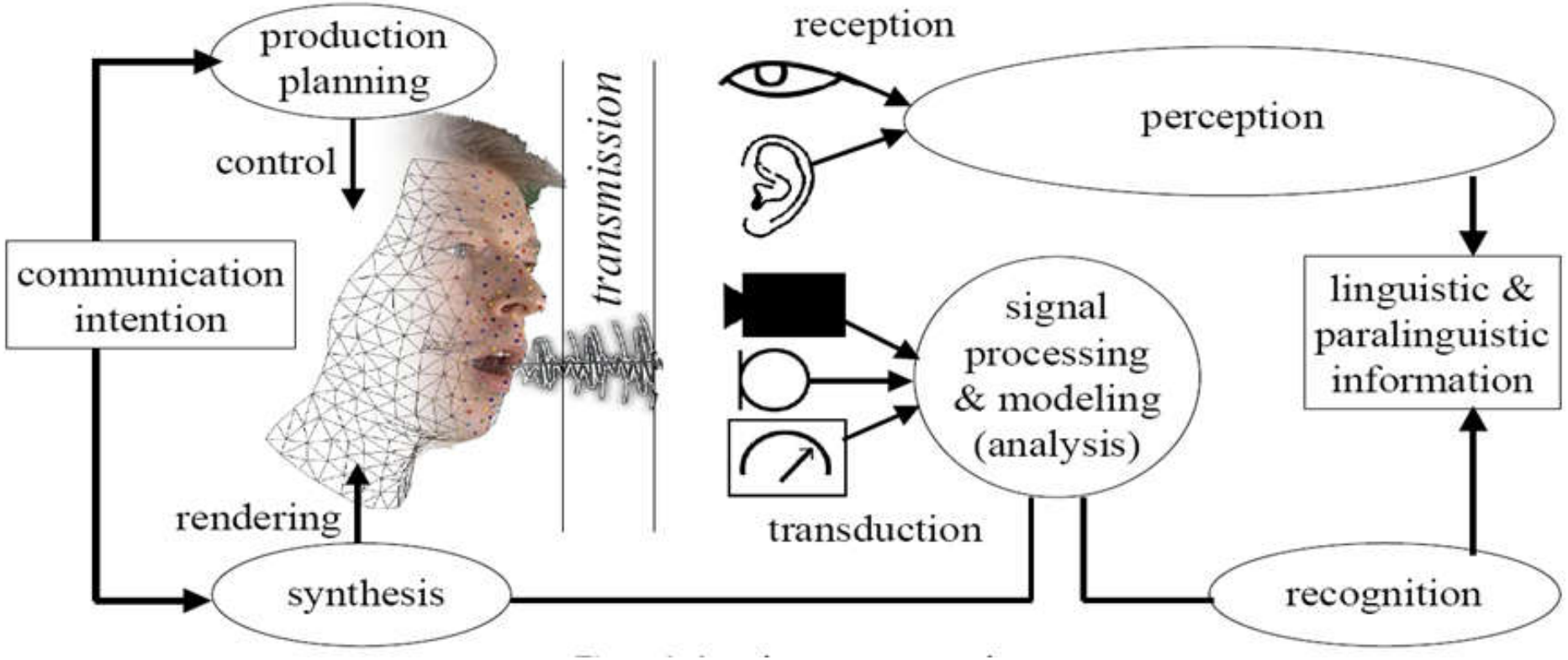
Human-Computer Interaction

BMEVITMMA11

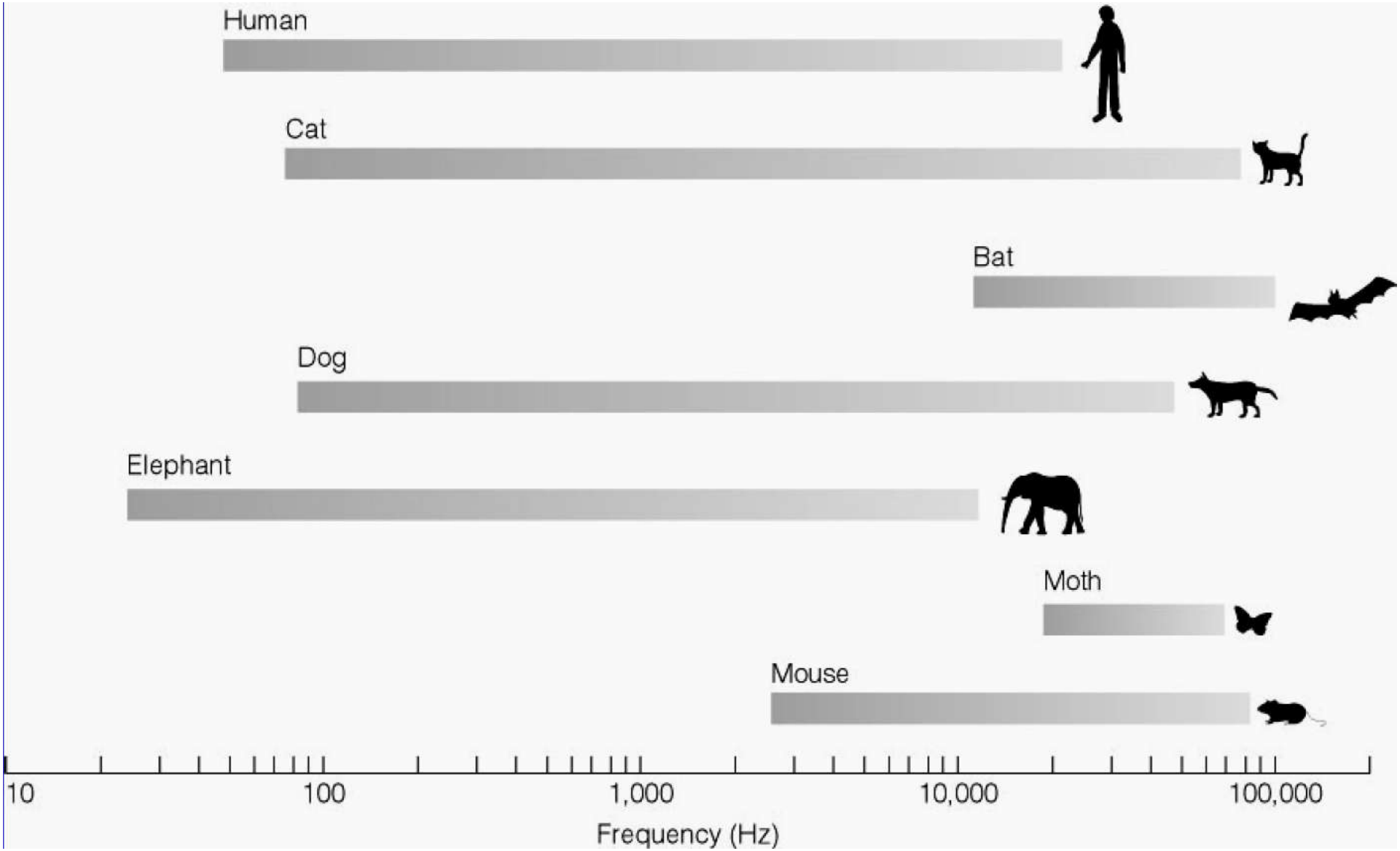
Natural speech communication chain

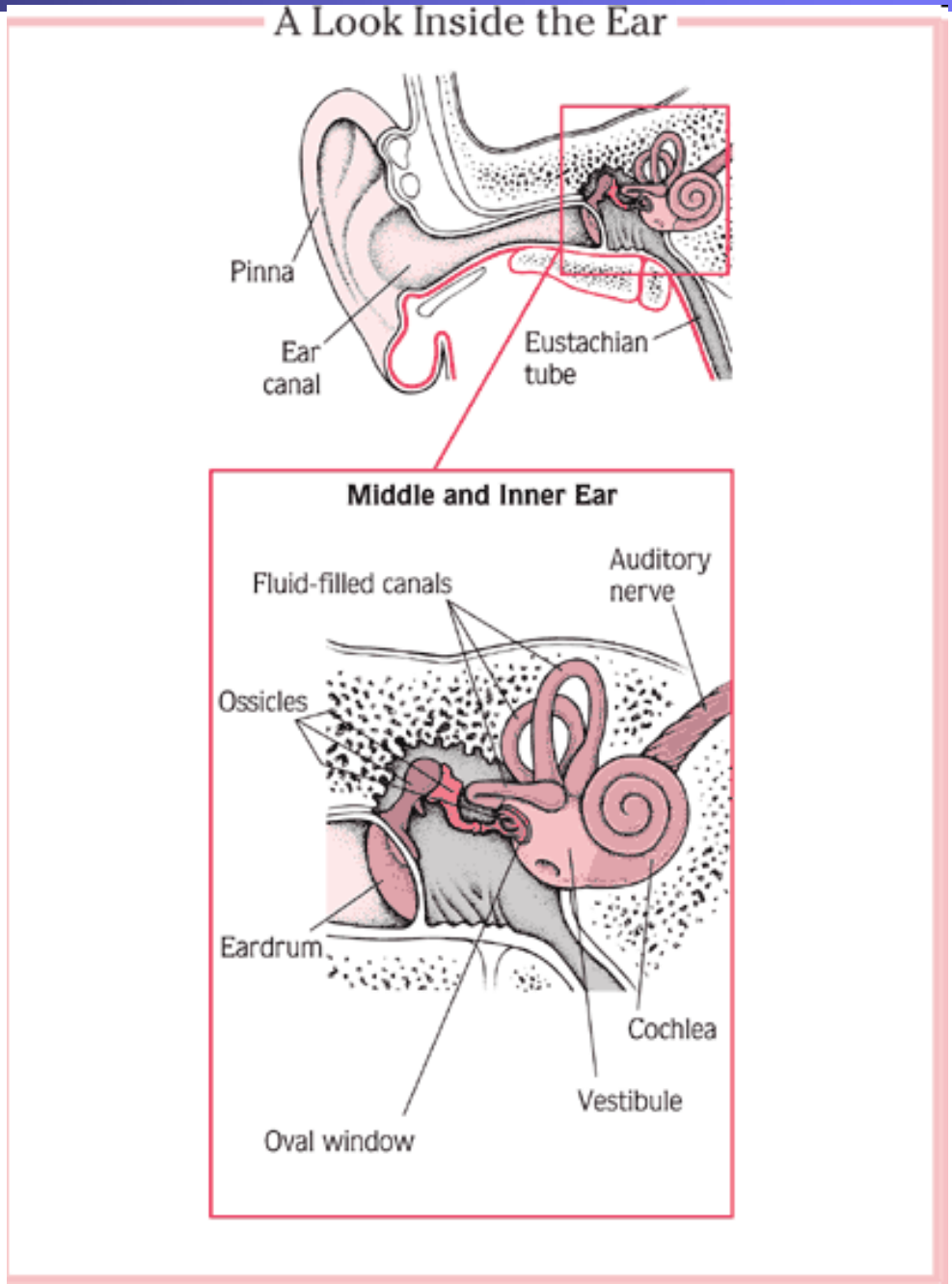


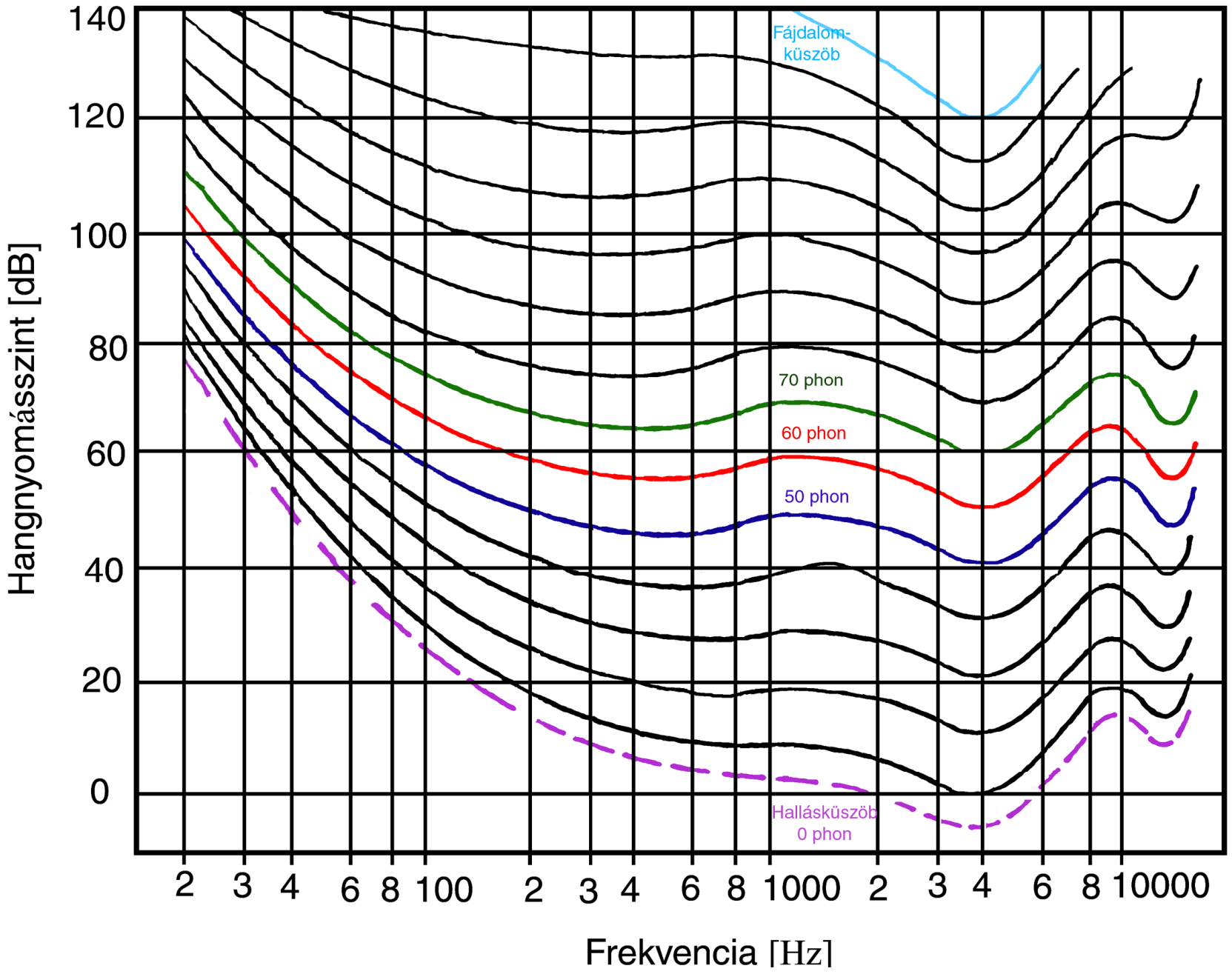
Speech (SUI, VUI ⇔ GUI)

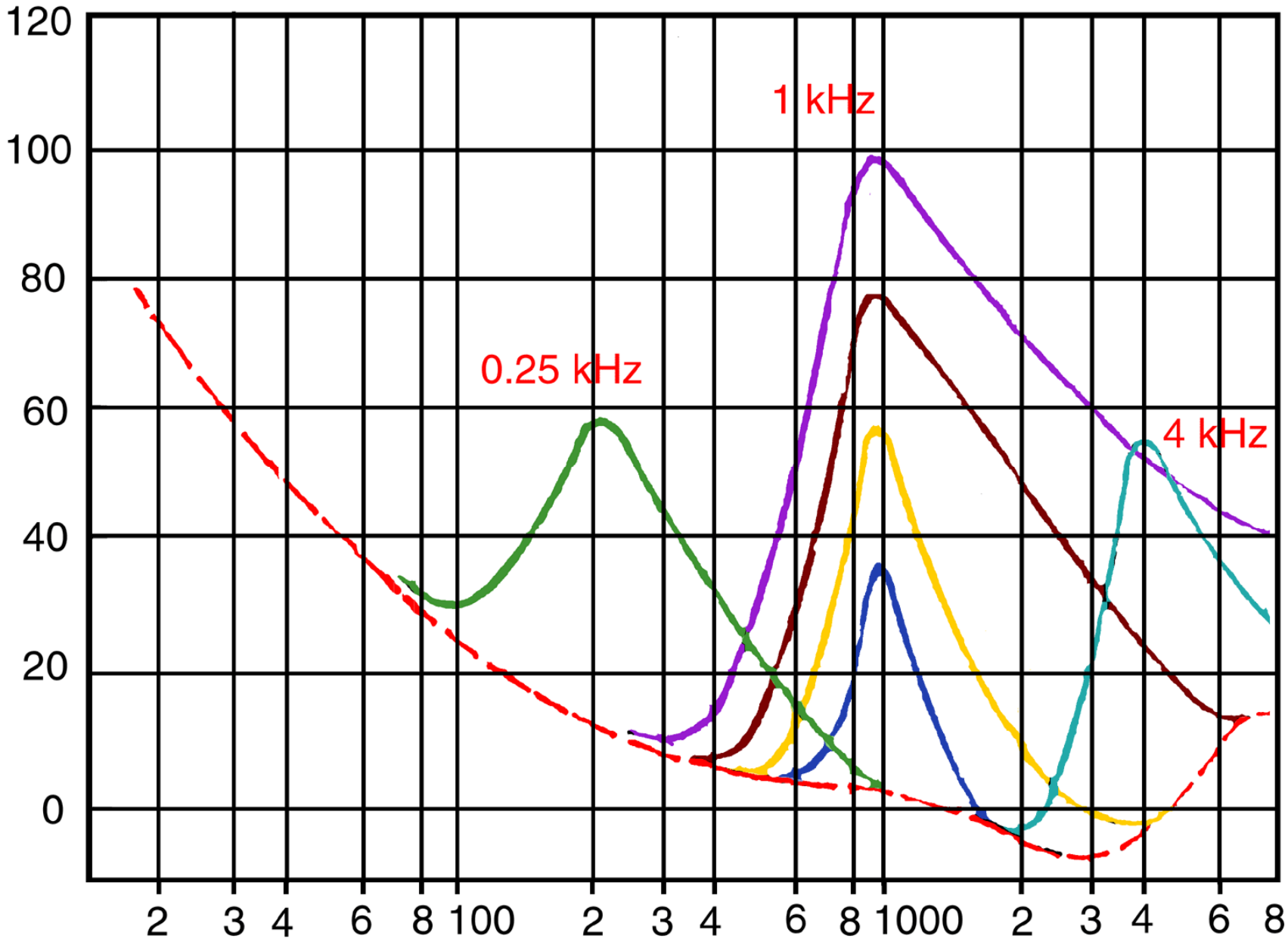
- Sounds
- F0
- Formant frequency
- Time structure

Hearing

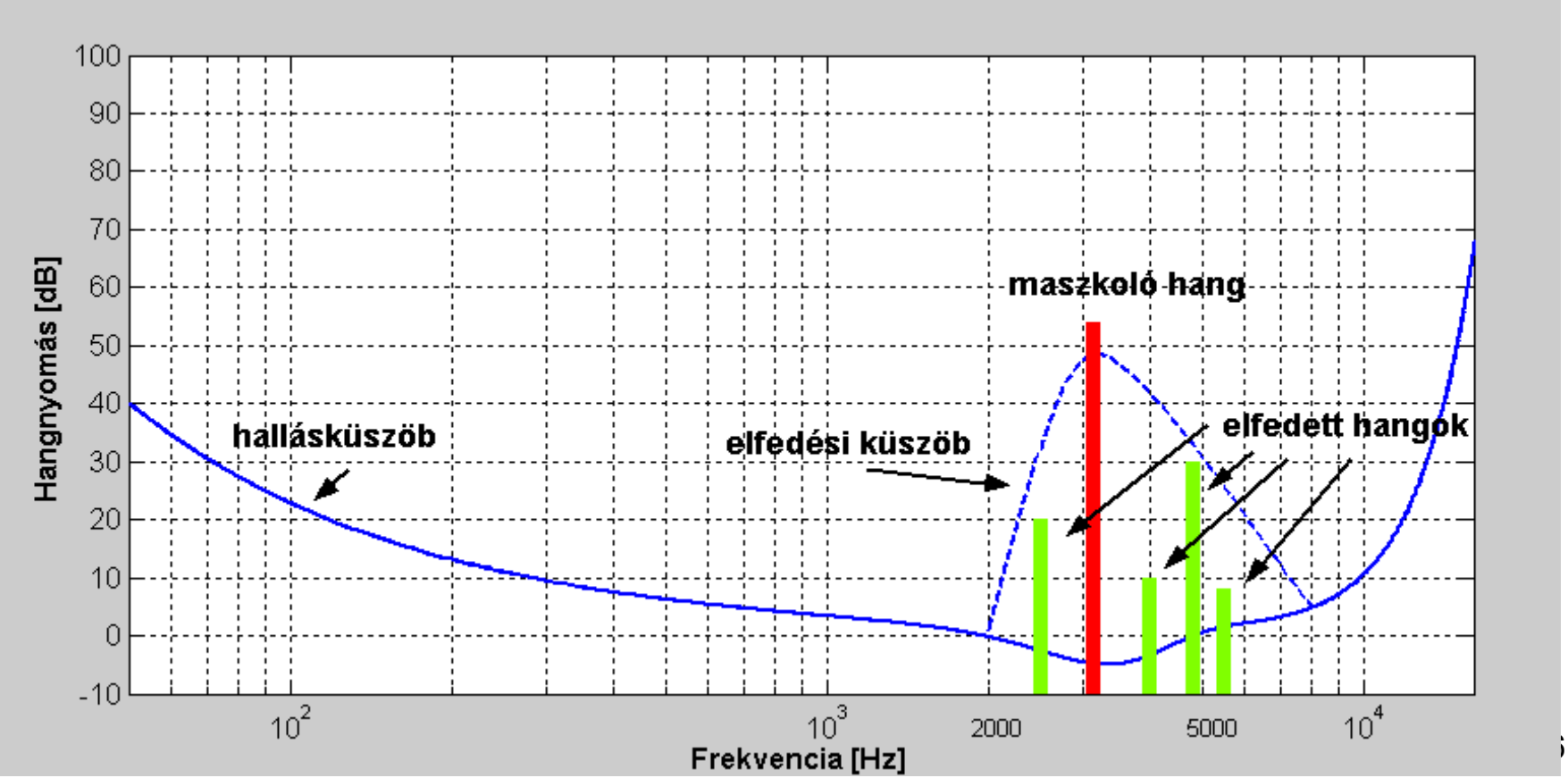

Human
Cat
Bat
Dog
Elephant
Moth
Mouse
10
100
1,000
10,000
100,000
Frequency (Hz)


A Look Inside the Ear
Pinna
Ear canal
Eustachian tube
Middle and Inner Ear
Fluid-filled canals
Auditory nerve
Ossicles
Eardrum
Cochlea
Vestibule
Oval window

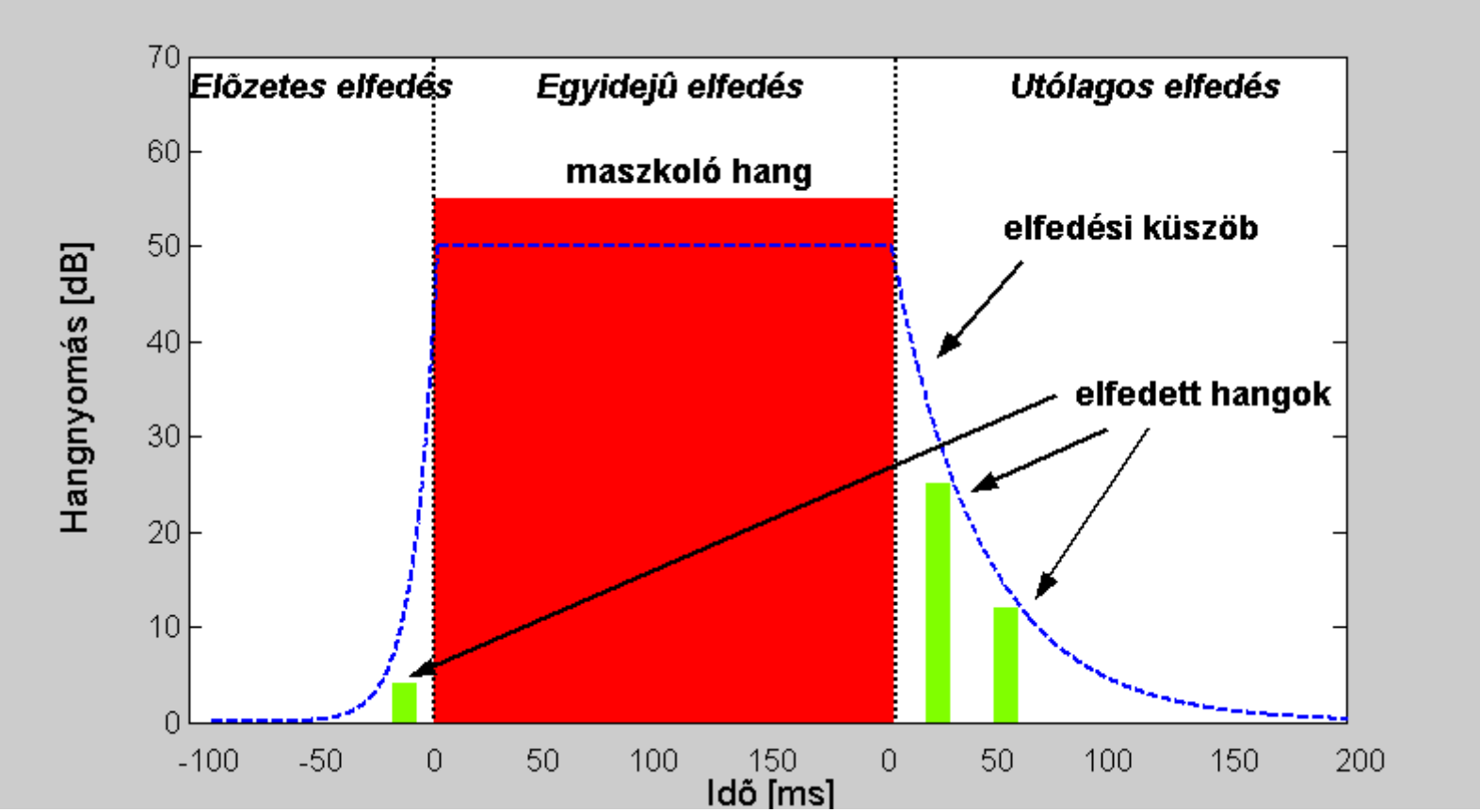

Hangnyomásszint [dB]
140
120
100
80
60
40
20
0
Fájdalom-
küszöb
70 phon
60 phon
50 phon
Hallásküszöb
0 phon
2 3 4 6 8 100 2 3 4 6 8 1000 2 3 4 6 8 10000
Frekvencia [Hz]


120
100
80
60
40
20
0
1 kHz
0.25 kHz
4 kHz
2
3
4
6
8
100
2
3
4
6
8
1000
2
3
4
6
8

Masking in the frequency domain



Masking in the time domain



Sound localization

- High-low
- Intensity
- Head movement
- Vision

Coding

- Standard
- Non-standard
- Speech and audio (music) coders
  - Quality
  - Delay
  - Jitter delay
  - Packet loss
  - Bandwidth

Multimodal

- Multiple modalities
- Advantages
- Disadvantages

Speech synthesizers

History

transport and speech technology

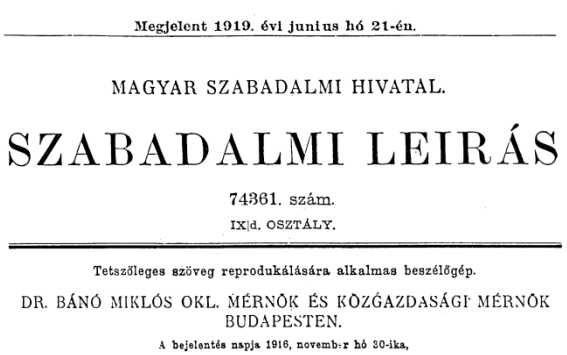
1791

2012

TTS from Kempelen to these days

Kempelen Farkas 1791

Bánó Miklós 1916



Megjelent 1919. évi junius hó 21-én.

MAGYAR SZABADALMI HIVATAL.

SZABADALMI LEIRÁS

74361. szám.

IX|d. OSZTÁLY.

Tetszőleges szöveg reprodukálására alkalmas beszélőgép.

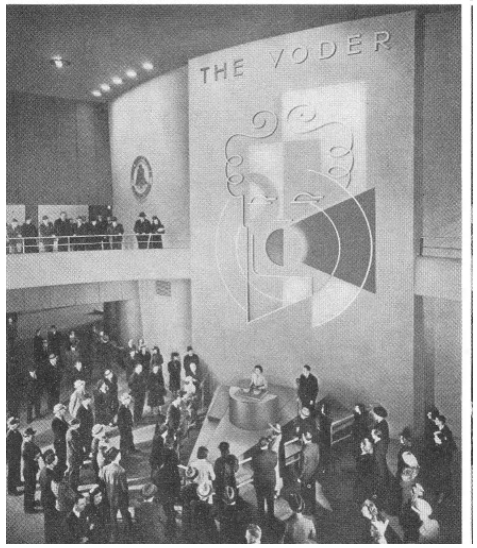
DR. BÁNÓ MIKLÓS OKL. MÉRNÖK ÉS KÖZGAZDASÁGI MÉRNÖK
BUDAPESTEN.

A bejelentés napja 1916. novemb:r hó 30-ika.

HungaroVox 1982
MultiVox 1986-2002

Some TTS milestones

Voder 1939



Dectalk 1982

Hungarian TTS progress

ProfiVox diphone 1995-

ProfiVox triphone 2000-

ProfiVox corpus 2002-

ProfiVox HMM 2005-

Some recent TTS examples

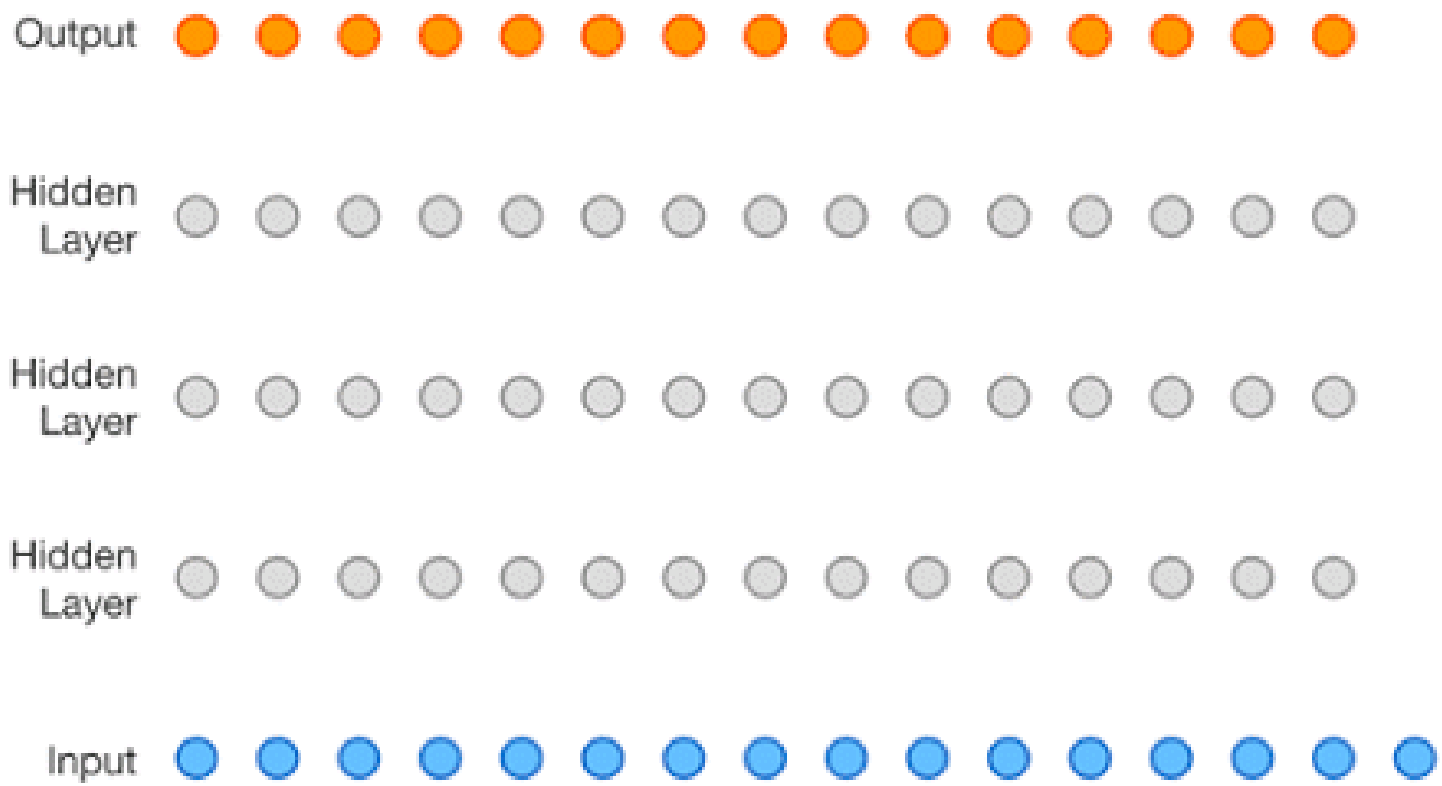
AT&T 2011 (US English)

AT&T 2011 (German)

Nuance (Loquendo) 2011 (US English)

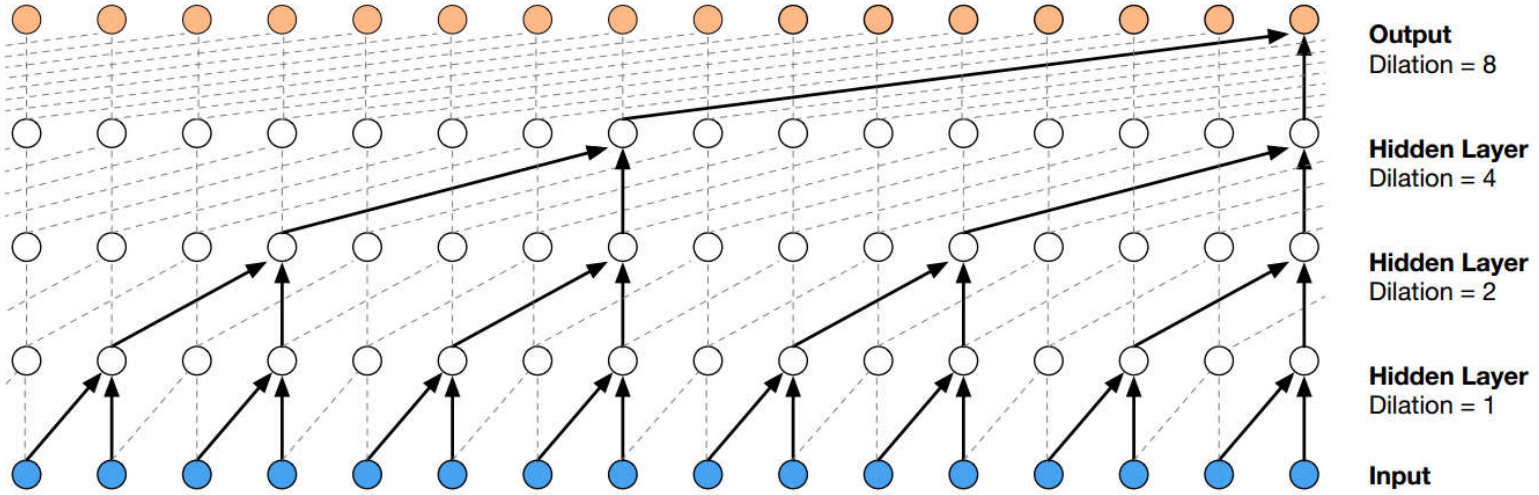
Nuance (Loquendo) 2011 (German)

Wavenet (Sept. 2016. -)



Source of Fig.: https://deepmind.com/blog/wavenet-generative-model-raw-audio/

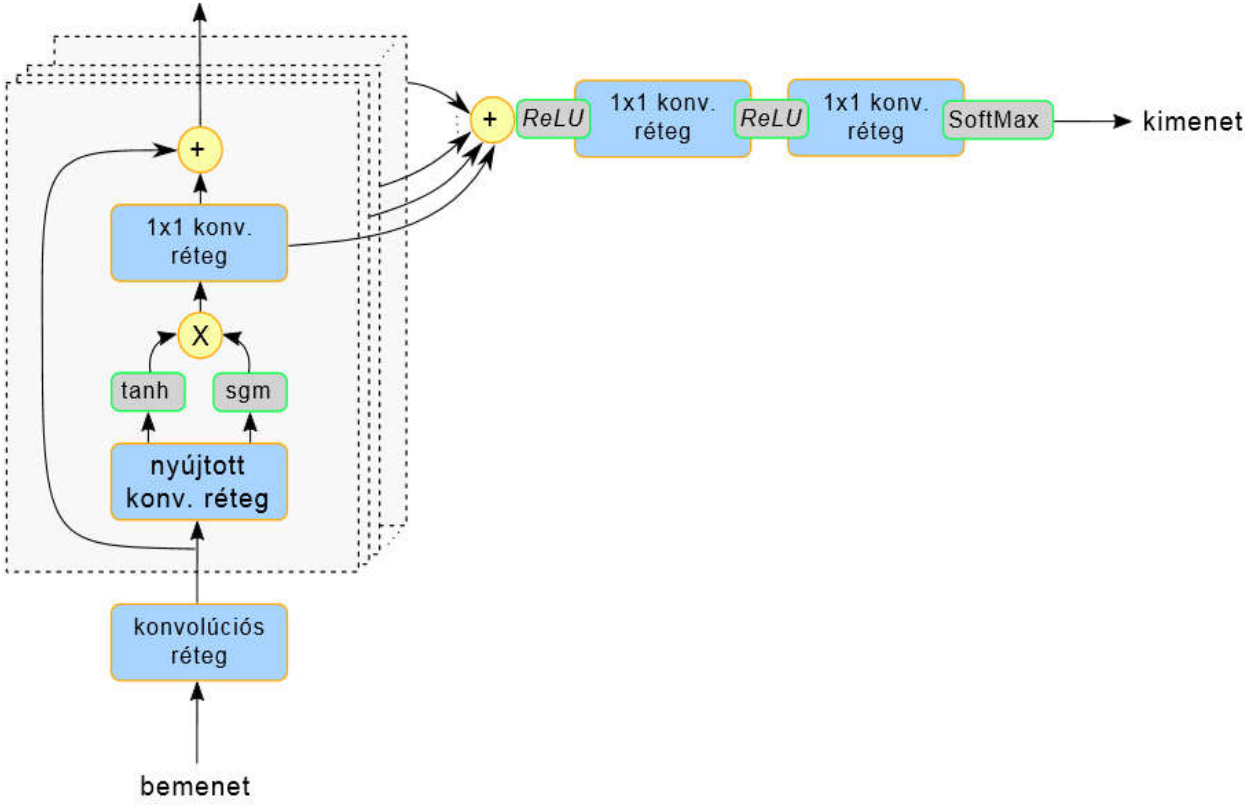
Generation



Google Oct. 2017.
(US angol és japán Google Assistant „production”)

Wavenet-based Hungarian

- Female Voice: Mátyus Kati
- Railways information
  - 3225 sentences
  - 44.1kHz, 16 bit
  - 27826s= 7 h 44 m



- TTS:

Amit nem tudsz egyszerűen elmagyarázni,
azt nem is érted egészen.

January 2017. November 2017

Albert Einstein

Prosody

- F0, intensity, length

- Intonation
- Stress/focus
- Pauses

What are the key factors?

Rule-based models
(articulation channel, prosody)
are agmented/replaced by
Natural units
in ever increasing size sets
statistical model creation
(minimized) signal processing
(approach to) Unified evaluation

Speech recognizers

- Barge-in
- Keyword
- Dictation
- Isolated word
- Speaker dependent/independent/adaptation

- Voice analytics

Growth predictions (1981-85)

	1981	1982	1983	1984	1985	AAGR (%) 1981-1985	1985 % OF TOTAL
SPEECH RECOGNITION							
Devices (Chips)	1	2	4	10	30	134%	20%
Products (Board Level)	10	17	36	70	100	78%	67%
Systems	4	6	9	13	20	50%	13%
Subtotal	$15	$25	$ 49	$ 93	$150	88%	100%
SPEECH SYNTHESIS							
Devices (Chips)	15	35	80	160	320	115%	65%
Products (Board Level)	5	12	25	50	100	111%	20%
Systems	3	9	20	40	75	124%	15%
Subtotal	$23	$56	$125	$250	$495	115%	100%
TOTAL	$38M	$81M	$174M	$343M	$645M	103%	

Source: Strategic, Inc.

Source:Voice Synthesis Nearing Growth Explosion, Computerworld, 31 August, 1981

Speaker identification

- Identification
  - Member of closed group
  - Who in the group
- Verification
  - The same as claimed

Non-verbal audio

- Background music
- Environment music
- Company image
- Event semaphore
- Limit signal

- Telephone

User environment

- Has to be defined
- Taken into account during design
- E.g:
  - Mobile ⇔ wireline
  - Home ⇔ other locations

Menu

- Number of levels/layers
- Mental load
- Consistency
- Interruptable, not interruptable

- Description:
  - Tree
  - Process flow-diagram
- Prompt design

Prompt

- Recording conditions
- Recording booklet
- Pauses, separation
- Stress labelling
- Full sentences
- Voice talent selection
  - Listening test in the application environment

Reading list

- Large letters
- 1.5 or double
- Organized text layout
- Prompts on one page
- Page numbering
- Free sheets

Recording sessions

- Max 4 hours / day
- The same time of day
- Drinks
- Breaks for rest
- Preparation, references

Based on usage

- Rarely
- Frequently
- Once
- Time frame

- Complexity
- No. of help requests
- Examples

Unique operation

- User specifique operation
- Identification
- User settings
- Personal agent

DTMF-recognizer

- Advantages /disadvantages

- Confirmation
  - Implicit
  - Explicit

Acoustic image

- - logo
  - simplified logo
  - slogen
  - business card
- - letter sheet
- - envelope
- - note block
- - dossier
  - publication cover
- - greeting card
  - advertisement plan
  - promotion material
- - fliers
  - annual reports
  - accounting printed forms
  - books, newspapers
  - presents
  - ……., etc.