					Mixed
			sz					Unvoiced
				zs				Mixed
				s				Unvoiced
					h*			Unvoiced
						h**		Unvoiced
							h	Unvoiced
							h***	Voiced
					j*			Unvoiced
Affricates			dz					Mixed
			c					Unvoiced
				dzs				Mixed
				cs				Unvoiced
Approximants				l				Voiced
					j			Voiced
Trill			r					Voiced
Nasals	m							Voiced
		mv*						Voiced
			n					Voiced
						ng*		Voiced
					ny			Voiced

SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

Notation and concept definition

- Vowel (V)
- Consonant (C)
- VC = vowel-consonant connection
- CCC = unit formed by three consonants
- Formant definition:
  - Formant frequency: the maximum of the envelope curve superimposed on the magnitude spectrum of voiced sounds (F1, F2, F3)
  - Formant bandwidth: 3dB bandwidth assigned to the magnitude spectrum corresponding to the formant frequency (B1, B2, B3)

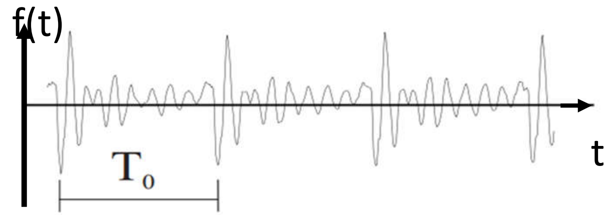


7. Spectral studies

Periodic (resonant) signals – Fourier series

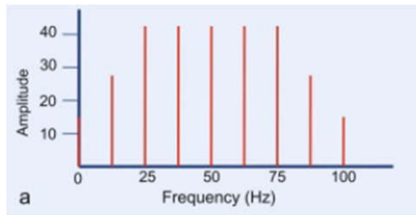
$f(t) = f(t + kT_0), \forall k \in \mathbf{Z}$

$$f(t) = c_0 + \sum_{n=1}^{\infty} c_n \cos(n\Omega_0 t + \phi_n), \Omega_0 = \frac{2\pi}{T_0}$$



Spectrum $\{n\Omega_0, c_n, \phi_n\}$:

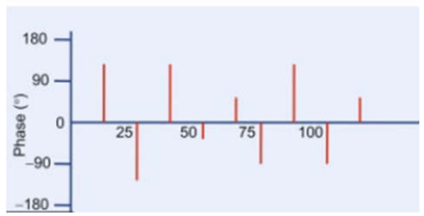
- Amplitude, Phase at any given frequency
- Periodic signals are described by a „line spectrum"



Complex form of Fourier-series:

$$\cos(x) = \frac{e^{jx} + e^{-jx}}{2} \Rightarrow f(t) = c_0 + \sum_{n=1}^{\infty}\left(\frac{c_n}{2} e^{j\phi_n} e^{jn\Omega_0 t} + \frac{c_n}{2} e^{-j\phi_n} e^{-jn\Omega_0 t}\right)$$

Introducing: $C_0 = c_0,\ C_n = \frac{c_n}{2} e^{j\phi_n}$ és $C_n^* = C_{-n} = \frac{c_n}{2} e^{-j\phi_n}$



$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn\Omega_0 t}, \text{ ahol } C_n = \int_{t_1}^{t_1+T_0} f(t) \cdot e^{-jn\Omega_0 t} dt$$

Fourier-series of a general periodic signal:

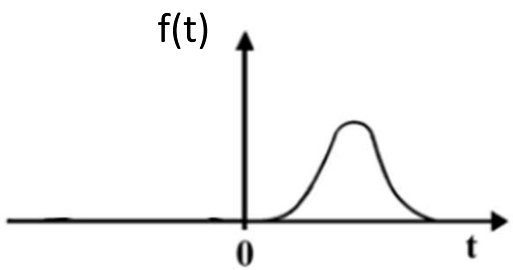
$$f(t) = \sum_{n=-\infty}^{\infty} D_n e^{j\Omega_n t} \qquad P = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f^2(t) dt$$

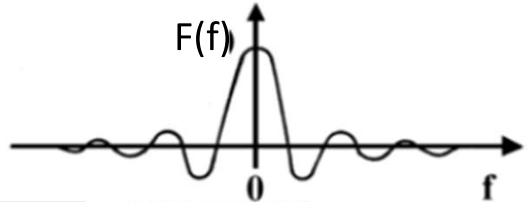
Single-shot signals (pulse excitation) Fourier transform/integral

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\, e^{-j\omega t}\, dt = \mathrm{F}\{f(t)\}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\, e^{j\omega t}\, d\omega$$

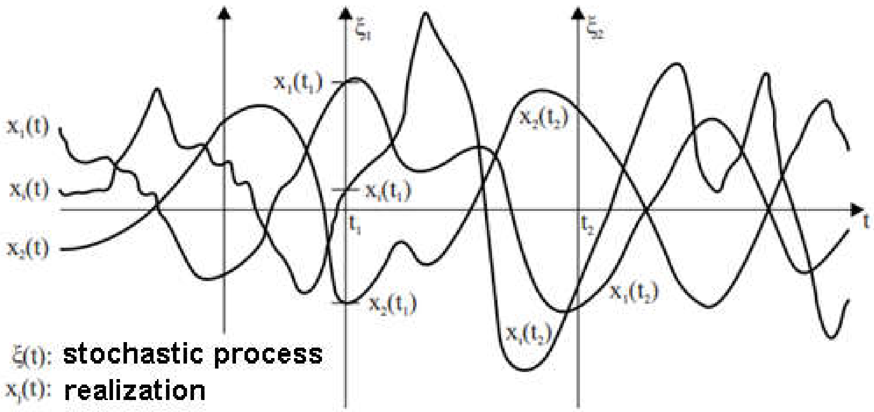
Where can we see such a signal in the figure above?



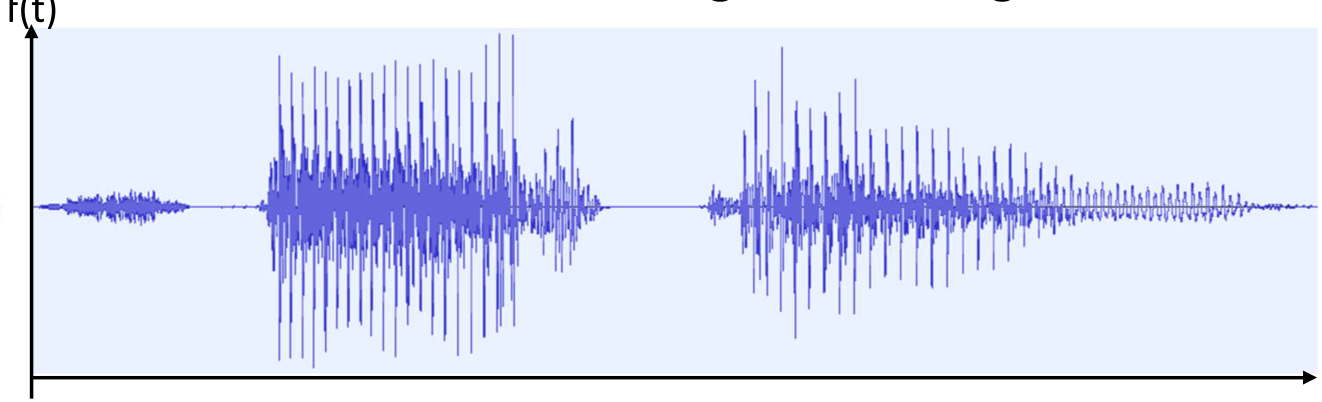


$$E = \int_{-\infty}^{\infty} f^2(t)\, dt$$

Stochastic process (noisy excitation) power density (spectrum) function



Where can we see such a signal in the figure below?



- arbitrary N-order distribution is known: P($\xi_1 < x_1$ and $\xi 2 < x2$ … and $\xi_N < x_N$) = $F^N(x_1, x_2, .. x_N, t_1, t_2, .. t_n)$
- Autocorrelation function:

$$R_\xi(t_1, t_2) = E(\xi_1, \xi_2),$$

$$R_\xi(t_1, t_2) = R_\xi(\tau), \quad \tau = t_2 - t_1$$

- Stationary stochastic process:

$$E(\xi) = \textbf{constant}$$

$$R_\xi(\tau)\ \textbf{continuos at}\ \tau = 0$$

Power density spectrum introduction

- Strongly stationary stoc . process : $F^N(x_1, x_2, .. x_N, t_1, t_2, .. t_n) = F^N(x_1, x_2, .. x_N, t_1+\tau, t_2+\tau, .. t_n+\tau)$
- Stationary processes are characterized by spectral density:

$$s_\xi(\omega) = \mathscr{F}\left\{R_\xi(\tau)\right\} = \int_{-\infty}^{\infty} R_\xi(\tau) e^{-j\omega\tau} d\tau$$

$$R_\xi(\tau) = \mathscr{F}^{-1}\left\{s_\xi(\omega)\right\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} s_\xi(\omega) e^{j\omega\tau} d\omega$$

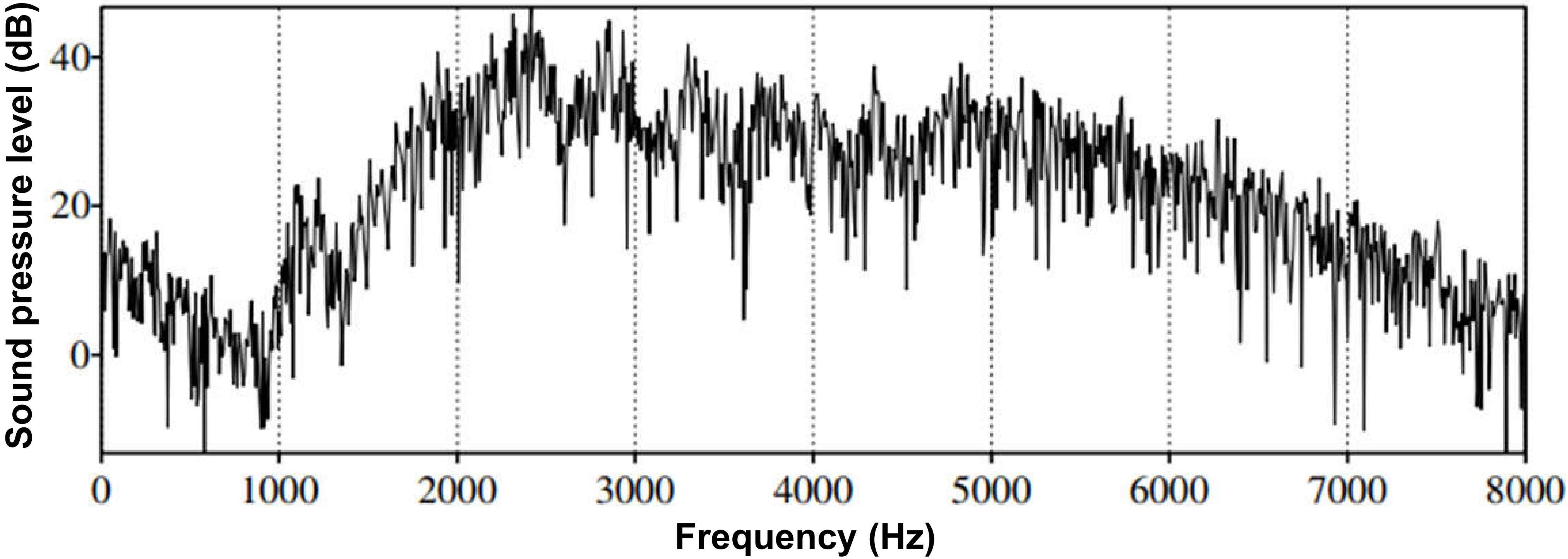
- If the stac. stoc . process is ergodic with respect to the autocorrelation function :

$$R_\xi(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x_j(t) x_j(t+\tau) dt,$$

- then the spectral density function can be interpreted as the spectral resolution of the signal powers:

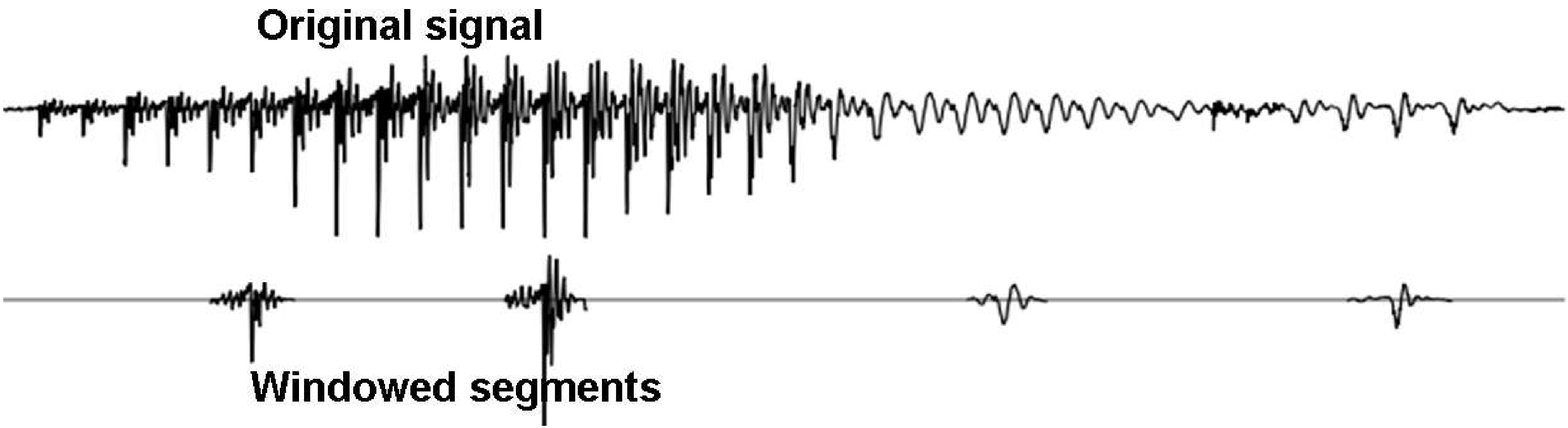
$$s_\xi(\omega) = |X(\omega)|^2$$

Power density spectrum example



Power spectral density of the [ʃ] sound from a single realization (100ms)

Windowing, example of windowed code segments

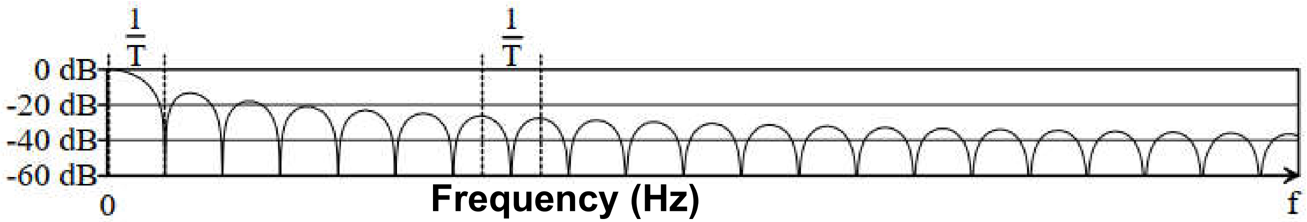


Effect of windowing: $\mathcal{F}\{f(t)\times w(t)\} = F(\omega) * W(\omega)$

Windowing (rolling spectrum)

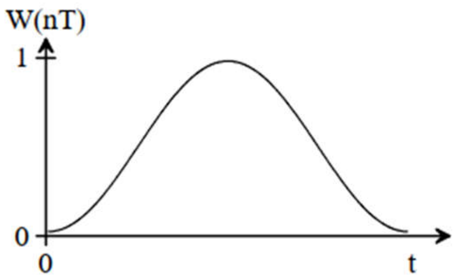
- Rectangular window, where $\hat{x}[nT] = x[nT]E[nT]$. $E[nT] = \begin{cases} 1 \text{ ,if } & i \le n \le i+N-1 \\ 0 \text{ , otherwise} \end{cases}$

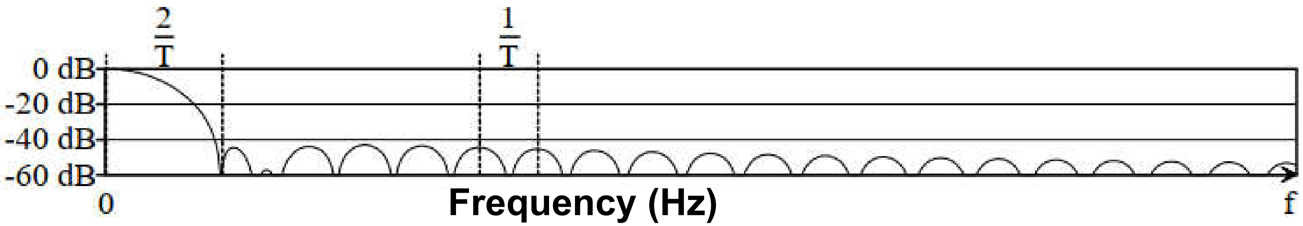
$$H_E(\omega) = NT \cdot \frac{\sin(\frac{1}{2}\omega NT)}{\frac{1}{2}\omega NT}$$



- Hamming window

$$W[nT] = \begin{cases} 0{,}54 - 0{,}46\cos(\frac{2\pi n}{N-1}) \text{ ,if } & i \le n \le i+N-1 \\ 0 & \text{, otherwise} \end{cases}$$

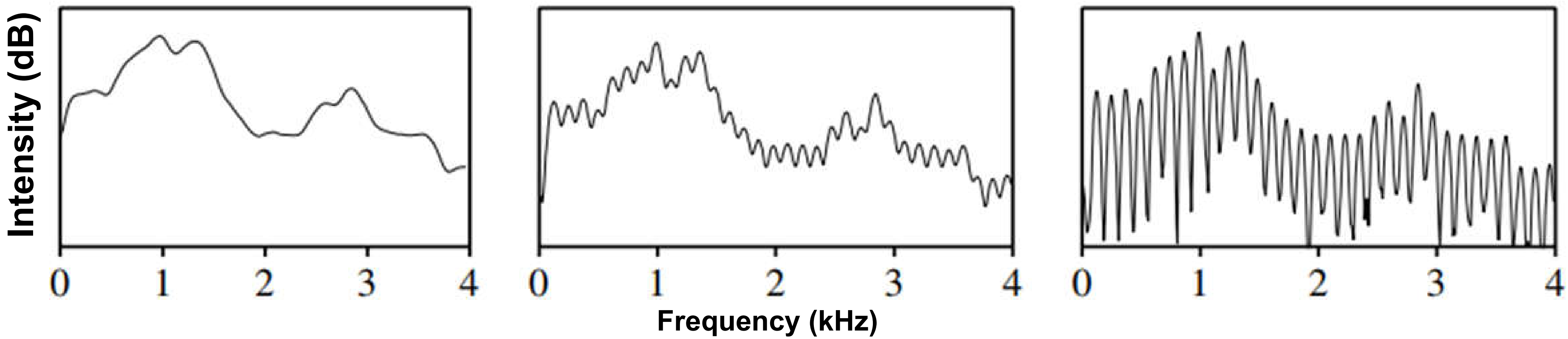




- Hann(ing) window

$$W[nT] = \begin{cases} 0{,}5(1 - \cos(\frac{2\pi n}{N-1})) \text{ ,if } & i \le n \le i+N-1 \\ 0 & \text{, otherwise} \end{cases}$$

Time and frequency domain resolution



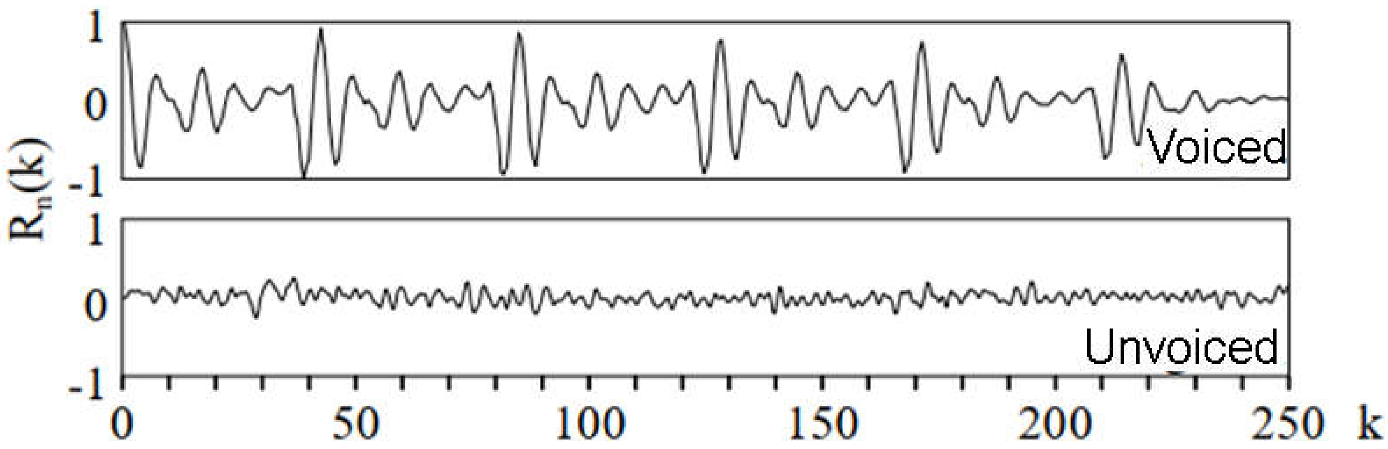
Frequency resolution limit achievable with N-point FFT : $df = f_s / N$

E.g. if the window duration is: 25ms, $f_s = 8kHz$, $N = 256$ $df = 31.25Hz$

Pitch tracker

- Autocorrelation

$$R_n(k) = \frac{1}{N} \sum_{i=n-N+1}^{n} x(i)x(i-k)$$



- Average Magnitude Difference Function (AMDF)

$$D_n(k) = \frac{1}{N} \sum_{i=n-N+1}^{n} |x(i) - x(i-k)|$$

8. Elements of Hungarian speech

Segmental structure

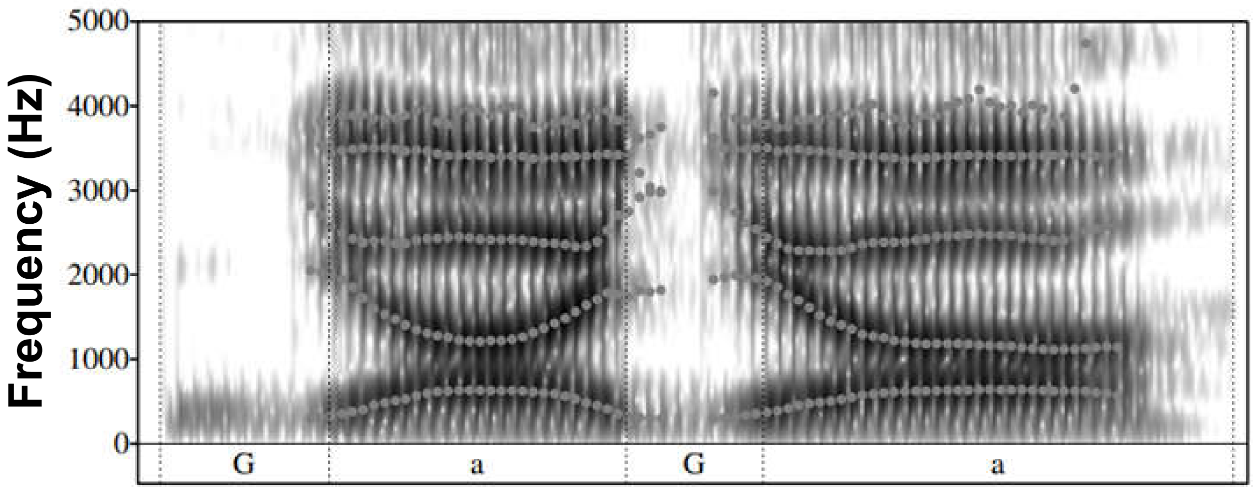
SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

BME TMIT

NVIDIA GPU EDUCATION CENTER

The Hungarian vowels

- 9 different sounds:
  + 5 short - long oppositions [aː ɔ o u i y eː ø ɛ].
  for [a:] : [a] *Fiat, strike*
- They can be characterized by three main basic articulation parameters: the position of the tongue, lips, and jaw.
- Effect of sound transitions



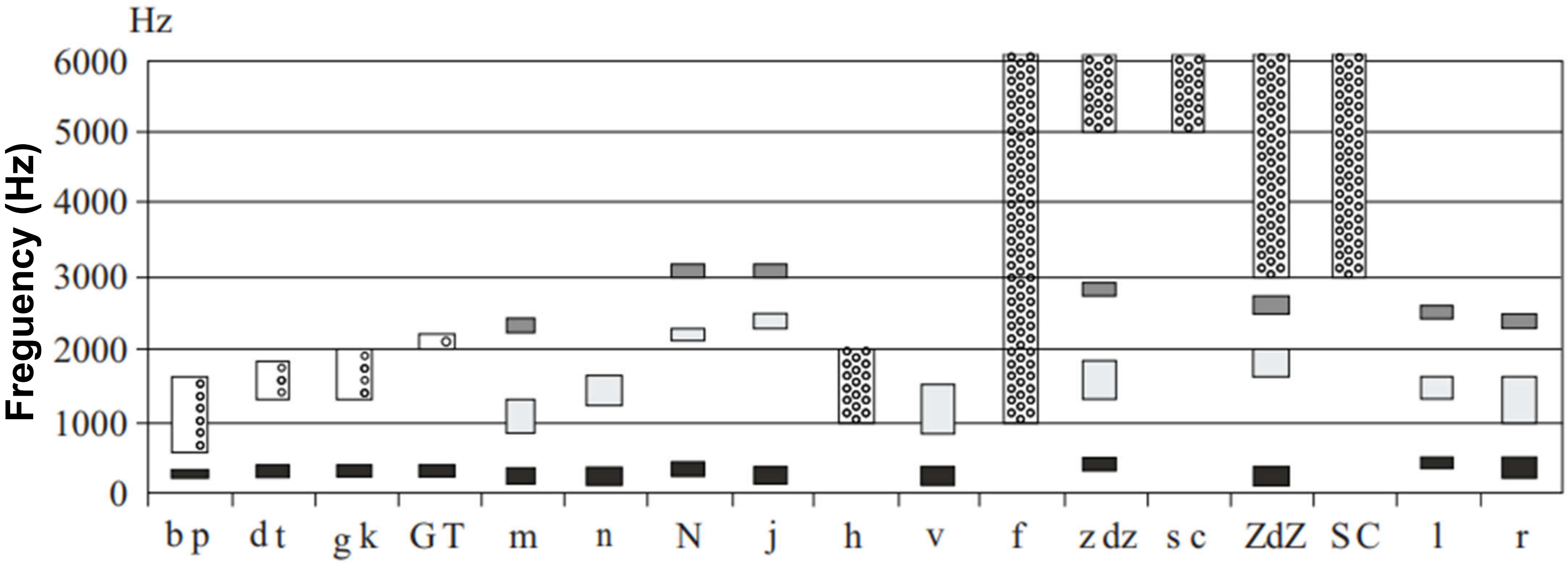
SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

BME TMIT

CENTER

The Hungarian consonants

- 55 Hungarian consonants. 25-25 short long, 5 only short



Frequency structure elements of Hungarian consonants. The dotted components represent noise, the others are voiced excitations

Hungarian consonant types

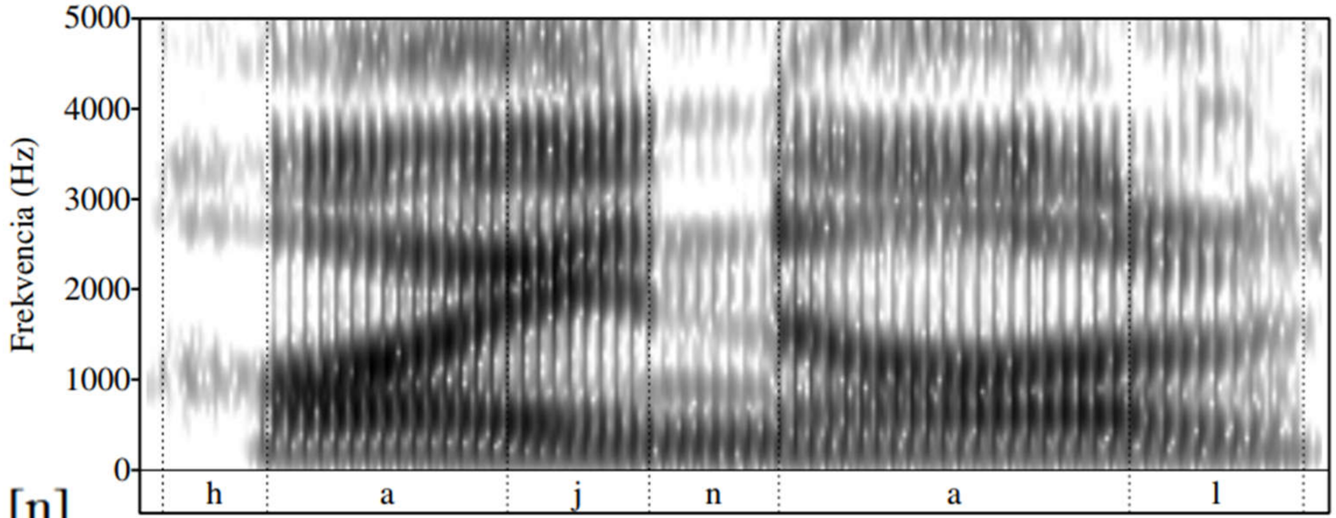
- Voiced final sounds: [b d ɟ g].
- Unvoiced plosive sounds (silence closure): [ptck]
  Burst release time (Voice Onset Time (VOT): the time elapsed from the onset of the voiceless stop sound in CV relationships to the onset of the next vowel sound, given in ms.
- Voiced fricatives: [v z ʒ], + voiced [h]
- Unvoiced fricatives: [f s ʃ h]
- Voiced affricates: [d͡z] [d͡ʒ].
- Unvoiced affricates: [t͡s], [t͡ʃ]

Special Hungarian consonants

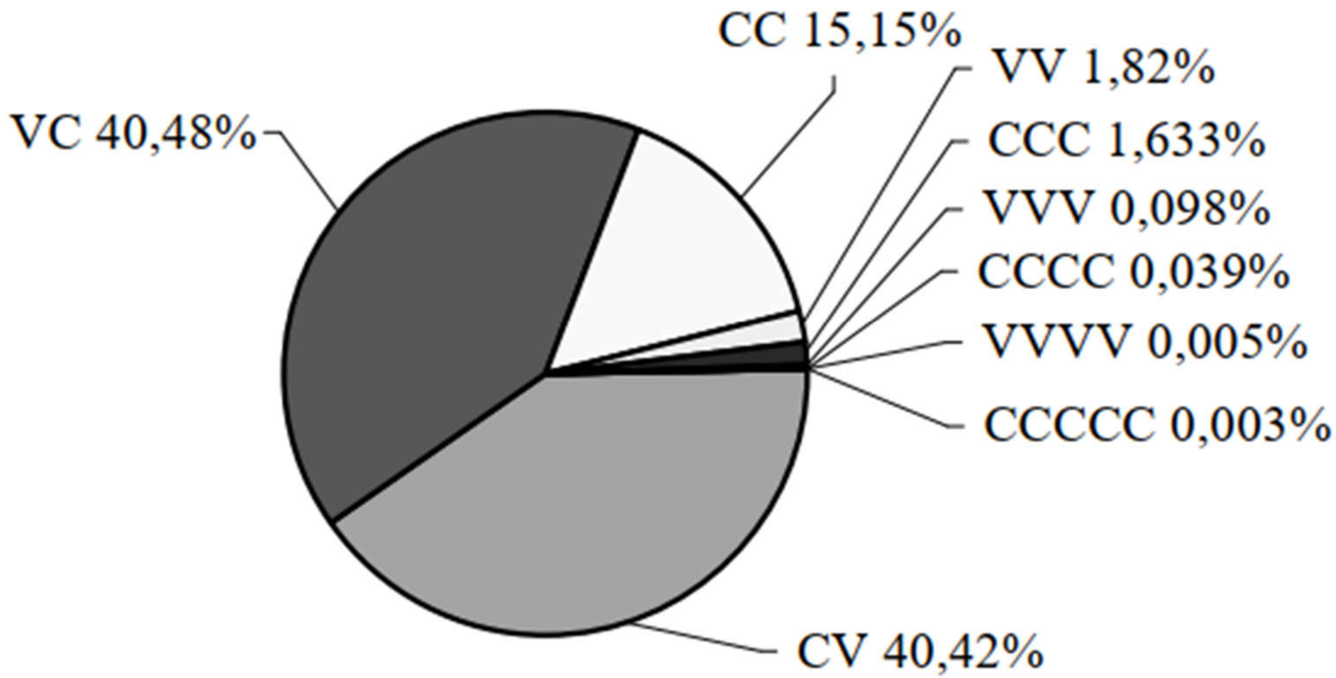
- Approximant sounds: [l], [j]

- Trill: [r]

- Nasal sounds: [m], [n], [ɲ]



Characteristics of the sound transition in the Hungarian language



Frequency of occurrence of tone sequence elements based on the texts of the Hungarian National Dictionary

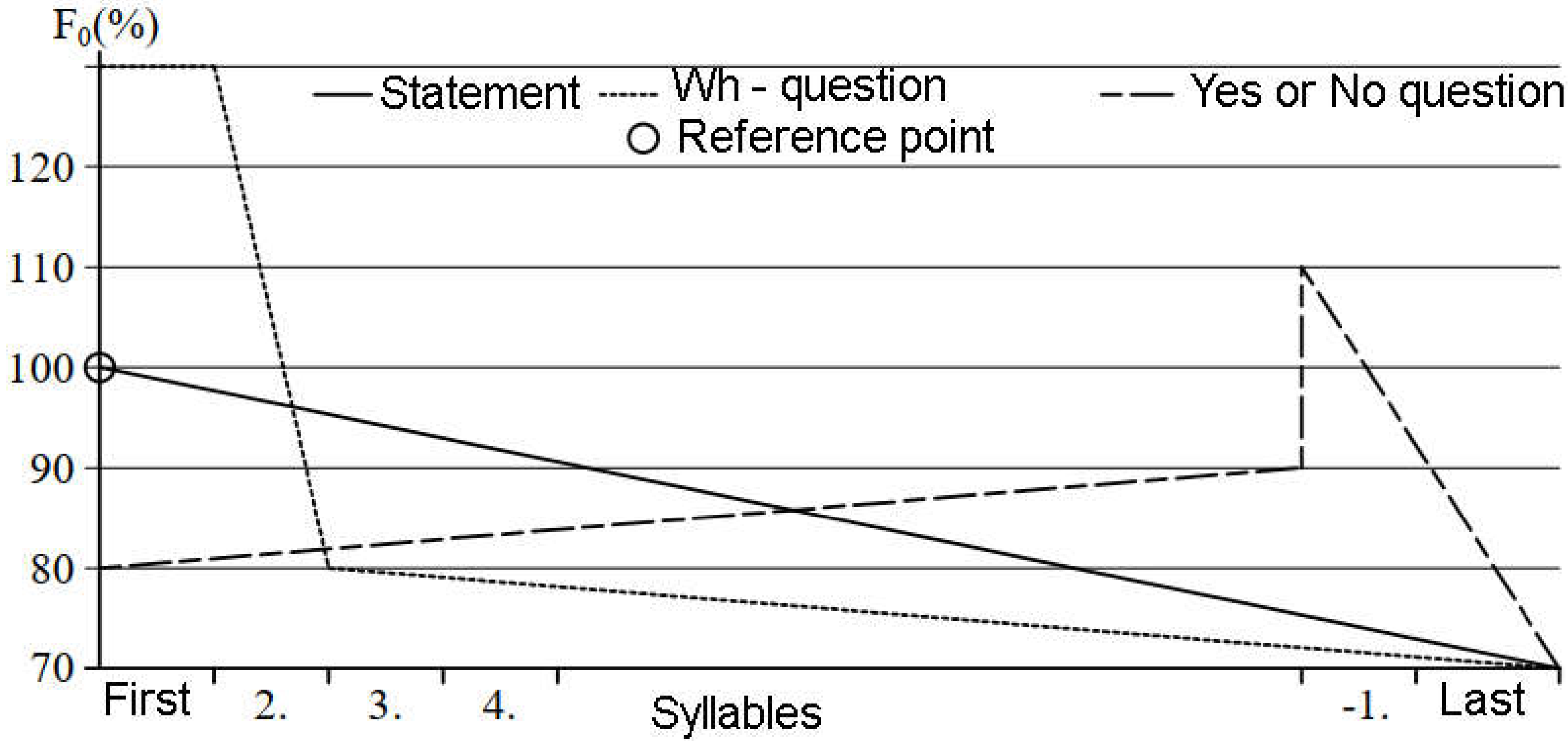
Acoustic characteristics of continuous speech

- Changes
  - Spectral
  - First harmonic (pitch , F_0)
  - Period
  - Intensity
- Approx. 13 sounds/s speech rate <> articulation rate
- Break structure determinant

The suprasegmental structure

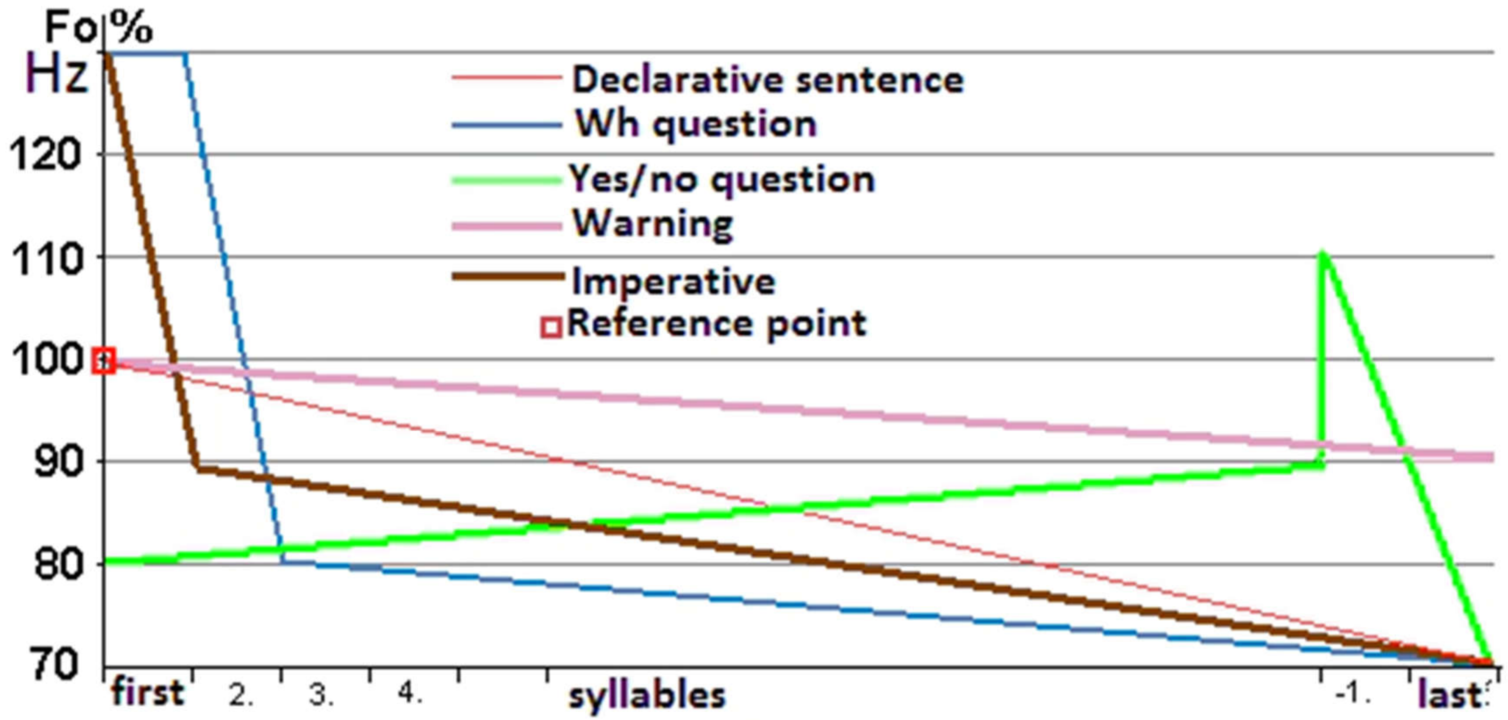
- A larger linguistic unit built on speech sounds (segments, segmental structure)
  - syllable
  - sentence
  - unity of thought
  - …………………..
- It is realized by the change in the intonation (fundamental frequency), duration and intensity of the sounds (physically F_0 is not continuous, but .... !!!)

The intonation - speech melody

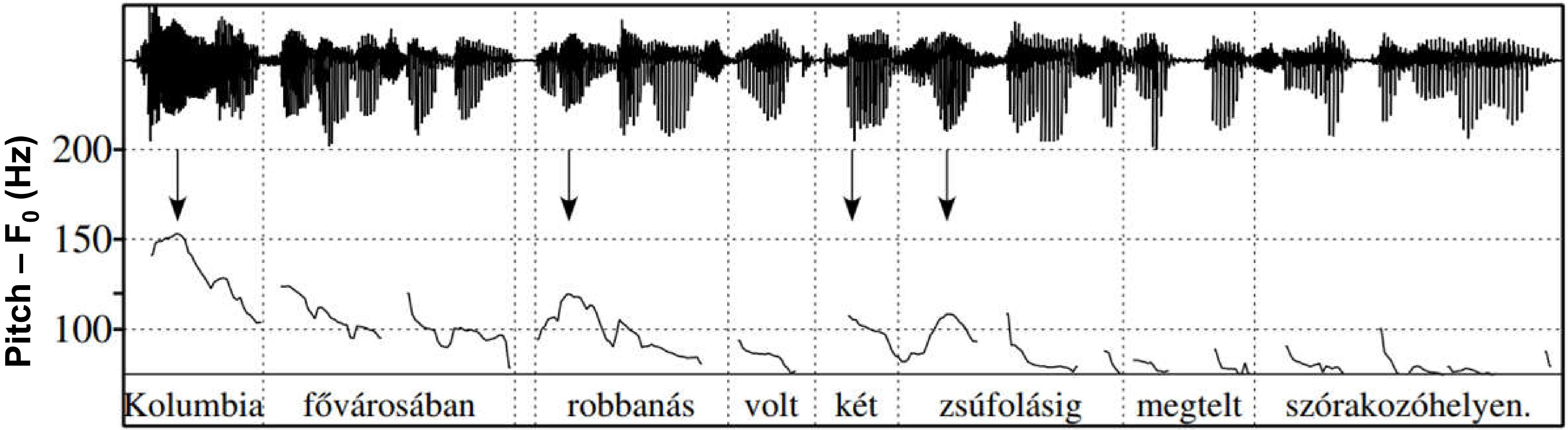


Schematic intonation/melody forms of Hungarian questions compared to statements depending on the syllable structure. The reference point is the beginning of the statement (e.g. 100% = 120 Hz)

Hungarian sentence melodies in the function of the melody of declarative sentence

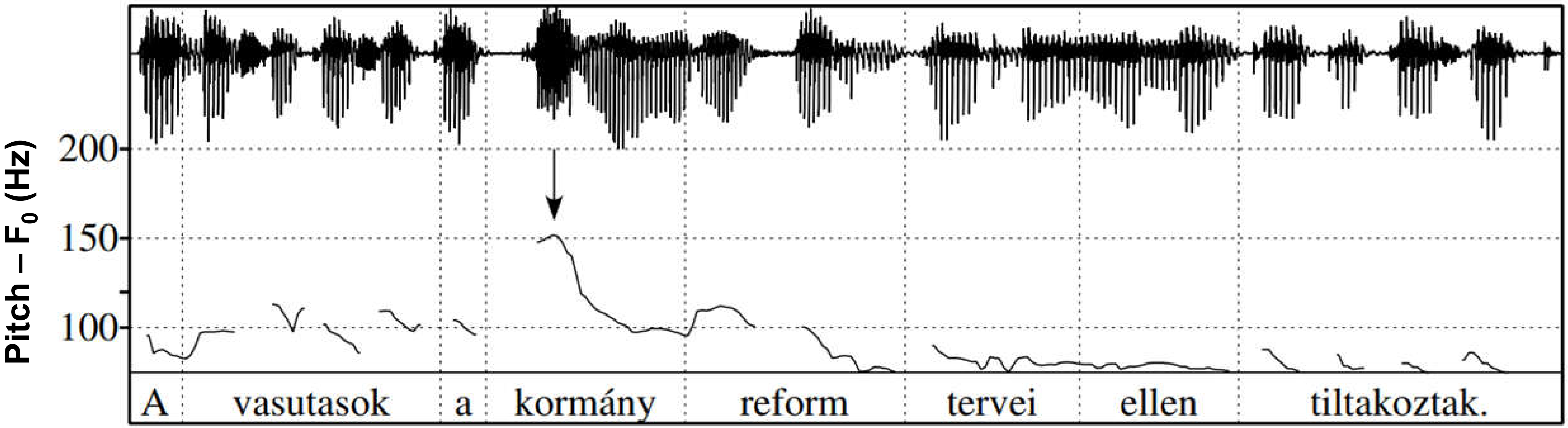


Word stress



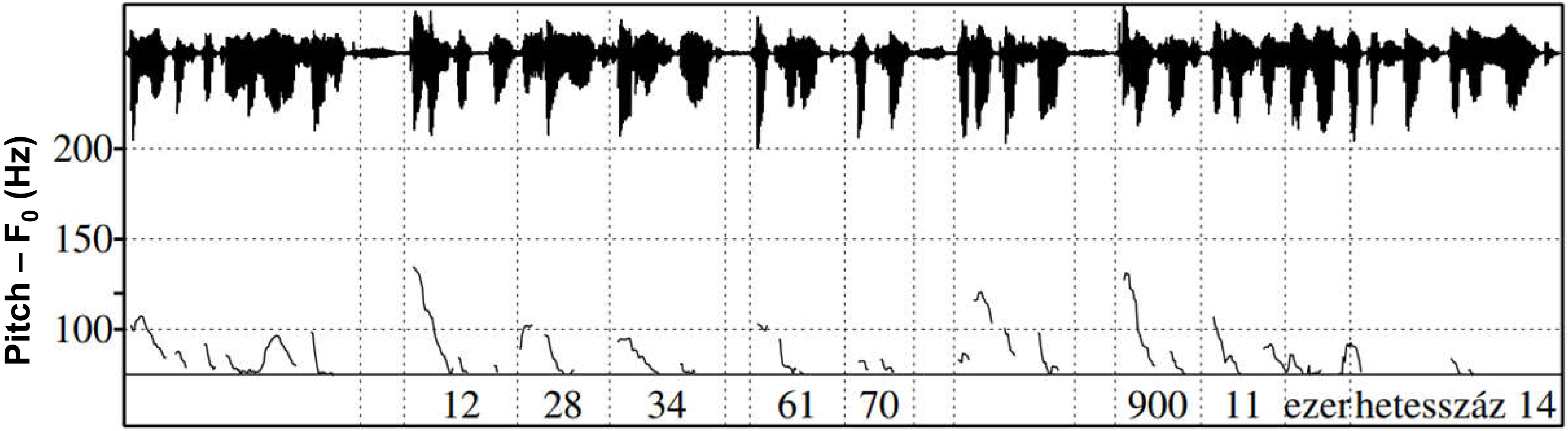
Falling intonation shapes characterize most of the sentence (4,4s).
The frequency span is larger at the beginning. Word stress is shown by F_0 peaks (and the arrows).

Sentence focus



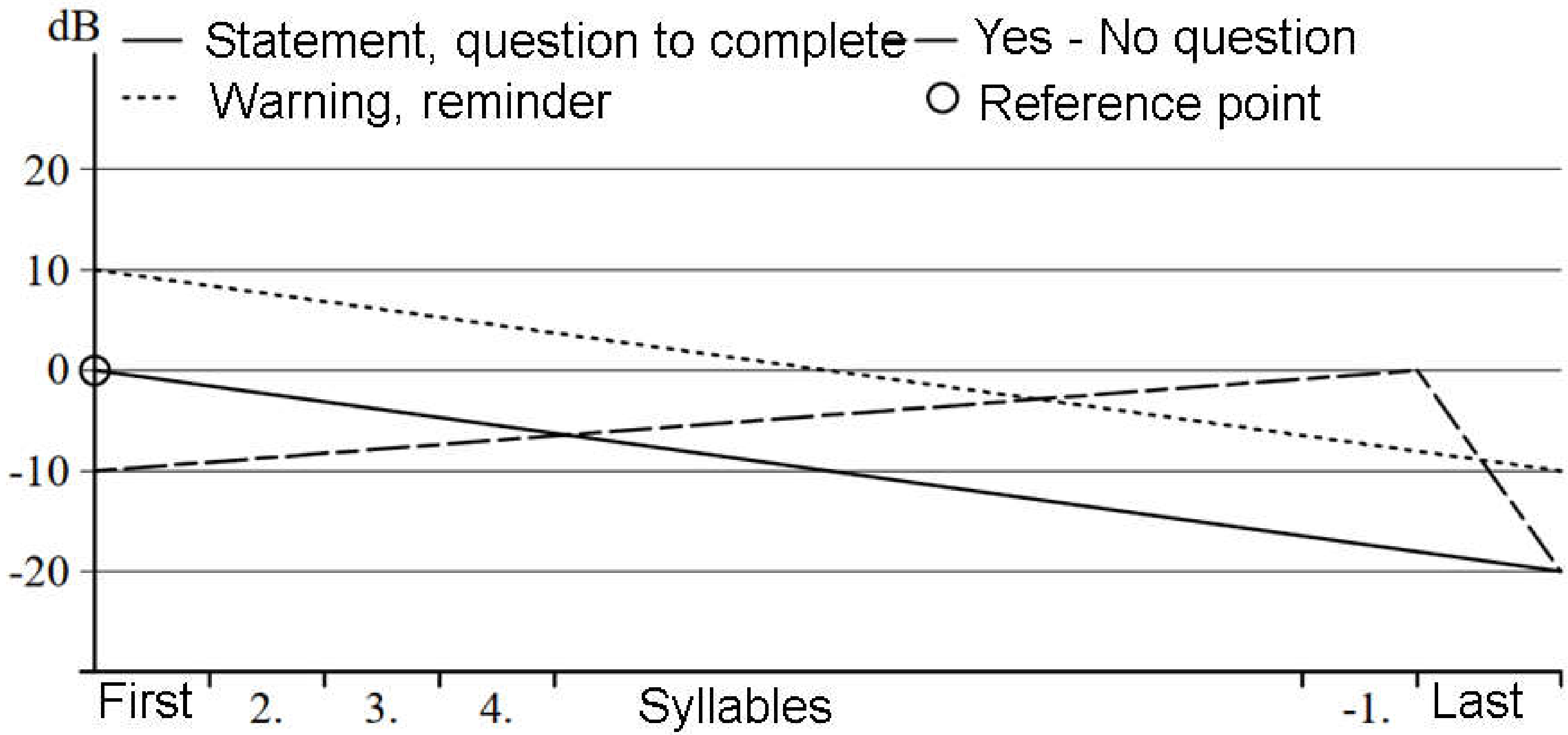
Sentence focus in a Hungarian statement (full length is 3,1s)

The role of the pause



Listing of lottery numbers. Combination of pause and intonation strategies.

The intensity structure



Intensity structure of typical sentence types
and their relationship compared to statements

The voice character

- Humans are good at identifying speakers based on their voices
  - Personal identification based on voice cannot be compared to, for example, fingerprints
  - Defining characteristics
    - Voice waveform/ glottal characteristics
    - Articulation channel parameters
    - Segmental and suprasegmental individual characteristics (these can be learned/copied quite well, see Stand- up)