Speech information systems

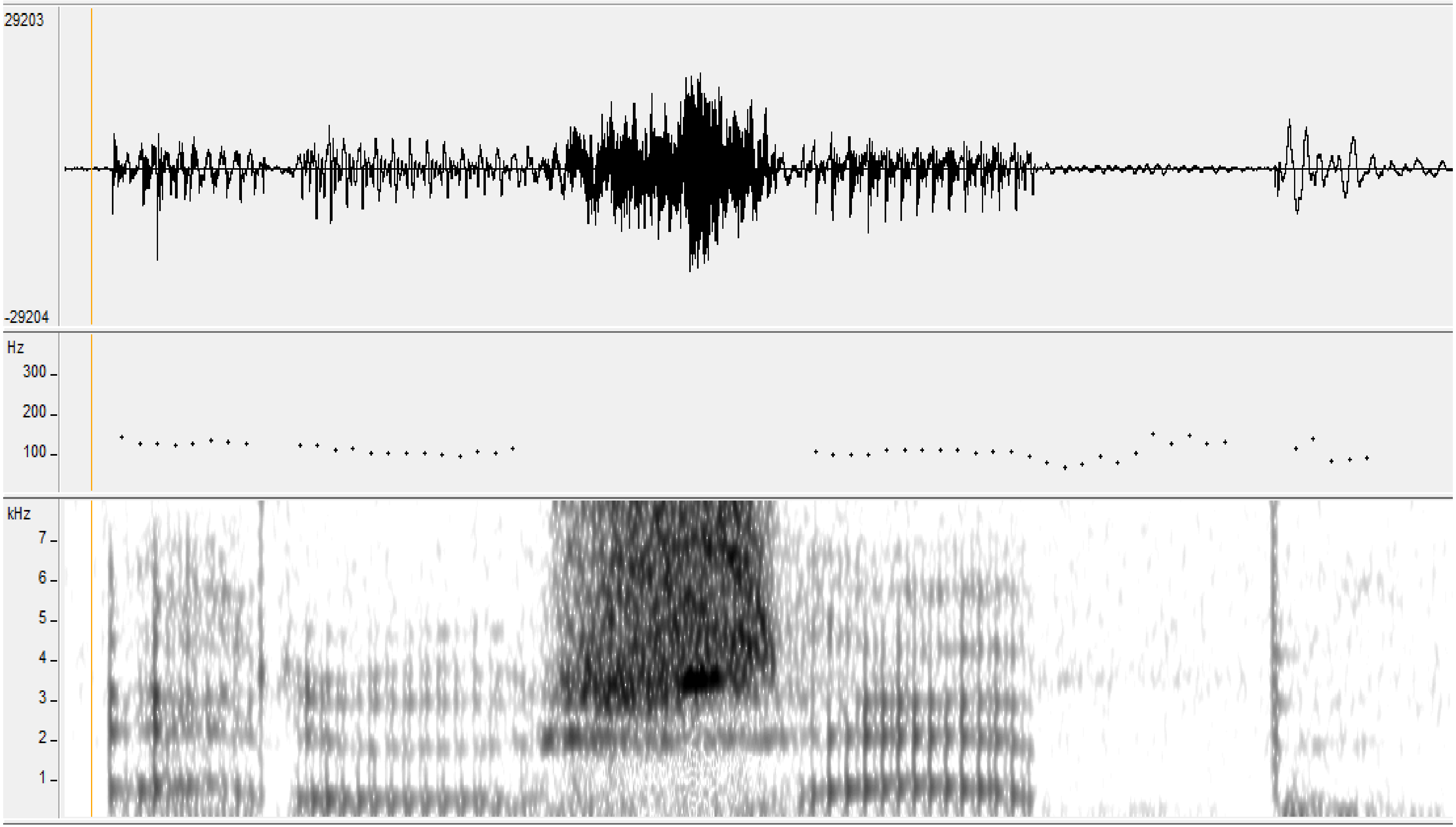
Speech processing

Gabor Kiss: kiss.gabor@vik.bme.hu

David Sztahó : sztaho.david@vik.bme.hu

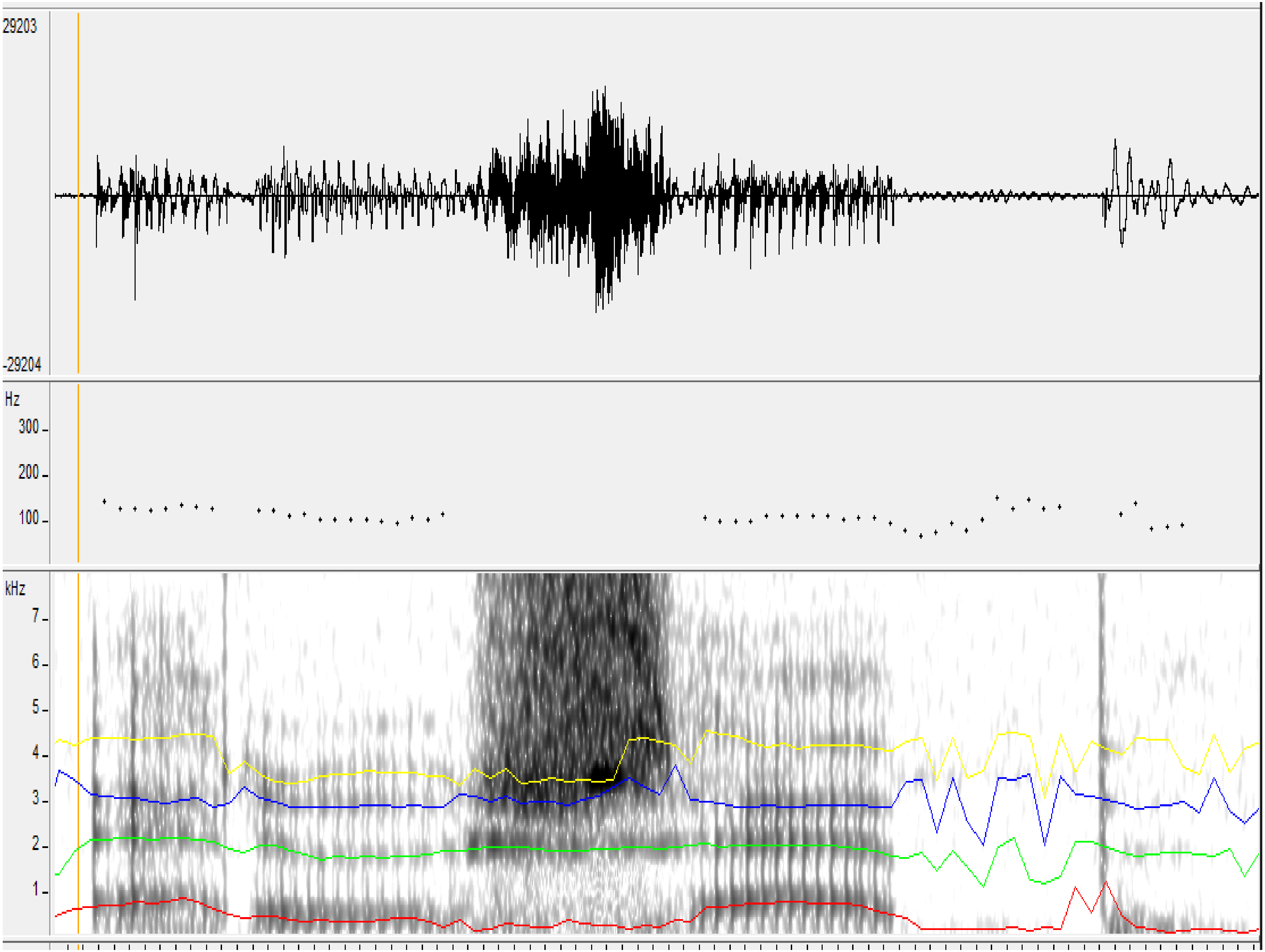
Today's practice material

- Speech analysis
- Fundamental frequency, formant measurement
- SAMPA character encoding
- Databases



1 : The figure shows a speech fragment.

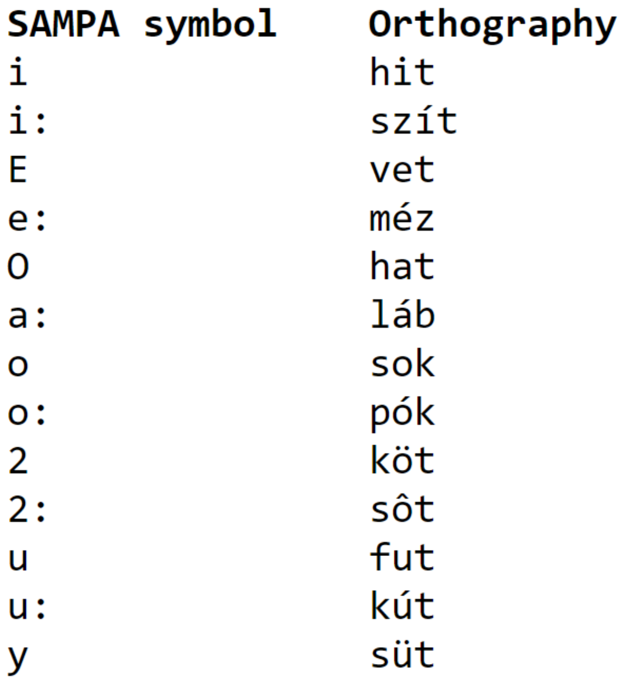
What is the f_0 of the first vowel? Could the speaker be a man or a woman? Are there any formants visible anywhere? If so, where, if not, why not?


29203
-29204
Hz
300
200
100
kHz
7
6
5
4
3
2
1

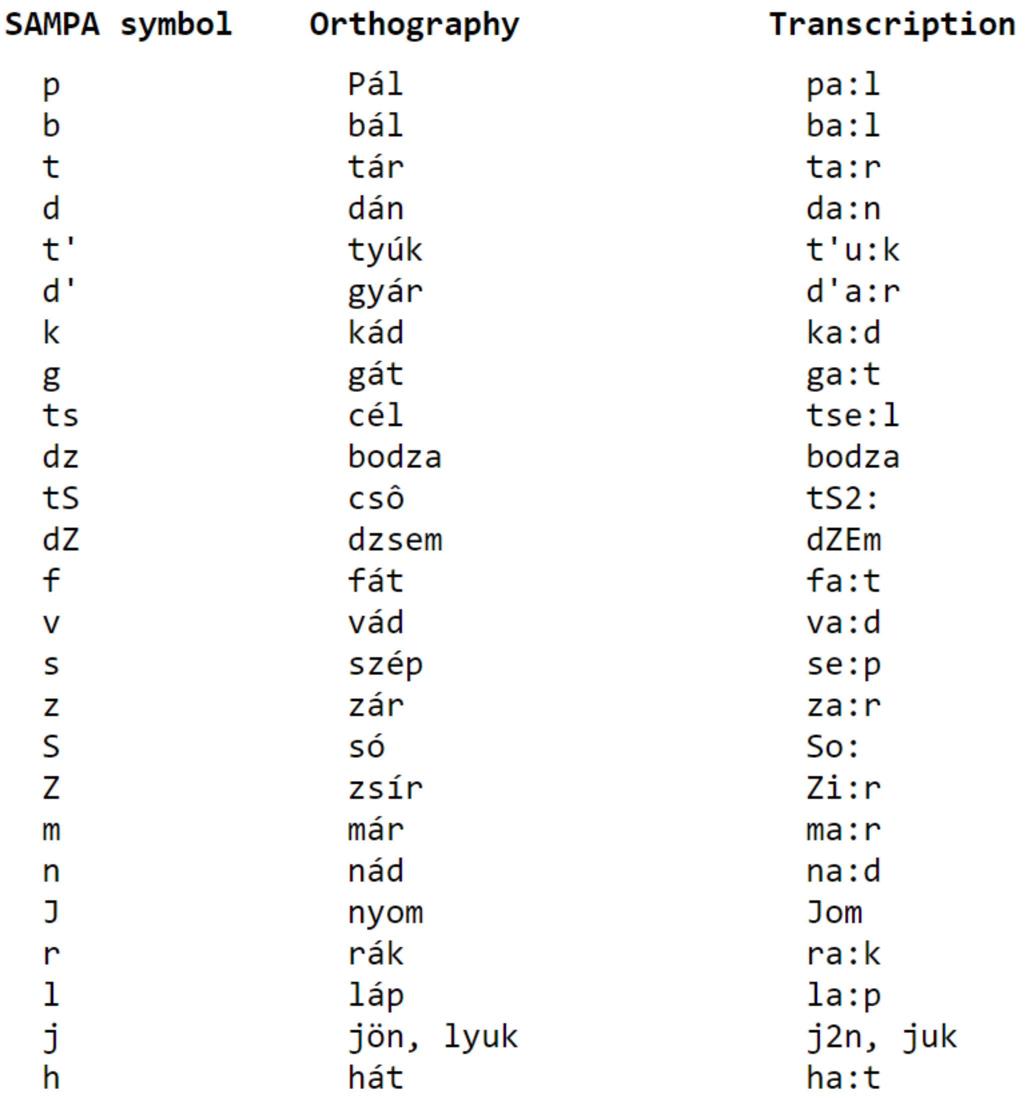
f_0 - measurement

2 : In a telephone system, the following harmonics we managed to measure: 400 [Hz], 800 [Hz], 1000 [Hz]. What can be the fundamental frequency?

- Solution options
  - 150Hz
  - 100Hz
  - 200Hz
  - 400Hz
  - 20Hz
  - 80Hz



SAMPA symbol	Orthography
i	hit
i:	szít
E	vet
e:	méz
O	hat
a:	láb
o	sok
o:	pók
2	köt
2:	sôt
u	fut
u:	kút
y	süt



SAMPA symbol	Orthography	Transcription
p	Pál	pa:l
b	bál	ba:l
t	tár	ta:r
d	dán	da:n
t'	tyúk	t'u:k
d'	gyár	d'a:r
k	kád	ka:d
g	gát	ga:t
ts	cél	tse:l
dz	bodza	bodza
tS	csô	tS2:
dZ	dzsem	dZEm
f	fát	fa:t
v	vád	va:d
s	szép	se:p
z	zár	za:r
S	só	So:
Z	zsír	Zi:r
m	már	ma:r
n	nád	na:d
J	nyom	Jom
r	rák	ra:k
l	láp	la:p
j	jön, lyuk	j2n, juk
h	hát	ha:t

SAMPA table

https://en.wikipedia.org/wiki/SAMPA_chart

Speech producing acoustics

3 : Analyze any word fragment according to the following criteria.

a) Write the pronunced sound in capital letters.

b) Perform character-to-sound conversion and enter the pronounced sound sequence into one of the learned sound code sets.

c) Specify the acoustic product for each sound (voicing, friction noise, stop release noise).

d) Indicate which note has a suppressed voicing.

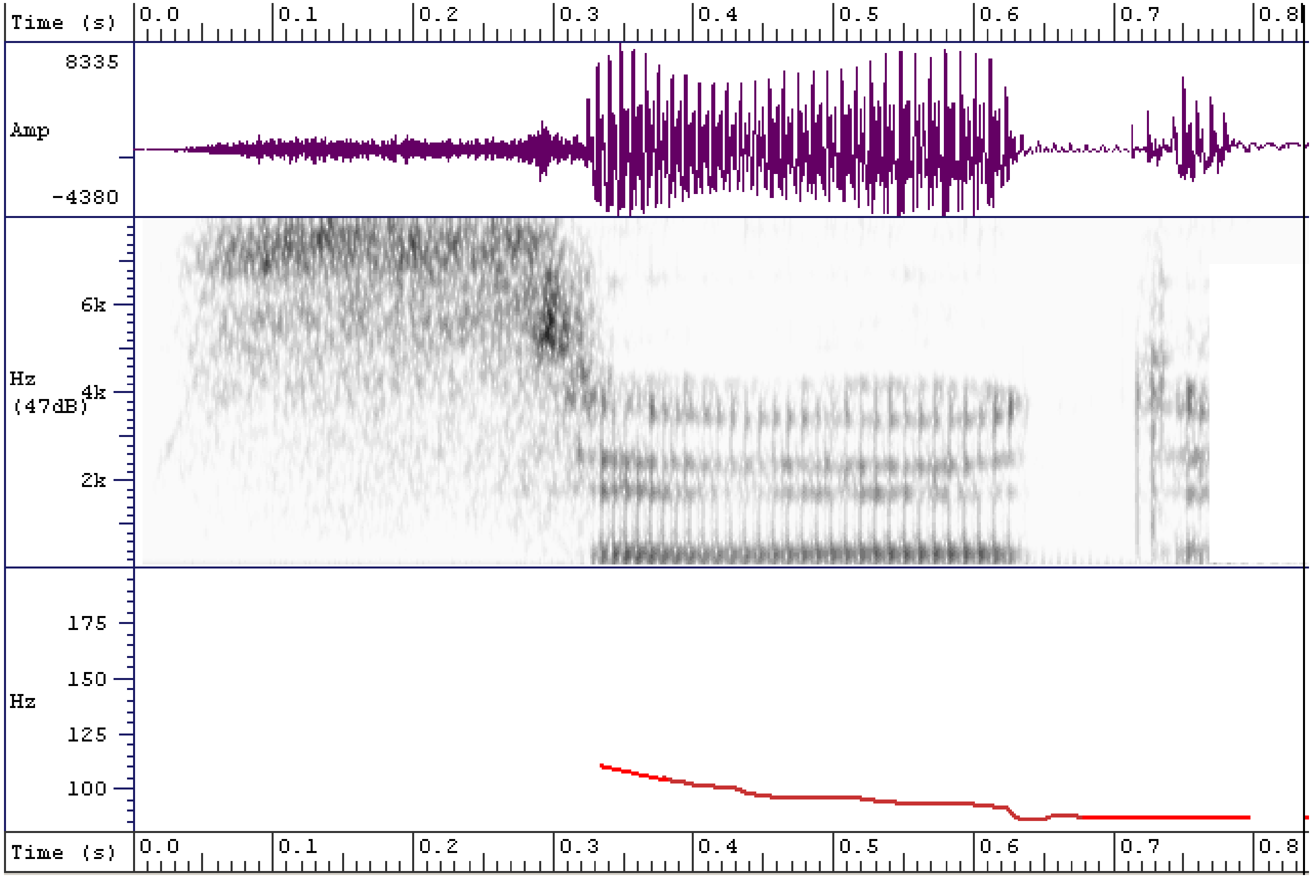
e) Indicate which sound has F1, F2 and F3 formants.

f) Classify sounds separately using at least 3 speech sound classification methods form according to the Gordos circles.

vovwels/consonants
turbulent flow noise: yes/no
voiced/unvoiced
impulse noise (stop sounds)

g) Draw the approximate intensity-time function of the sound pressure change of the uttered sounds . Also mark the sound boundaries . Time division: 100 ms = approx. 2 cm.

h) What is the average fundamental frequency when you pronounce your own name? Give reasons!



4 : The figure shows the time function (upper band) and spectrogram of a word consisting of three speech sounds. (middle band) and fundamental frequency curve (bottom band) visible.

- Draw the boundaries of speech sounds. Explain! Describe the characteristics of speech sounds based on what you have learned.
- Try to determine the audio sequence. Explain!
- For each note, specify the excitation type!
- female, male, or child speech on the spectrogram?

Transcription - segmentation

5 : What does it mean to annotate the text material of a speech database and what does it mean to segment it?

- *Annotation:* general information about the content of the audio recording is provided in a label file. E.g. part-of-speech tags, orthographic notation, canonical form

- *Segmentation:* separation of parts with signals/event boundary markers in the time function of speech, where the specific properties change in the audio material.
  - ❑ Most often at the phoneme, word and sentence level

6 : Complete the phonotypification of the text below using the SAMPA symbol system!

- He said he was going to rest.

- hi s{d hi w{z goUIN t@ rEst

- Breakdown:
  - He: /hi/
  - said: /s{d/
  - he: /hi/
  - was: /w{z/
  - going: /"goUIN/
  - to: /t@/
  - rest: /rEst/

Annotation

- **What parameters are recorded during annotation? What is the purpose of this? significance?**
  - ❑ A specific task requires the labeling of different types of information: e.g. emotions, prosodic features, stress.
  - ❑ It is also important for training acoustic-phonetic models.
- **What noises should be marked when annotating? Give a typical noise grouping for annotation!**
  - ❑ Speaker-source: e.g. coughing, swallowing, mouth movement noises
  - ❑ Environmental noises: e.g. vehicles, doors creaking, birds chirping
- **What is intra-speaker and inter-speaker variation? What do we do with them in speech synthesis? And what upon speech recognition?**

Annotation

- **Variation of speech**

= variation of the characteristic physical parameters of speech

- It affects the physical characteristics of speech sound
  - the size of the vocal production organs
  - tension of the control muscles
  - elasticity of the cavity walls.
  - environmental, acoustic conditions (e.g. noisy environment, echoes)

Annotation

Intra-speaker variation

- *The set of constantly changing movements during sound production.*
  - **Coarticulation effect:** Due to continuous vocal organ movements, the physical properties of one sound affect the physical properties of the sounds that precede and follow it
  - **Formant frequencies** of the same speech sounds produced in different sound environments differ from each other.
  - Differences in **rhythm, volume, pitch, intonation, and stress**
  - A **cold** greatly changes the acoustic parameters of sounds, as the resonance frequencies of fluid-filled cavities shift.
- *Environmental **effects**, excitement, surprise, etc. also affect the acoustic properties of the produced speech.*

Annotation

- **Inter-speaker variation**
- **Biological factors**, e.g. the difference in size of speech organs (between men, women, children, but also within a group)
- **Linguistic differences, pronunciation differences between** groups of people belonging to a language community .

→ Intra-speaker: speech synthesis

- ❑ The essential feature of the database used is the collection of material with the greatest possible phonetic variation.
- ❑ The material emphasizes the microsegmental characteristics of speech. Few, but precisely modeled speakers are needed.

→ Inter-speaker: speech recognition

- ❑ For training speaker-independent recognizers
- ❑ Large variety of speaking styles and recording conditions is required

Databases - general

- What could be the reason why a German - language speech synthesis voice database containing 50 voices was created in 4 different versions, 7 Mbytes, 3.5 Mbytes, 2.54 Mbytes and 1.27 Mbytes in size ?
  - Values are double each other. This could be an 8 and 16 bit or kHz version. The other is that the smaller one takes up less space and is faster, but it comes at a price in terms of quality.
- What requirements would you have for a speech synthesis system and what requirements would you have for a database used for speech recognition?
  - Synth: professional announcer, announces all sounds and voice connections in many variations
  - ASR: many announcers, no need to clean up noises or errors if it fits into normal usage conditions
- There is a speech database containing 15 speakers, 1.5 hours per speaker, with varied text, phonetically transcribed speech database recorded over the phone. What is this suitable for and what is it not? Give justification to your answers!
  - Speech synthesis using concatenation technique
  - Speech recognition for office dictation
  - Speech recognition in a callcenter
  - Speech recognition adaptation

Databases - general

- What is the purpose of speech database?
  - **Speech database:** a large, linguistically and acoustically processed, stored set of audio data, with explanatory notes, labels, and transcriptions. It allows for the study of the variability and invariance of speech.
  - Different sampling frequency and quantization according to requerments
  - E.g. good sound quality vs. language independence

- **What does it mean that the database is phonetically rich? What is triphone? What do they have to do with each other?**
  - **Phonetically rich sentences –>** audio material that meets the teaching criteria of phoneme-based recognizers should be sufficient for Hungarian;

❑ the phonemes and phoneme relationships (e.g. bi- and triphones : double, triple sounds – most often CVC-) occurring in Hungarian occur in sufficient numbers

  - **The speech database must contain all phonemes and phoneme relationships occurring in the language, providing statistically adequate coverage.**
  - **CVC sound sequences: ‘alma’ Hungarian word contains ‘_al’ and ‘ma_’ triads and the ‘lm’ dyad**

Thank you for your attention!

Questions?
Test!

Speech processing

https://forms.office.com/e/sXy7EXE3gg