Speech Production and Perception

TTT4185 Machine Learning for Signal Processing

Giampiero Salvi

Department of Electronic Systems
NTNU

HT2021

Why use speech?

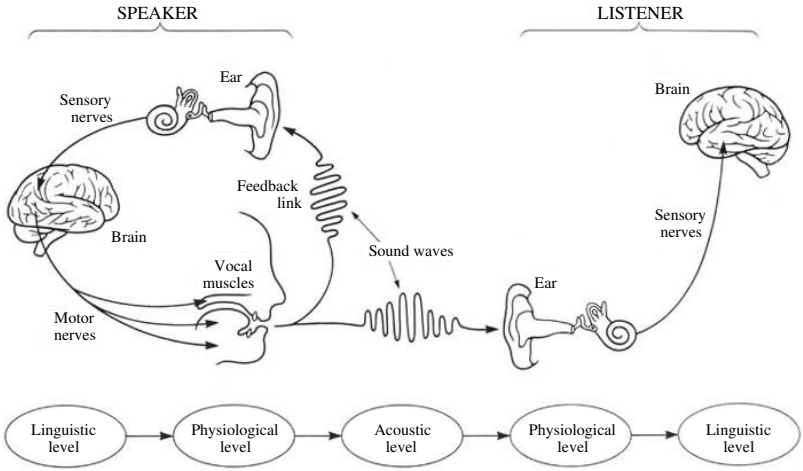
Human-Computer (or -Robot) Interaction

- Natural way of communication (No training needed)
- Leaves hands and eyes free (Good for functionally disabled)
- Effective (Higher data rate than typing)
- Can be transmitted/received inexpensively (phones)

Surveillance/Search

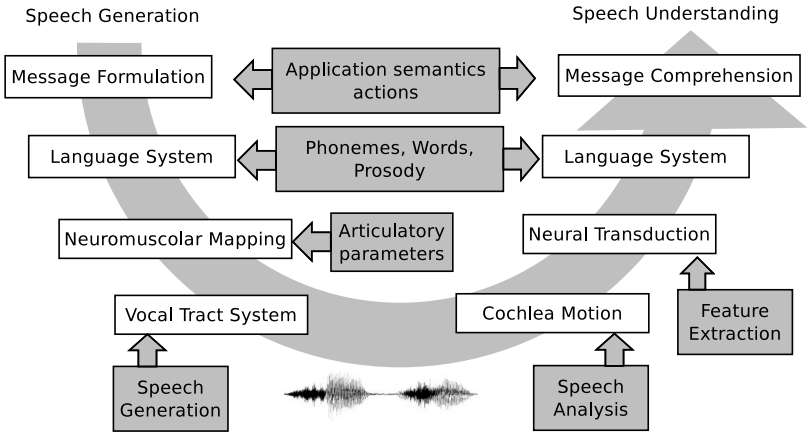
- Transcribe human-human conversations
- produce indexing for broadcast material
- produce subtitles for movies/news

The Speech Chain



Peter Denes, Elliot Pinson, 1963

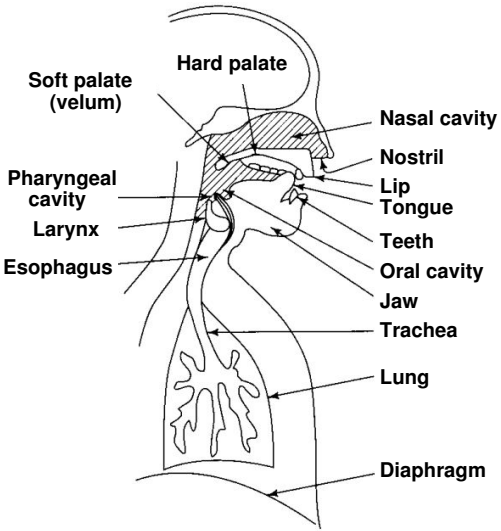
The Speech Chain revisited

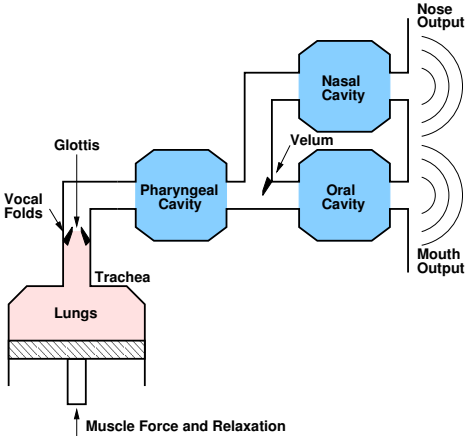


from Huang, Acero and Hon's book

Outline

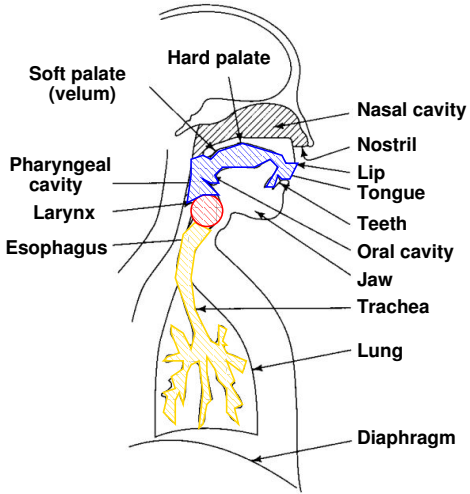
Physiology

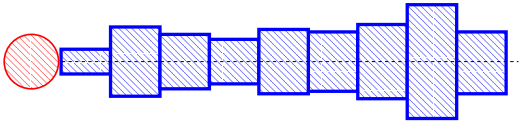

Soft palate (velum)
Hard palate
Nasal cavity
Nostril
Lip
Tongue
Pharyngeal cavity
Larynx
Esophagus
Teeth
Oral cavity
Jaw
Trachea
Lung
Diaphragm


Nose Output
Nasal Cavity
Velum
Glottis
Pharyngeal Cavity
Oral Cavity
Vocal Folds
Mouth Output
Trachea
Lungs
Muscle Force and Relaxation

Source/Filter Model, Vowel-like sounds

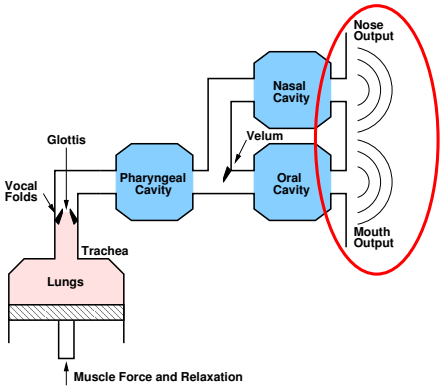
Vowels





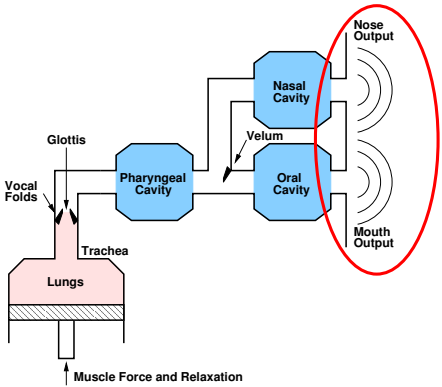
- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Radiation form the Lips/Nose

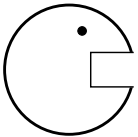


Problem of radiation at the lips plus diffraction about the head too complicated.

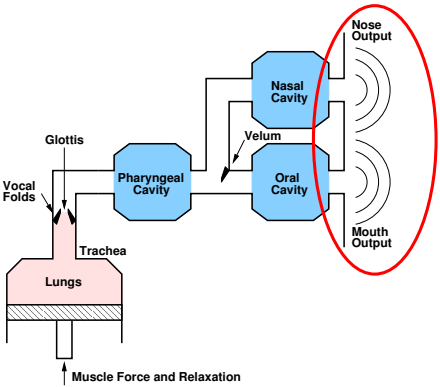
Radiation form the Lips/Nose



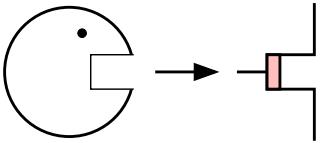
Approx. with a piston in a rigid sphere: solved but not in closed form



Radiation form the Lips/Nose



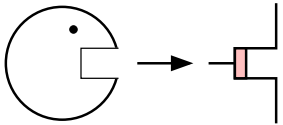
2nd approx: piston in an infinite wall



$$R(z) \approx 1 - \alpha z^{-1}$$

Radiation form the Lips/Nose

2nd approx: piston in an infinite wall



$$R(z) \approx 1 - \alpha z^{-1}$$

Question:

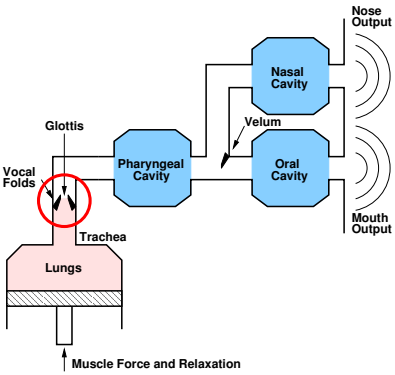
Given $R(z) = 1 - \alpha z^{-1}$ is the transfer function of a linear system, what is the relationship between the system input $x[n]$ and its output $y[n]$?

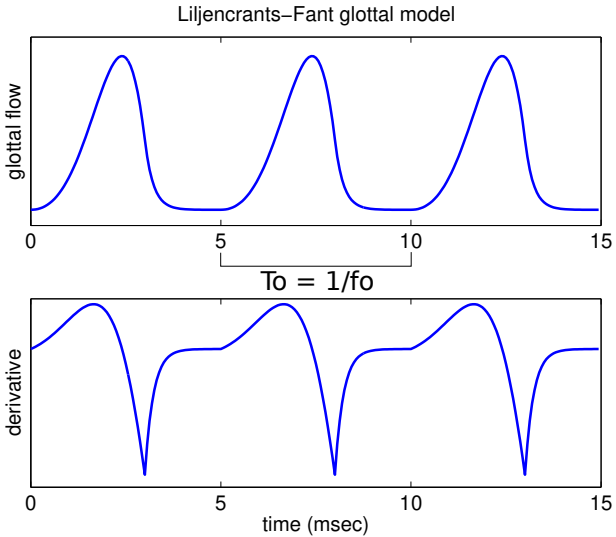
a) $y[n] =$ constant
b) $y[n] = 1 - \alpha x[n]$
c) $y[n] = x[n] - \alpha x[n-1]$
d) $y[n] = 1 - \alpha x[n-1]$

Glottal Flow: Laryngoscopy

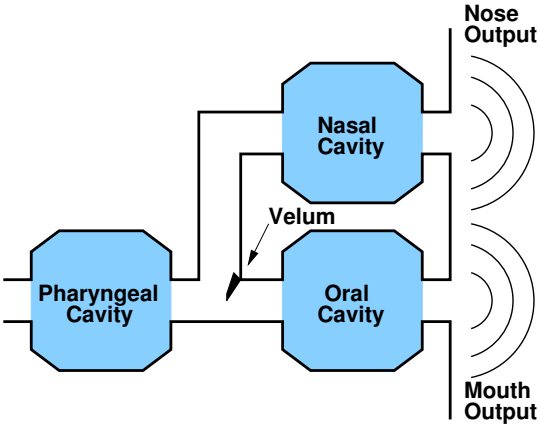
https://youtu.be/iYpDwhpILkQ

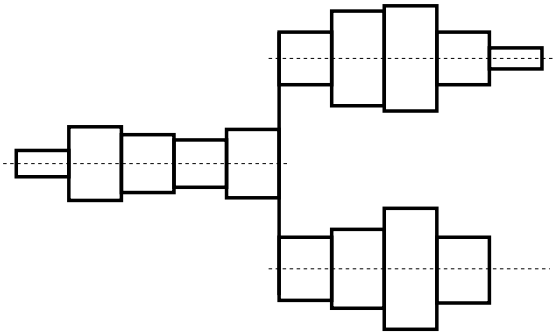
Glottal Flow



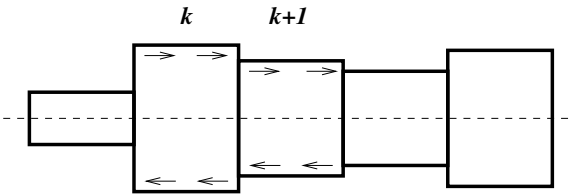


Tube Model of the Vocal Tract





Tube Model (cntd.)

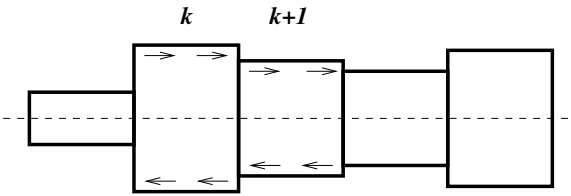


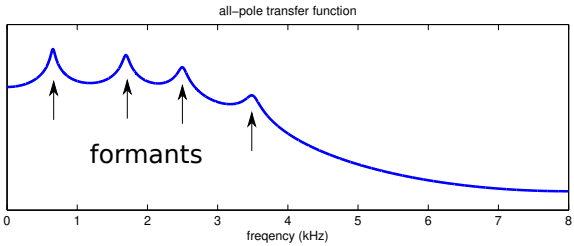
- assume planar wave propagation and lossless tubes
- solve pressure $p(x,t)$ and velocity $u(x,t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

$\Rightarrow$ all-pole transfer function ($N =$ number of tubes)

$$V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

Tube Model (cntd.)



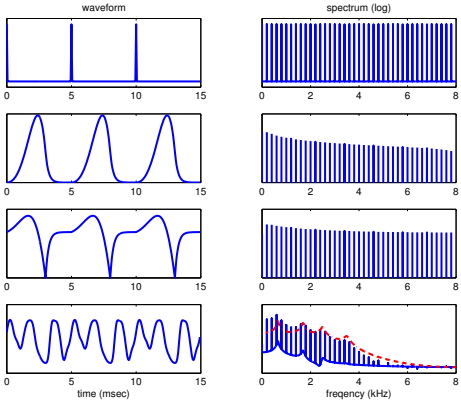


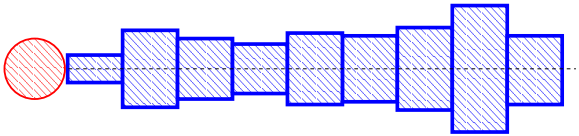
- assume planar wave propagation and lossless tubes
- solve pressure $p(x,t)$ and velocity $u(x,t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

$\Rightarrow$ all-pole transfer function ($N =$ number of tubes)

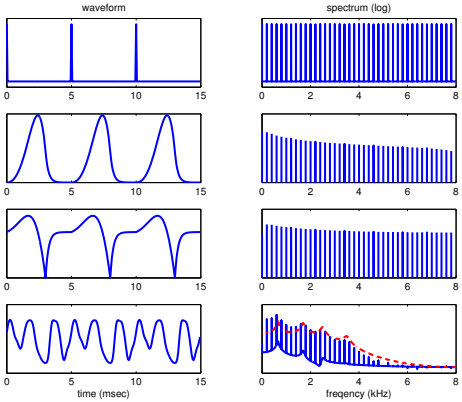
$$V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

Source/Filter Model: vowel-like sounds

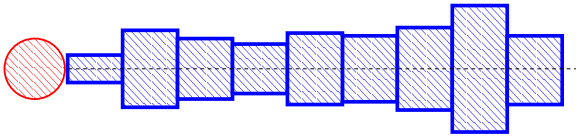




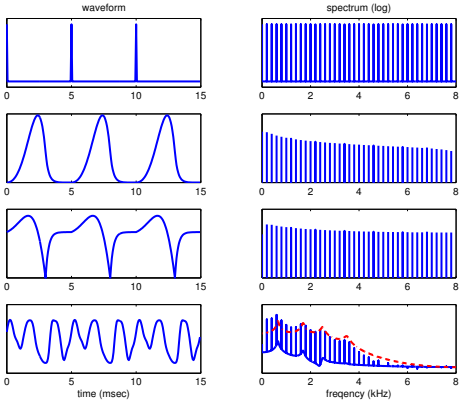
Source/Filter Model: vowel-like sounds



$\leftarrow p[n]$

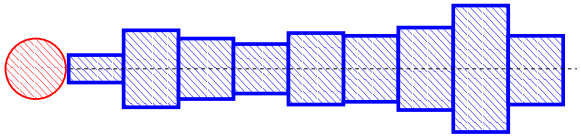


Source/Filter Model: vowel-like sounds

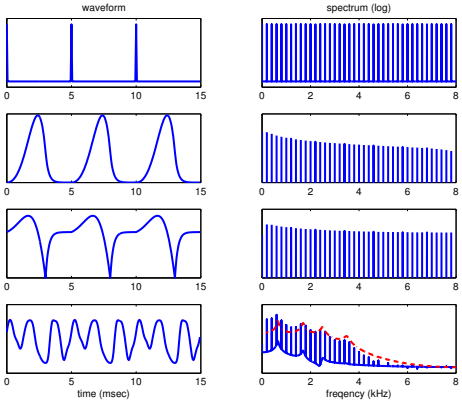


$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$



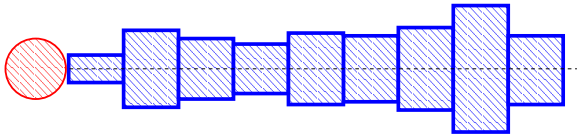
Source/Filter Model: vowel-like sounds



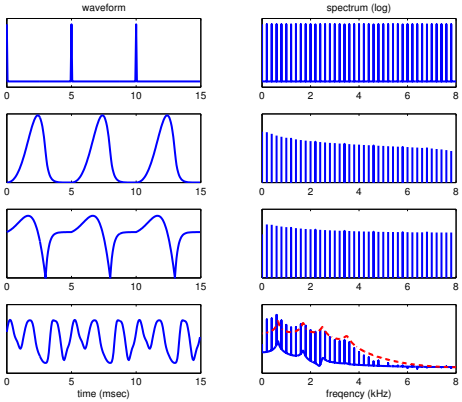
$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$

$\leftarrow p[n] * g[n] * r[n]$



Source/Filter Model: vowel-like sounds

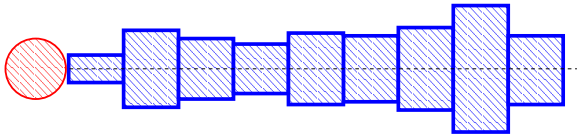


$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$

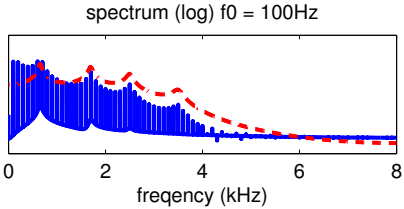
$\leftarrow p[n] * g[n] * r[n]$

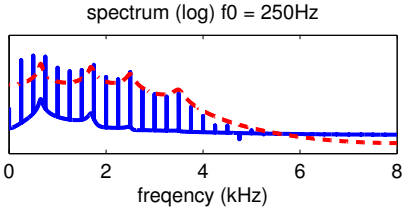
$\leftarrow p[n] * g[n] * r[n] * v[n]$



f_0 and Formants

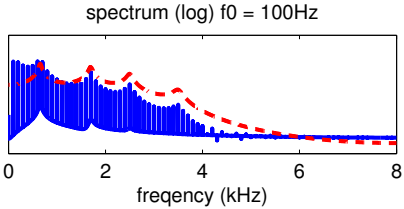
- Varying f_0 (vocal fold oscillation rate)

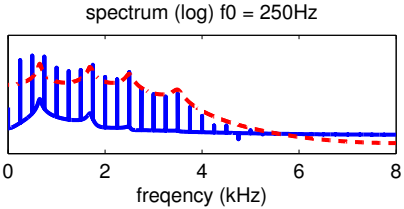




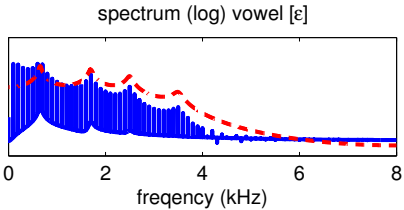
f_0 and Formants

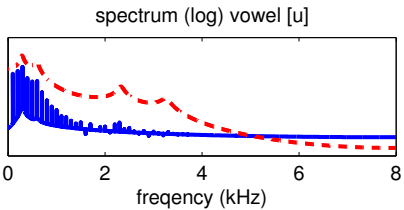
- Varying f_0 (vocal fold oscillation rate)





- Varying Formants (vocal tract shape)





Demonstration of speech sounds

Wavesurfer

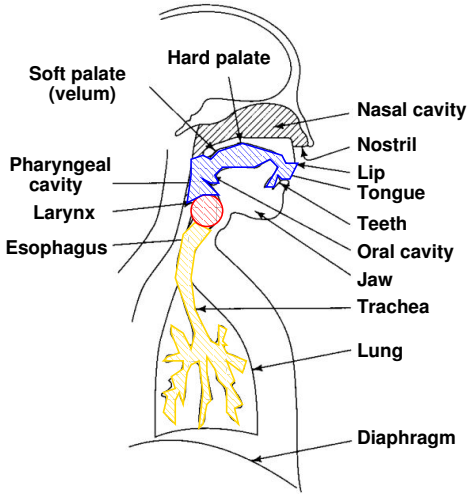
https://sourceforge.net/projects/wavesurfer/

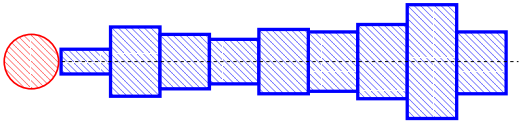
Praat

https://www.fon.hum.uva.nl/praat/

Source/Filter Model, General Case

Vowels

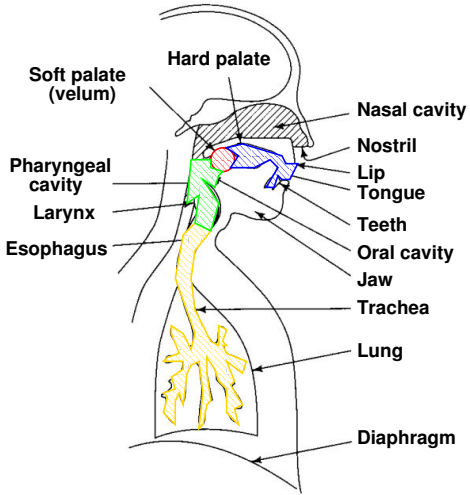


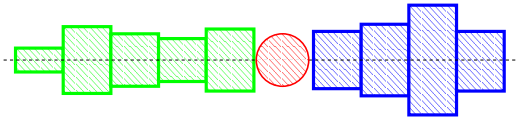


- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

Fricatives (e.g. sh) or Plosive (e.g. k)

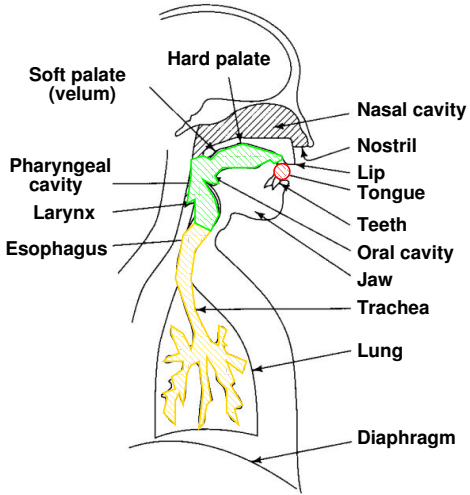


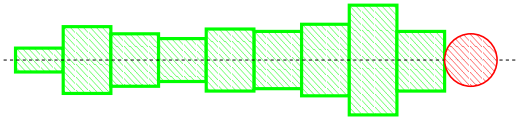


- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

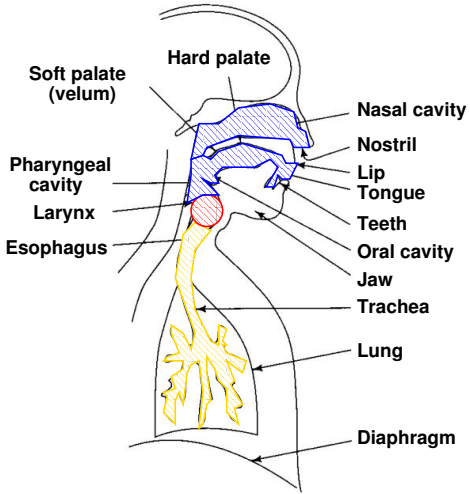
Fricatives (e.g. s) or Plosive (e.g. t)



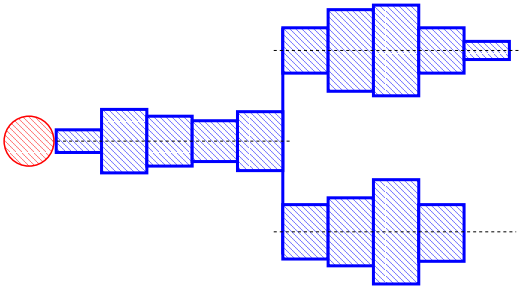


- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

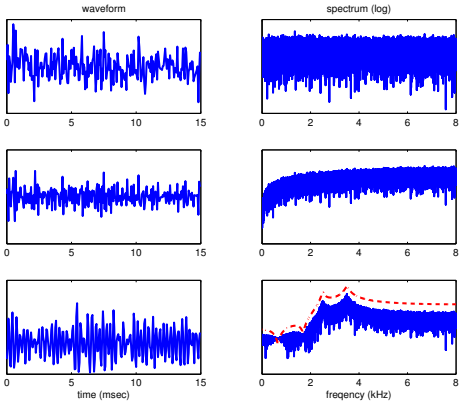


Nasalised Vowels



- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

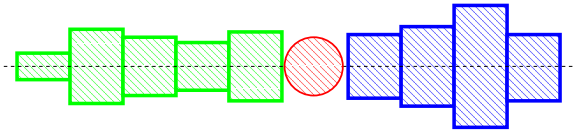
Source/Filter Model: fricative sounds



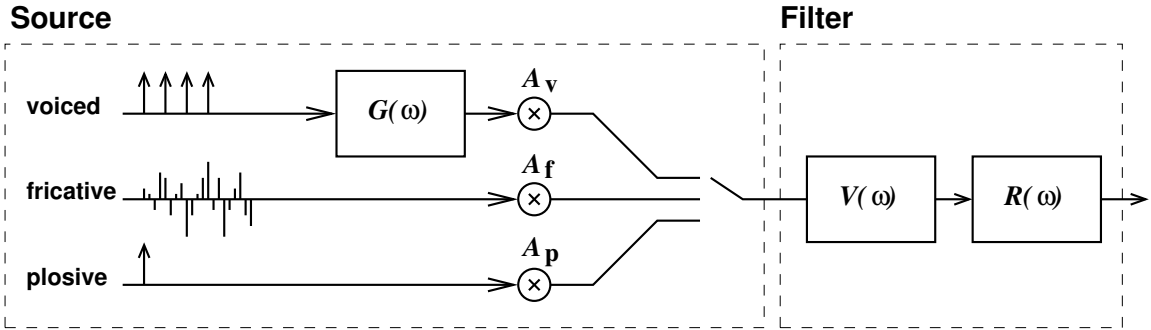
$\leftarrow p[n]$

$\leftarrow p[n] * r[n]$

$\leftarrow p[n] * r[n] * v[n]$

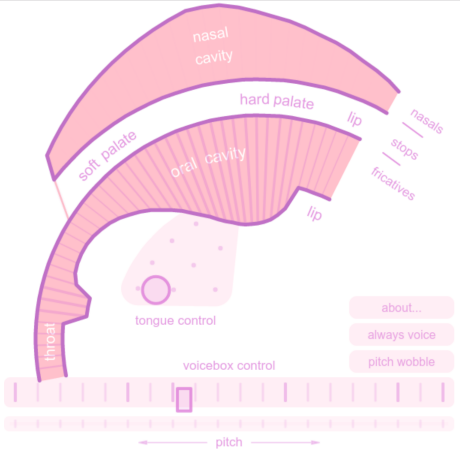


Complete Source/Filter Model

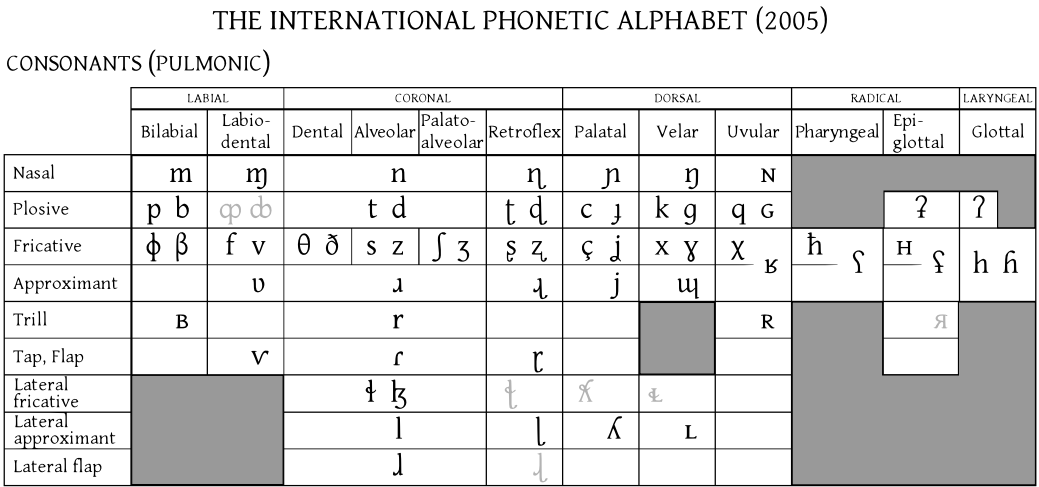


Pink Trombone!

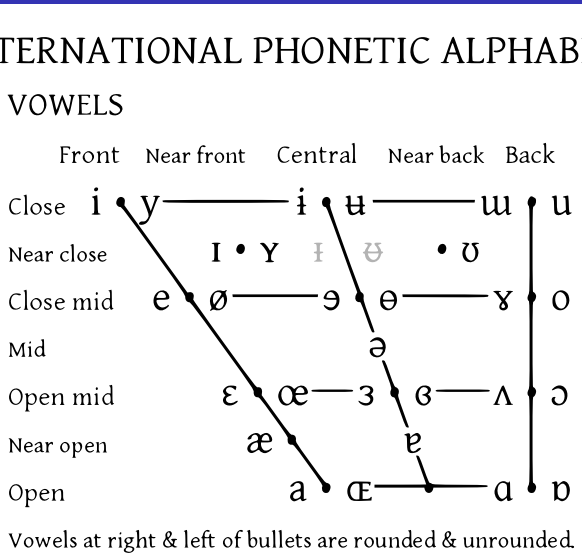
http://dood.al/pinktrombone/



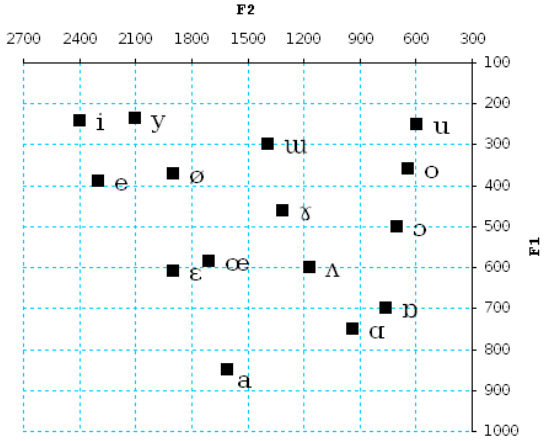
IPA Chart: Consonants



IPA Chart: Vowels

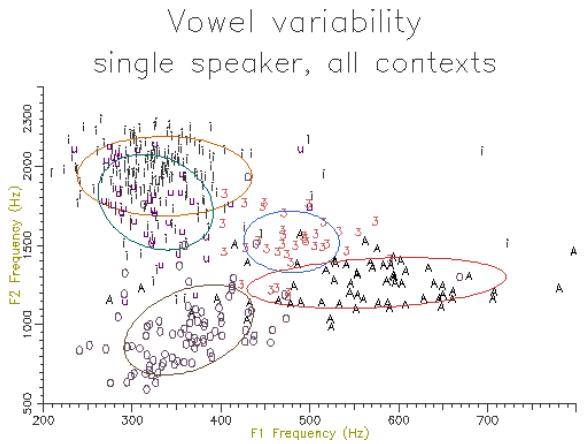


Average formant frequencies (vowels)



Source: `https://en.wikipedia.org/wiki/Formant`

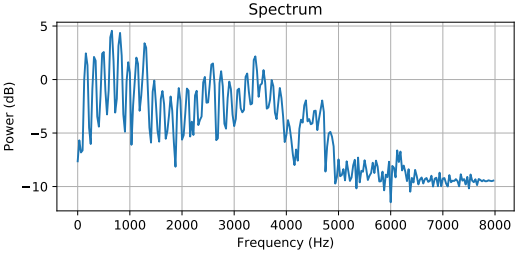
Formant frequencies, single speaker



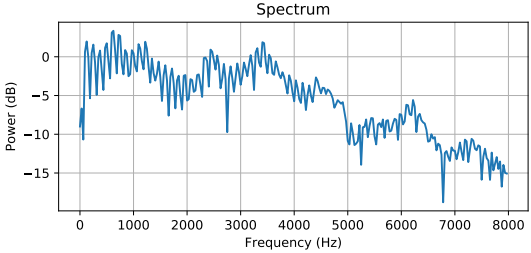
Source: `https://www.phon.ucl.ac.uk/resource/sfs/howto/formant.php`

Question: which is the speech sound with higher f_0?

a)

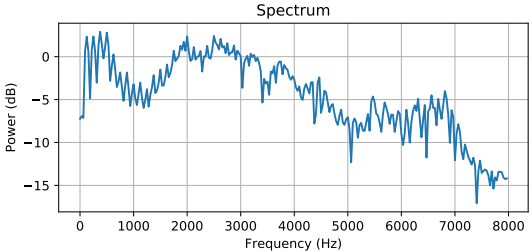


b)

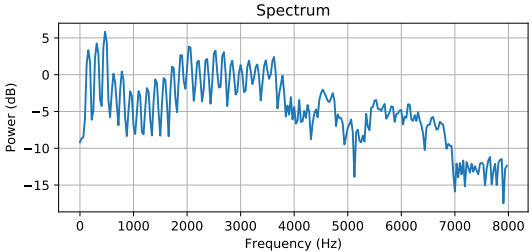


Question: which is the speech sound with higher f_0?

a)

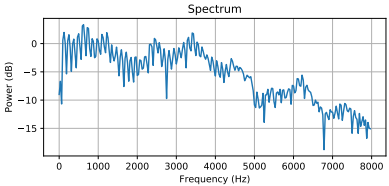


b)

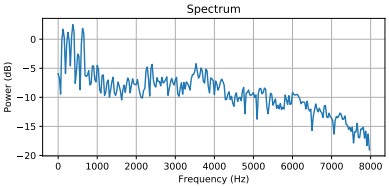


Question: which spectrogram depicts an /a/?

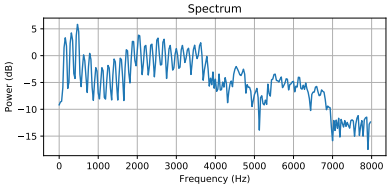
a)



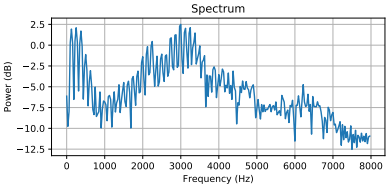
b)



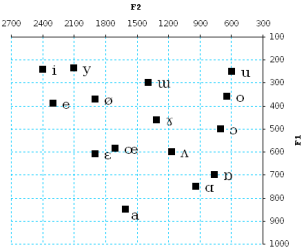
c)



d)

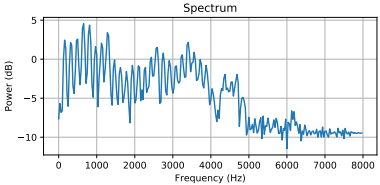


Average formant frequencies

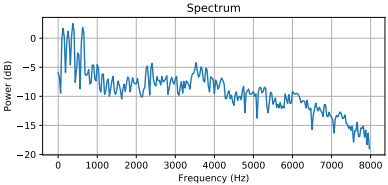


Question: which spectrogram depicts an /i/?

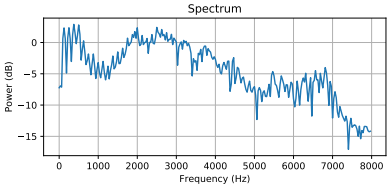
a)



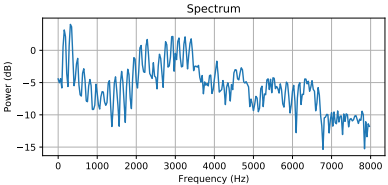
b)



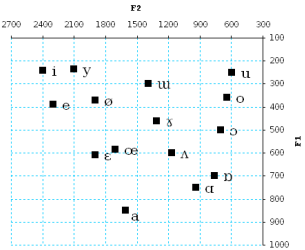
c)



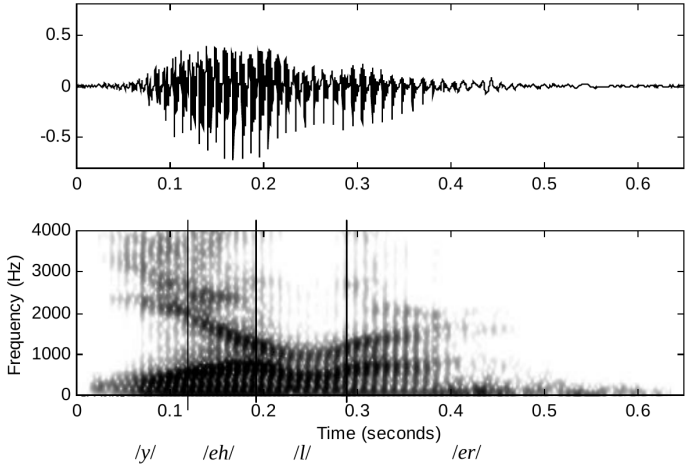
d)



Average formant frequencies

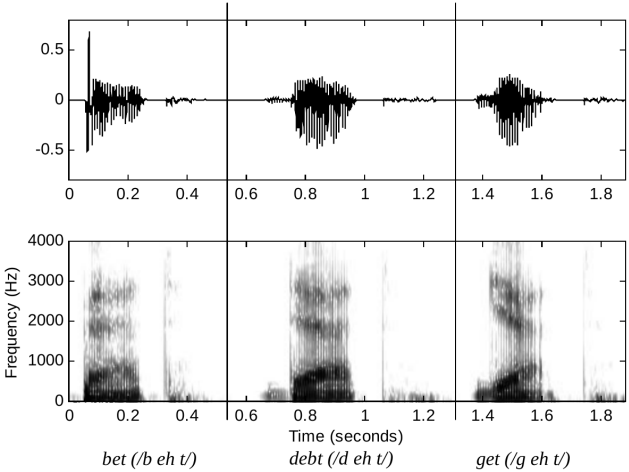


Coarticulation: vowels-semivowels



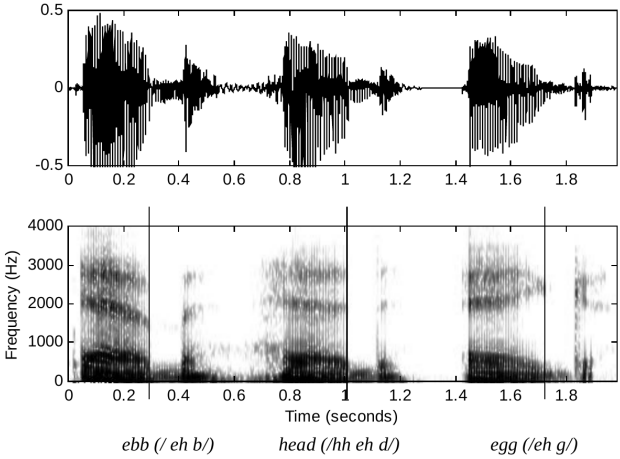
from Huang, Acero and Hon's book

Coarticulation: plosive-vowel



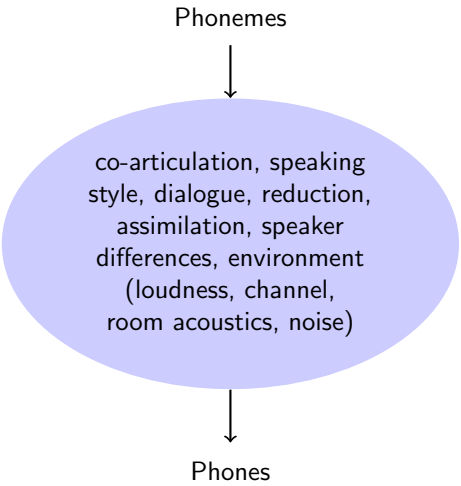
from Huang, Acero and Hon's book

Coarticulation: vowel-plosive

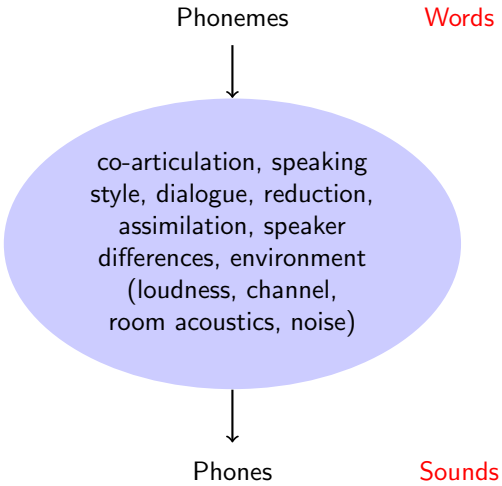


from Huang, Acero and Hon's book

Phonology vs Phonetics

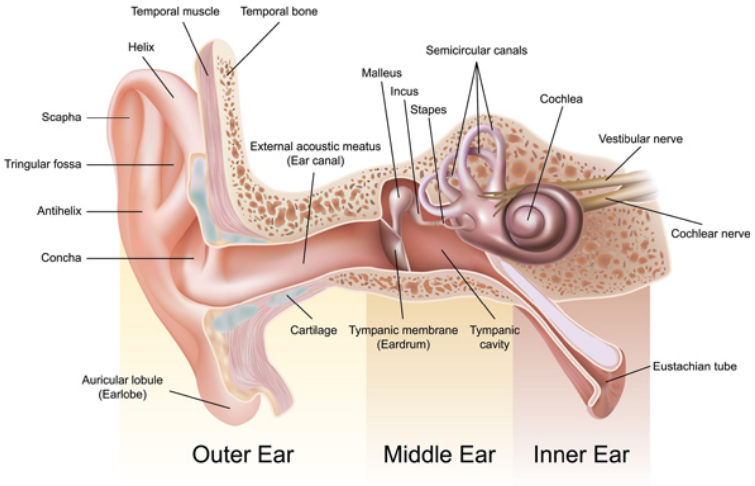


Phonology vs Phonetics



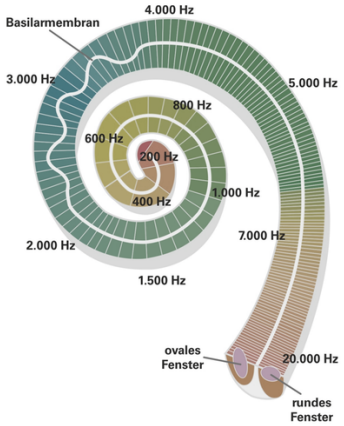
Outline

Anatomy of the ear



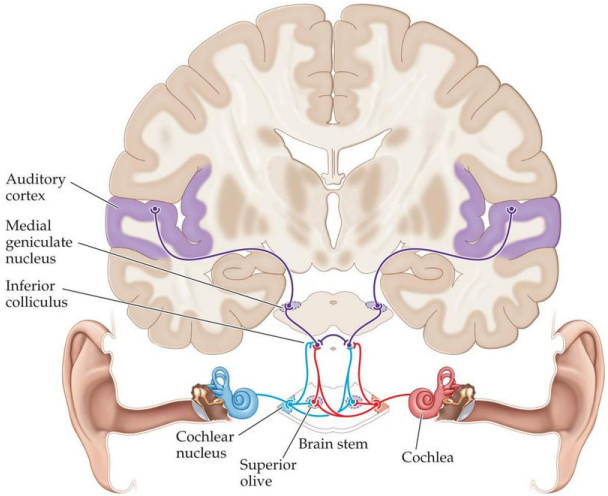
from `https://drjillgordon.com/how-hearing-works`

Cochlea



from `https://tu-dresden.de/ing/elektrotechnik/ias/aha/forschung/akustik/psychoakustik?set_language=en`

Auditory Pathways

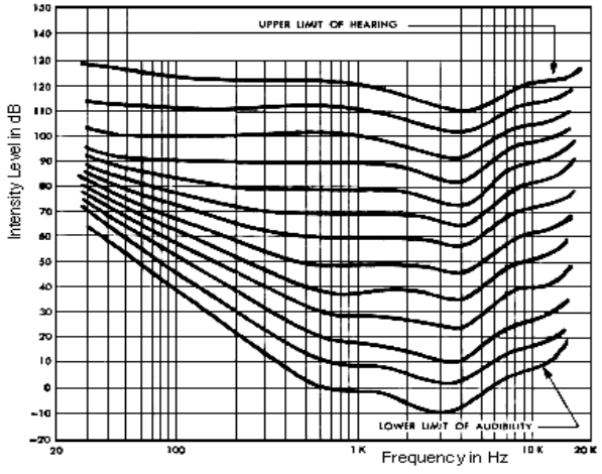


SENSATION & PERCEPTION 5e, Figure 9.20

Physical vs perceptual attributes

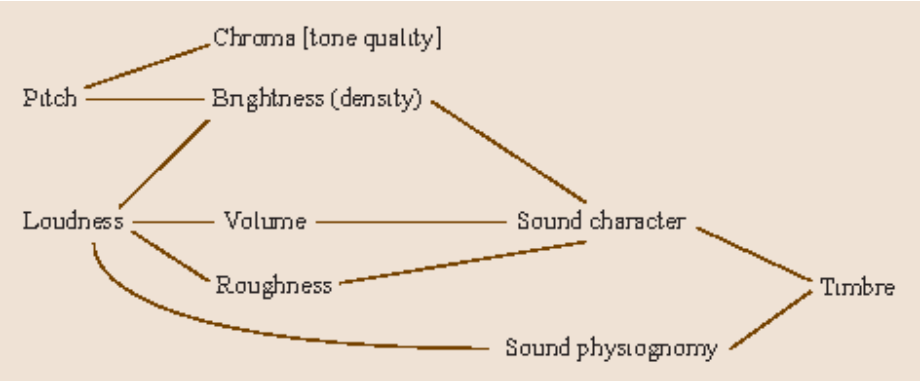
Physical Quantity	**Perceptual Quantity**
Intensity (SPL)	Loudness
Fundamental frequency (f_0)	Pitch
Spectral shape	Timbre
Onset/offset time	Timing
Phase difference in binaural hearing	Location

Equal-loudness curves (pure tones)



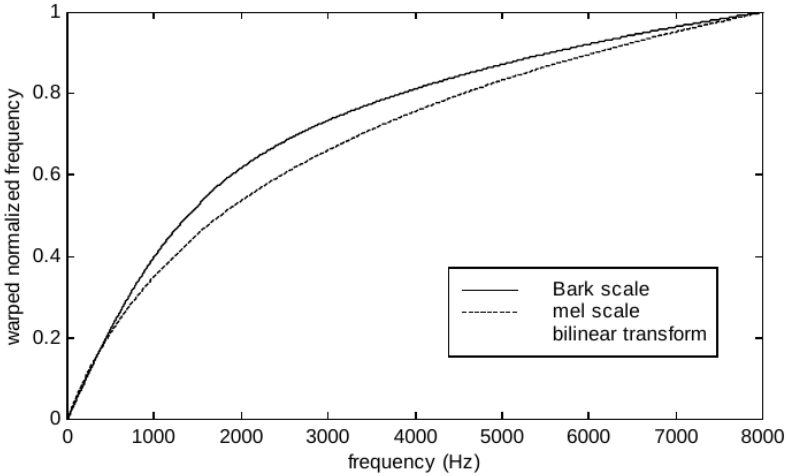
from Huang, Acero and Hon's book

Loudness perception (complex sounds)



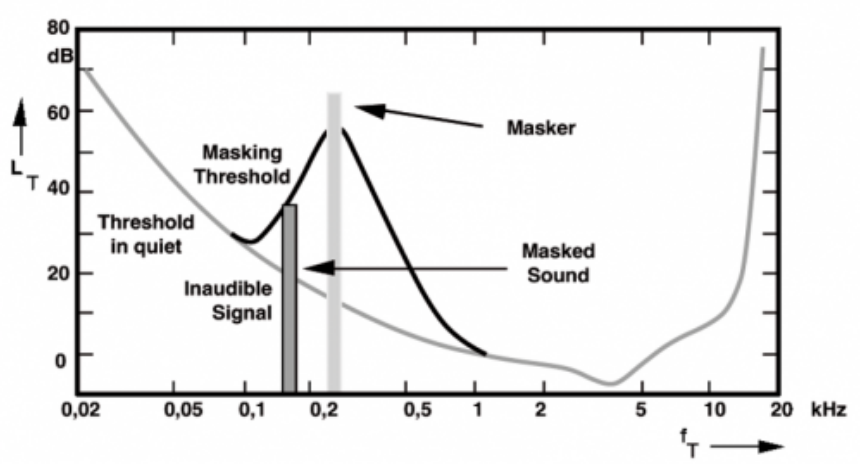
from https://link.springer.com/chapter/10.1007/978-3-662-55004-5_33

Bark and Mel frequency scales



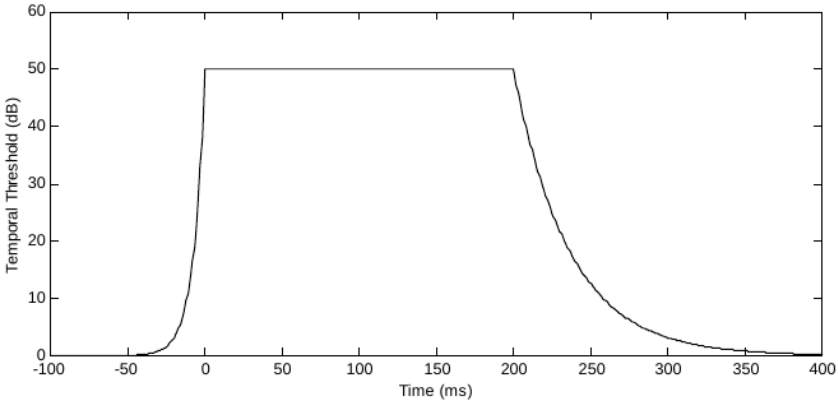
from Huang, Acero and Hon's book

Masking in frequency



from http://hephaestusaudio.com/delphi/category/psychoacoustics/

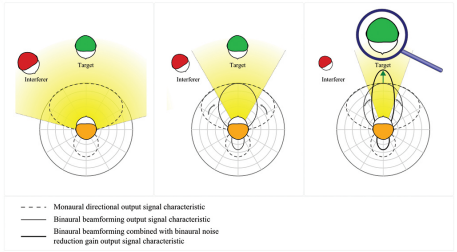
Masking in time



from Huang, Acero and Hon's book

Binaural hearing

- mainly used for localization
- main cue: inter-aural time difference (ITD)
- main cue: inter-aural level difference (ILD)
- filtering from ear pinna
- head-related transfer function HRTF



Check `http://auditoryneuroscience.com/spatial-hearing/binaural-cues`

Outline

Challenges — Variability

Between speakers

- Age
- Gender
- Anatomy
- Dialect

Within speaker

- Stress
- Emotion
- Health condition
- Read vs Spontaneous
- Adaptation to environment (Lombard effect)
- Adaptation to listener

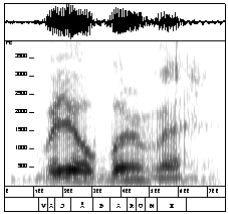
Environment

- Noise
- Room acoustics
- Microphone distance
- Microphone, telephone
- Bandwidth

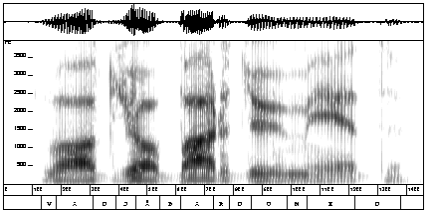
Listener

- Age
- Mother tongue
- Hearing loss
- Known / unknown
- Human / Machine

Example: spontaneous vs hyper-articulated



Va jobbaru me



Vad jobbar du med

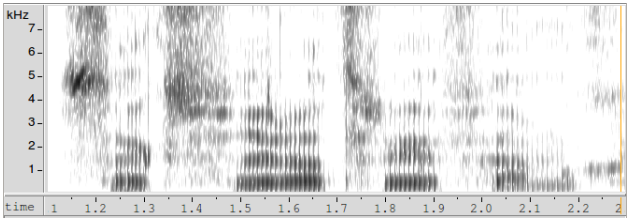
"What is your occupation"
("What work you with")

Examples of reduced pronunciation

Spoken	Written	In English
Tesempel	Till exempel	for example
åhamba	och han bara	and he just
bafatt	bara för att	just because
javende	jag vet inte	I don't know

Microphone distance

Headset



2 m distance