Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

Prof. Timo Gerkmann

Speech Signal Processing

Signal Processing Group
Department of Informatics
Universität Hamburg

SS 2018/2019

Lecture

Dates: Mon 12-14, G-210 Tue 16-18, D-018
Contact: timo.gerkmann@uni-hamburg.de

Exercises

- A protocol of the exercises is to be handed in July 16th 2018 via Email.

Literature

- P. Vary, R. Martin, "Digital Speech Transmission," Wiley, 2006.
- P. Vary, U. Heute, W. Hess: "Digitale Sprachsignalverarbeitung", Teubner Verlag, 1998
- Hendriks, Gerkmann, Jensen: "DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art", Morgan & Claypool, 2013
- R.F. Lyon, "Human and Machine Hearing", Cambridge, 2017

What Do We Learn?

Speech production: How is speech produced by humans?

Speech perception: How do we perceive speech signals?

Speech synthesis: How can we produce speech synthetically?

Speech analysis: What are the most important parameters of speech and how can we represent them?

Speech coding: How can we code speech efficiently?

Speech enhancement: How can we improve noisy speech?

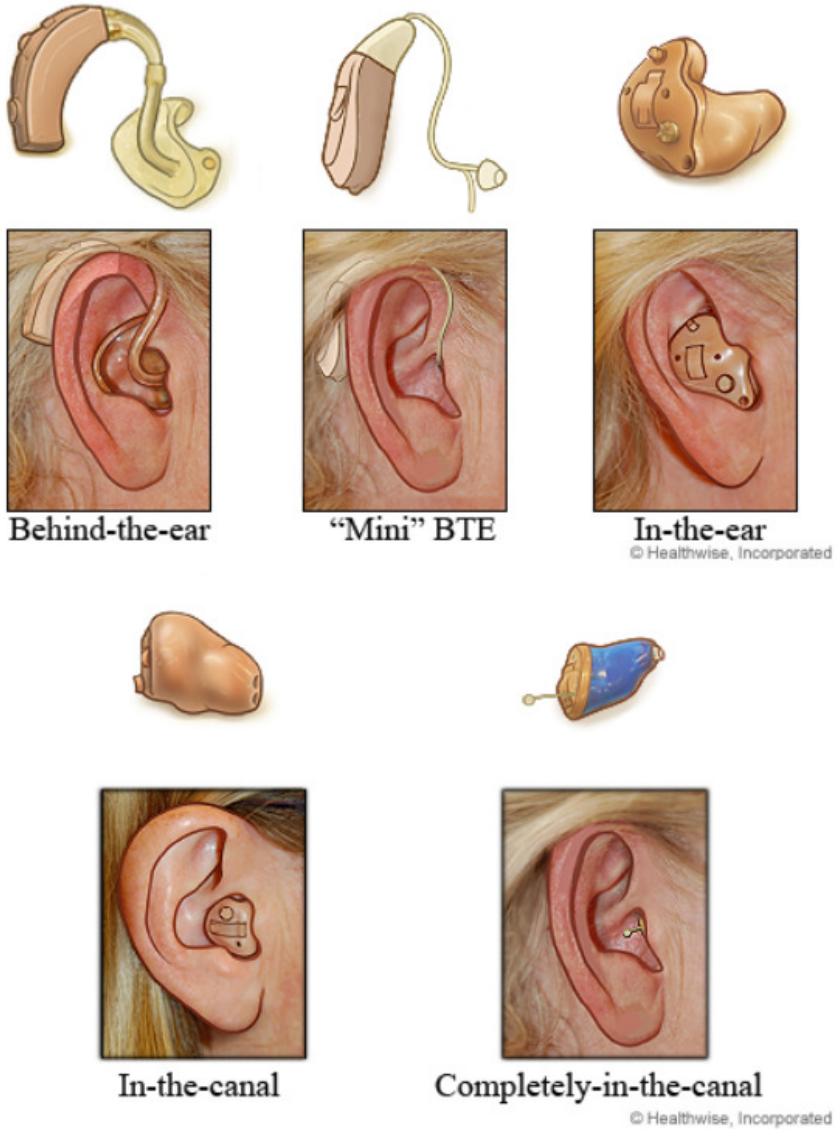
Speech recognition: How can computers automatically recognize speech?

Outline

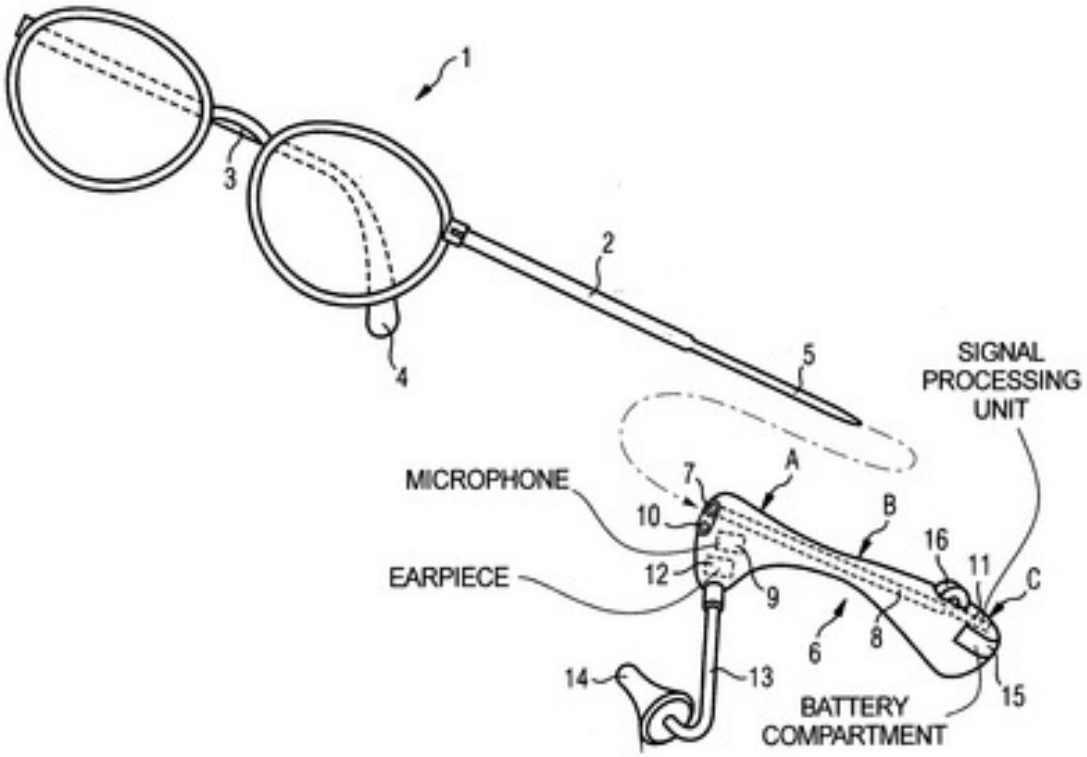
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

1. Introduction


Behind-the-ear
"Mini" BTE
In-the-ear
© Healthwise, Incorporated
In-the-canal
Completely-in-the-canal
© Healthwise, Incorporated

Speech signal processing for hearing devices



United States Patent 7103192

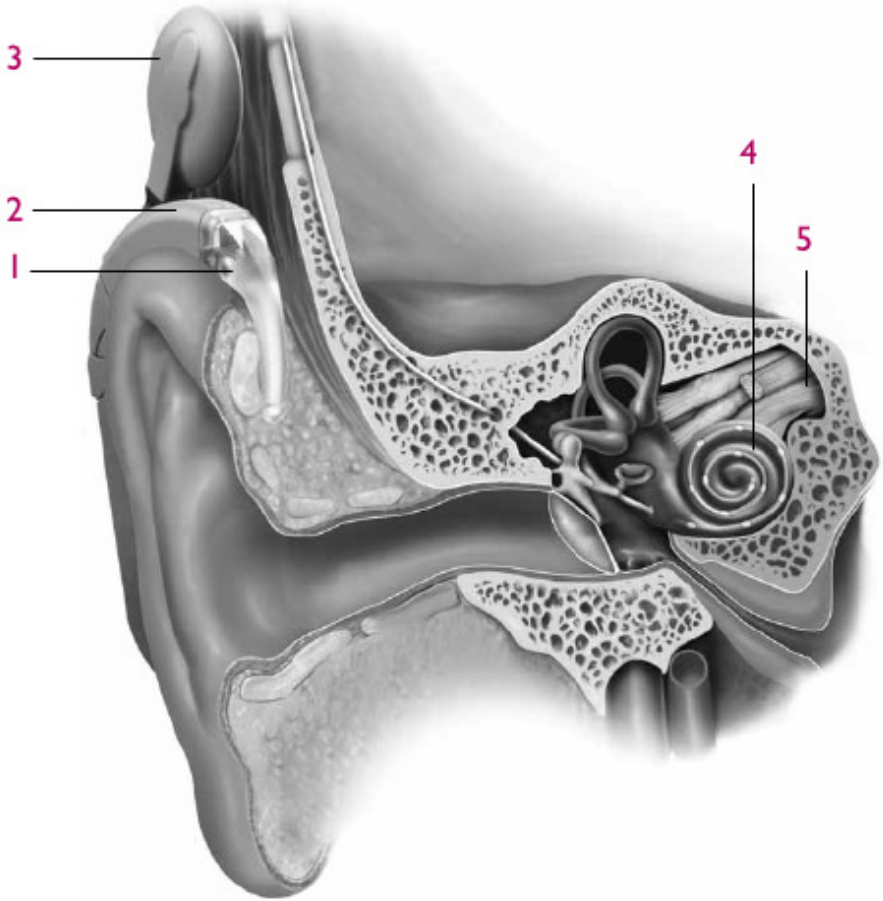
Speech signal processing for hearing devices

varibel.nl

Speech signal processing for hearing devices

Quelle: designaffairs.com

Cochlear Implants



1 Sounds are picked up by the microphone.

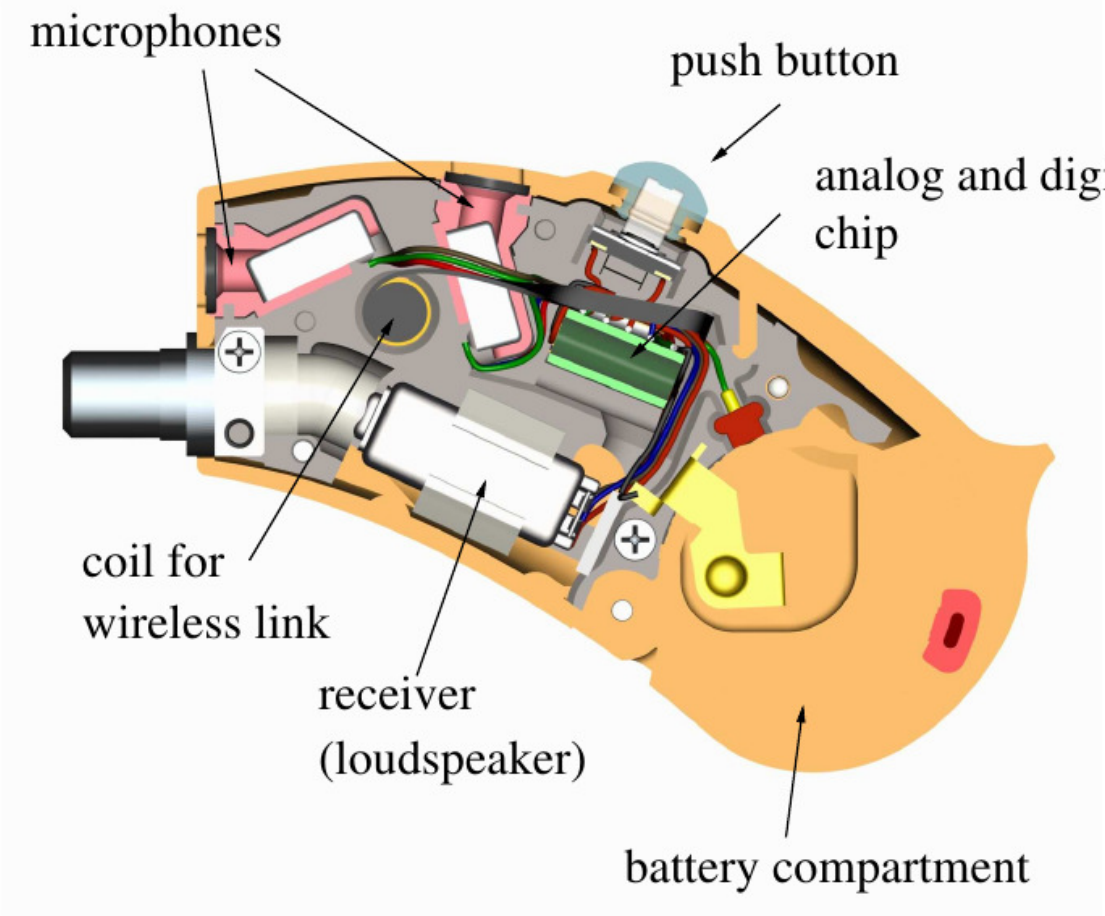
2 The signal is then "coded" (turned into a special pattern of electrical pulses).

3 These pulses are sent to the coil and are then transmitted across the skin to the implant.

4 The implant sends a pattern of electrical pulses to the electrodes in the cochlea.

5 The auditory nerve picks up these electrical pulses and sends them to the brain. The brain recognizes these signals as sound.

Quelle: Handbook for Educators, Med-El



Quelle: Siemens Audiologische Technik

There's Work To Do

- Successful speech communication requires good speech perception
- Hearing loss impedes inter-human communication and thus social contacts
- 19% of the German population is hearing impaired, of which
  - mild hearing loss: 56.5%
  - medium hearing loss: 35.2%
  - large hearing loss: 7.2%
  - deaf or almost deaf: 1.6%
  - source: http://www.schwerhoerigen-netz.de
- Unlike glasses, hearing aids can only partly compensate for hearing loss
- Speech understanding in noise remains difficult

➡ Only 20% of all hearing impaired in the EU use a hearing aid.

Quelle: http://www.nuheara.com/

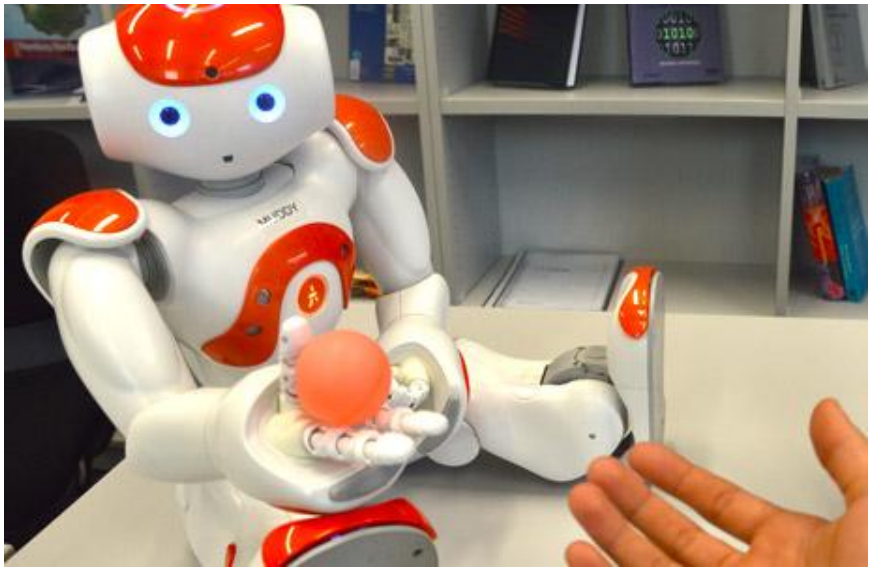
Wireless earbuds for assisted listening

- no prescription needed ➡ much faster time to market
- computations can be done on smartphone / cloud

Typical Algorithms/Functionality

- Music Streaming
- Blended Audio Worlds
- Noise Cancellation
- Advanced Speech Amplification (like a hearing aid)

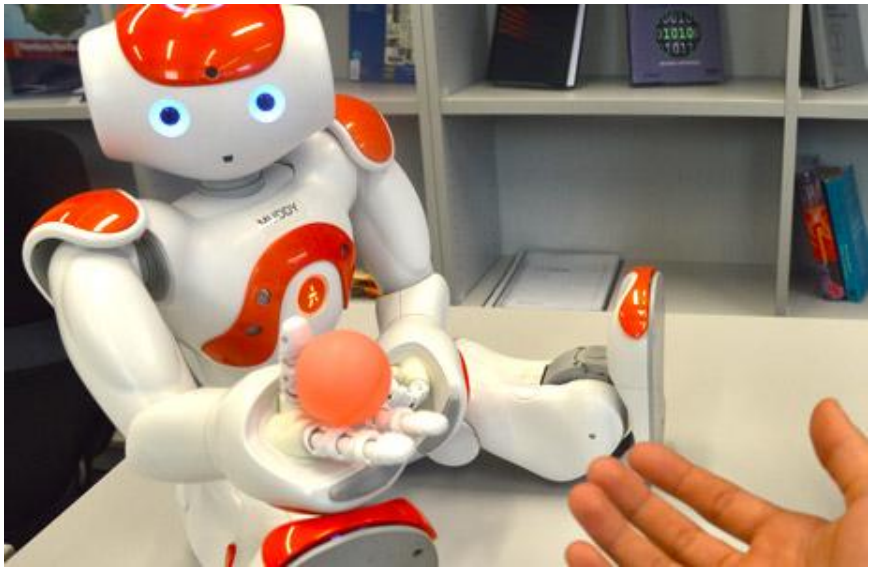
Human-Machine Interfaces



- Robust speech recognition required

Video recordings from http://robot-ears.eu

Human-Machine Interfaces



- Robust speech recognition required

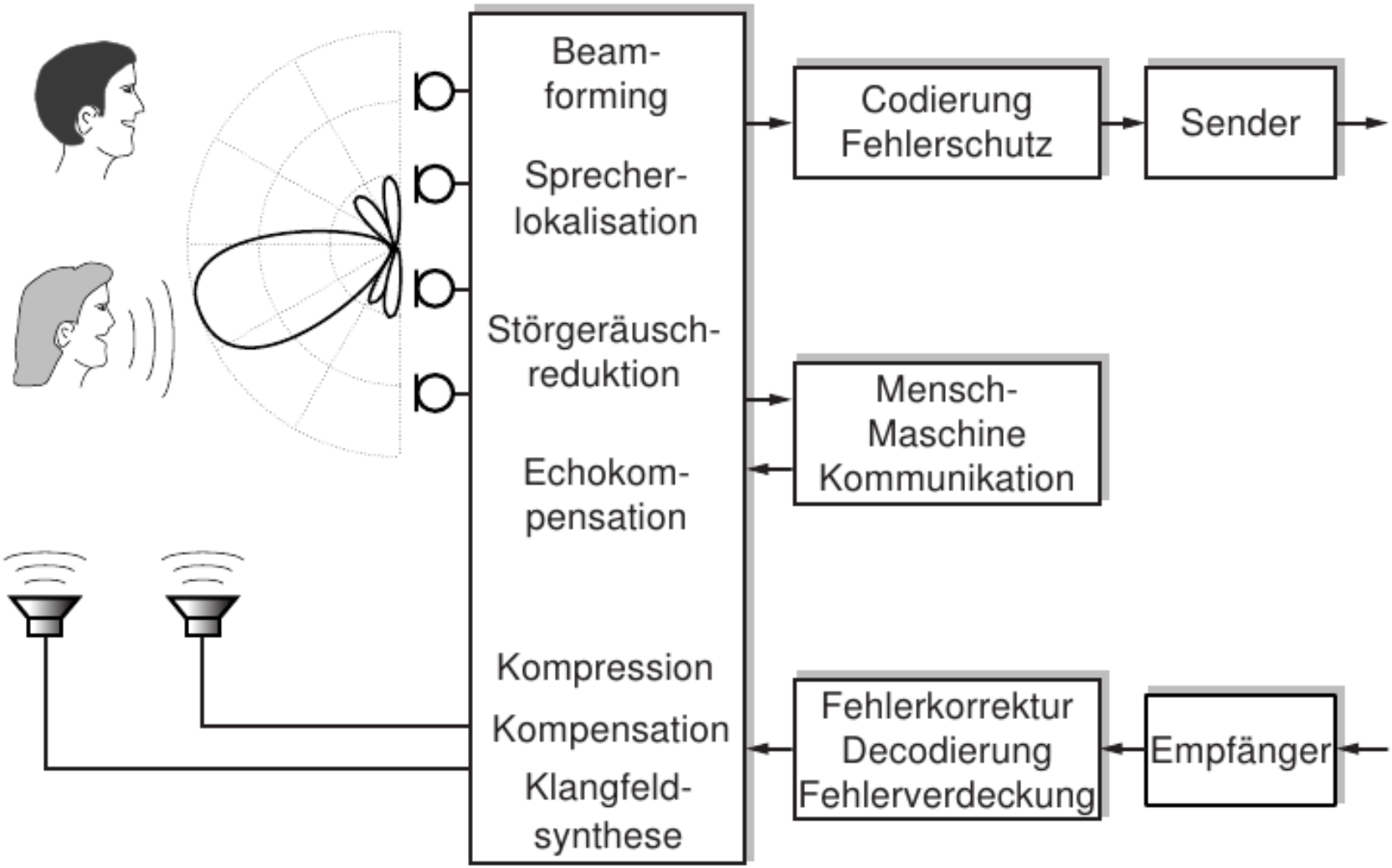
Video recordings from http://robot-ears.eu

Speech Signal Processing in Smartphones

- speech coding
- noise reduction
- speech recognition
  - speech control
  - virtual assistant (incld. speech synthesis)

The beginning: Dudleys Voder (1939)

Algorithms for Speech Signal Processing

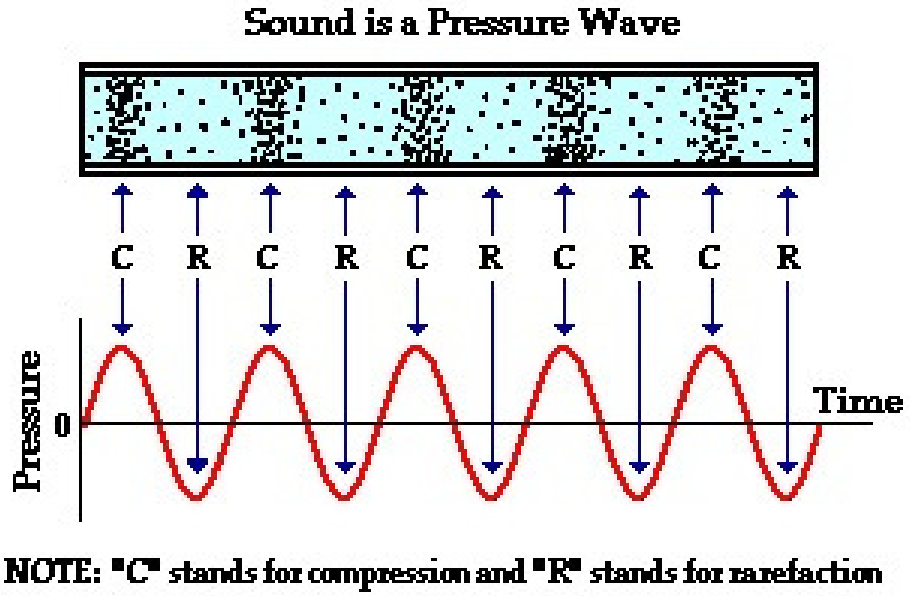


Quelle: R. Martin, Ruhr-Universität Bochum, 2009

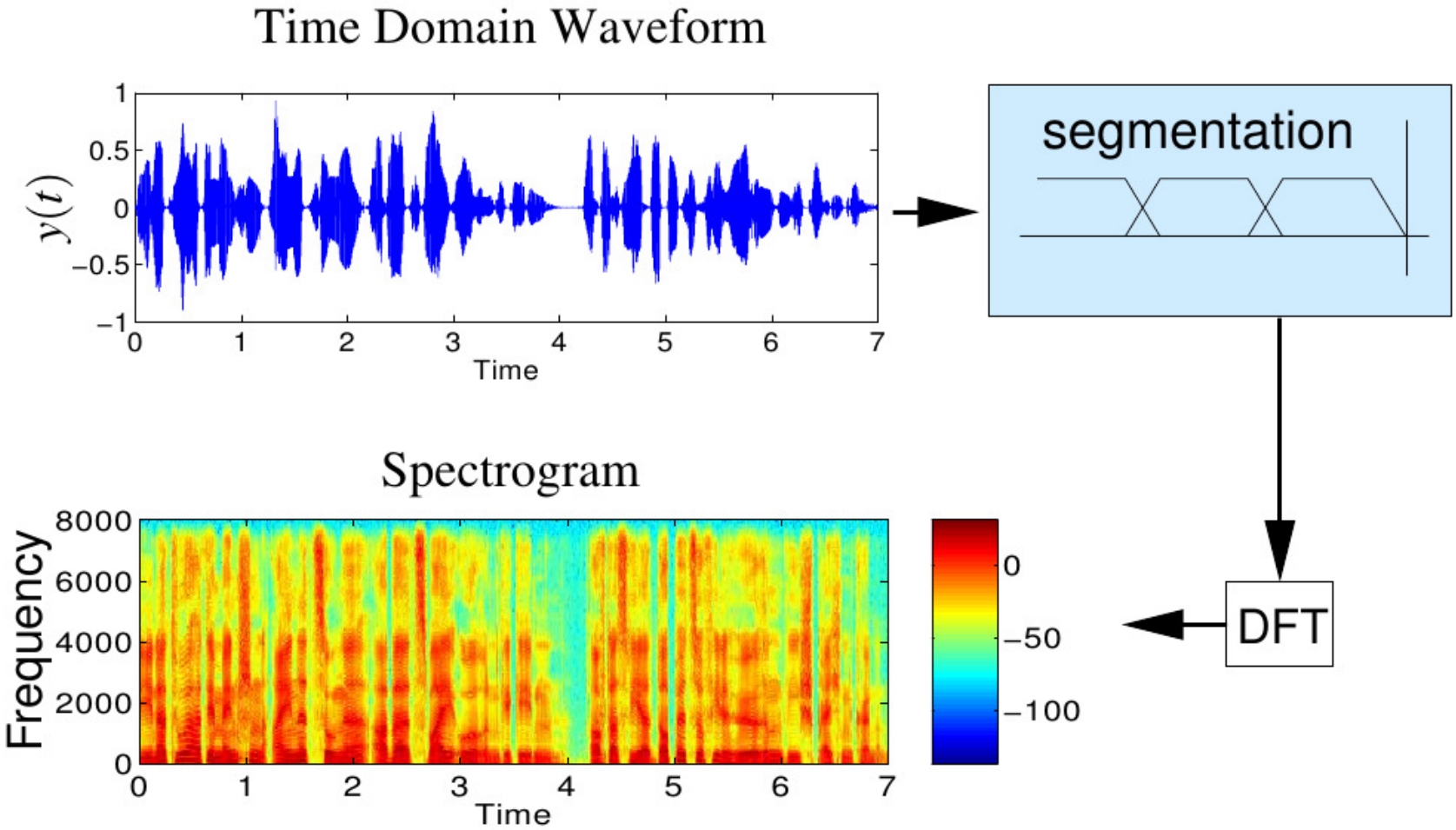
Applications I

Apple Siri
Google Now / Google Glasses
Microsoft Cortana
Amazon Echo

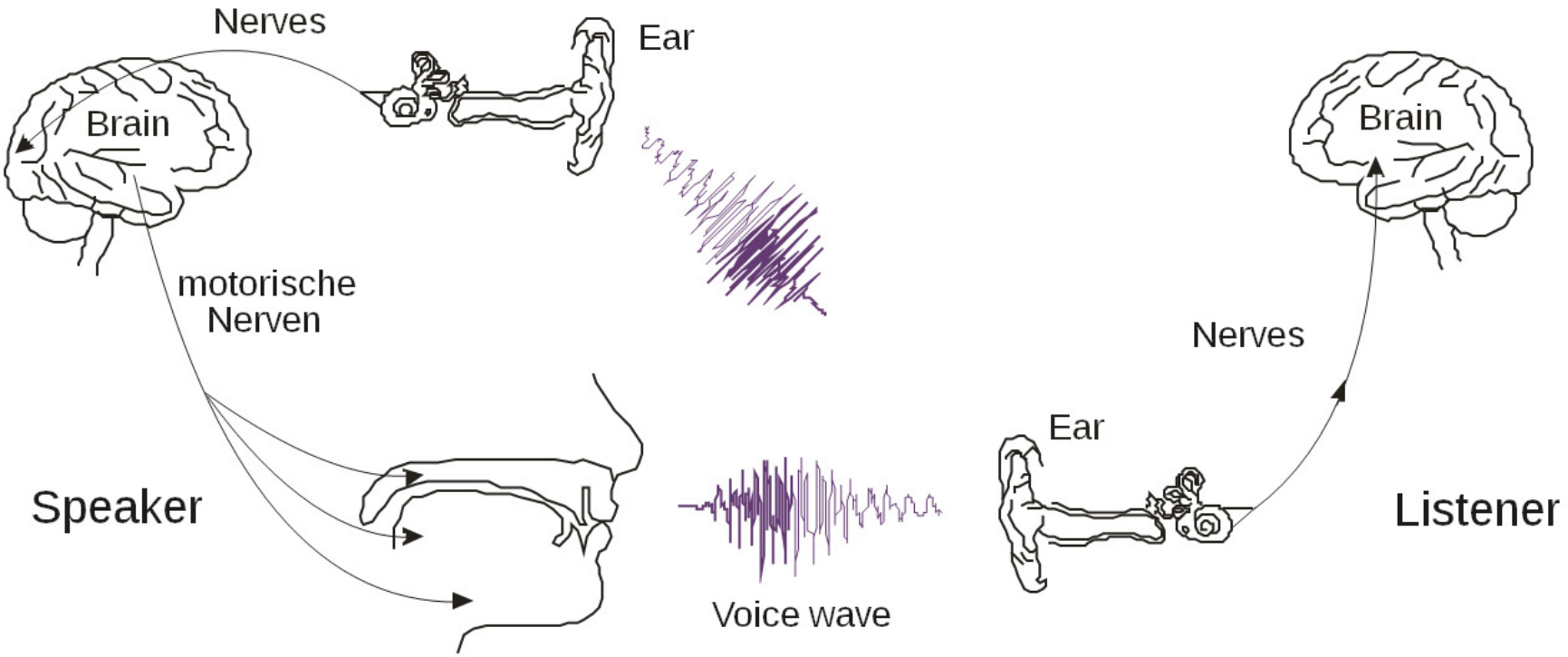
Speech is a Specific Pressure Wave



Speech is a Specific Pressure Wave



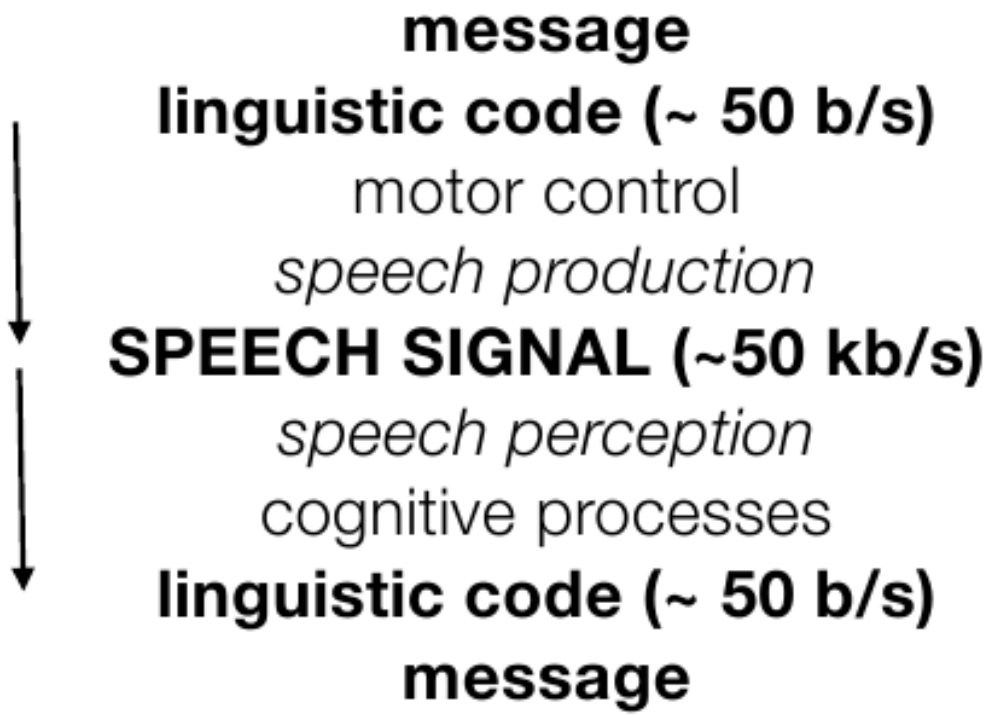
Communications Chain



Chain:

Quelle: Vary, Heute, Hess, Digitale Sprachsignalverarbeitung

Human Speech Communication



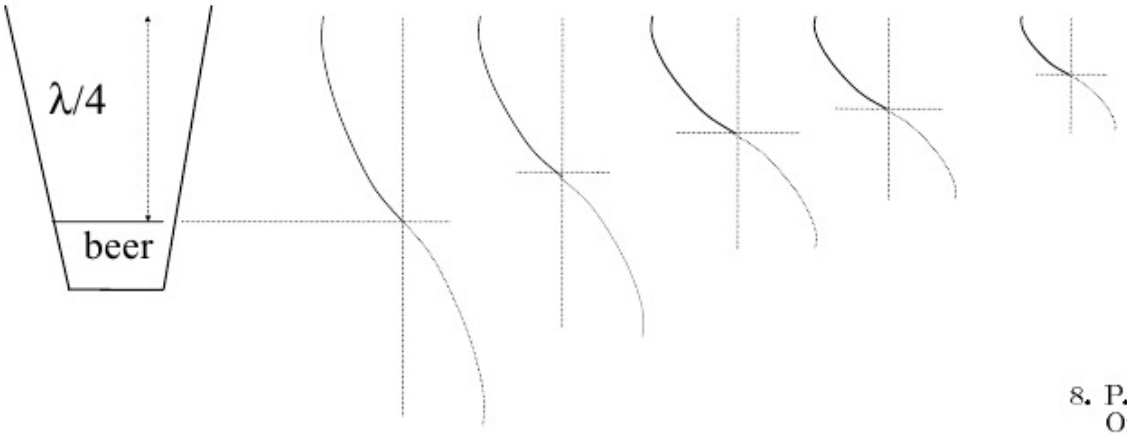
Quelle: Hermansky, lecture on feature extraction, 2005

Evolution and Constraints

- Using speech we can transmit information also under challenging conditions
  - Noise,
  - Long distances between speaker and listener,
  - Constraints due to other tasks of the vocal tract
    - Eating
    - Breathing
    - Smelling

Historic roots of speech research: Newton

In 1665 Isaac Newton made the following observation: *'The filling of a very deepe flaggon with a constant streame of beere or water sounds yer vowells in this order w, u, ω, o, a, e, i, y'* [8]. What young Newton observed was the spectral resonance peak which enhanced the spectrum of the beer pouring sound and moved up in frequency as the "deepe flaggon" was filling up. Since then, attempts to find acoustic correlates of phonetic categories mostly followed Newton's lead and studied the spectrum of speech.



(Hermansky & Sharma, 1998)

8. P. Ladefoged. *Three Areas of Experimental Phonetics.* Oxford University Press, 1967.

Historic Roots of Speech Research: Dudley

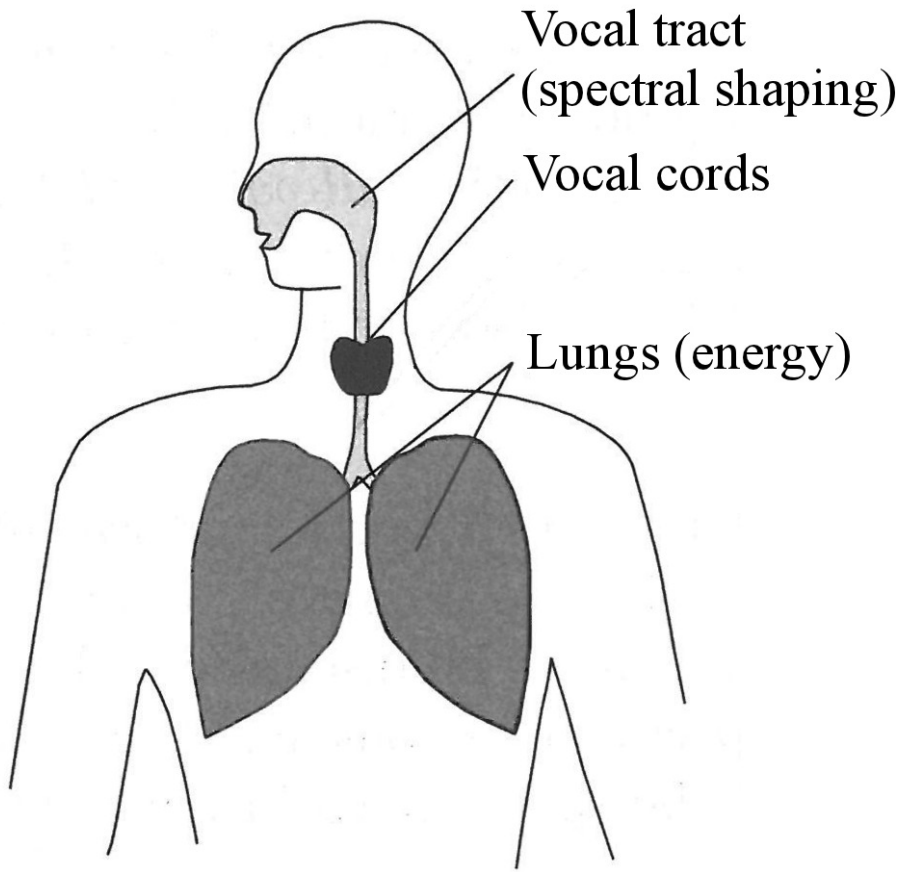
- Homer Dudley (1898 – 1981)
- Changes of sound pressure as a function of time

Outline

1. Introduction
   - Speech Production
   - Source-Filter Model
   - Hearing
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Automatic Speech Recognition

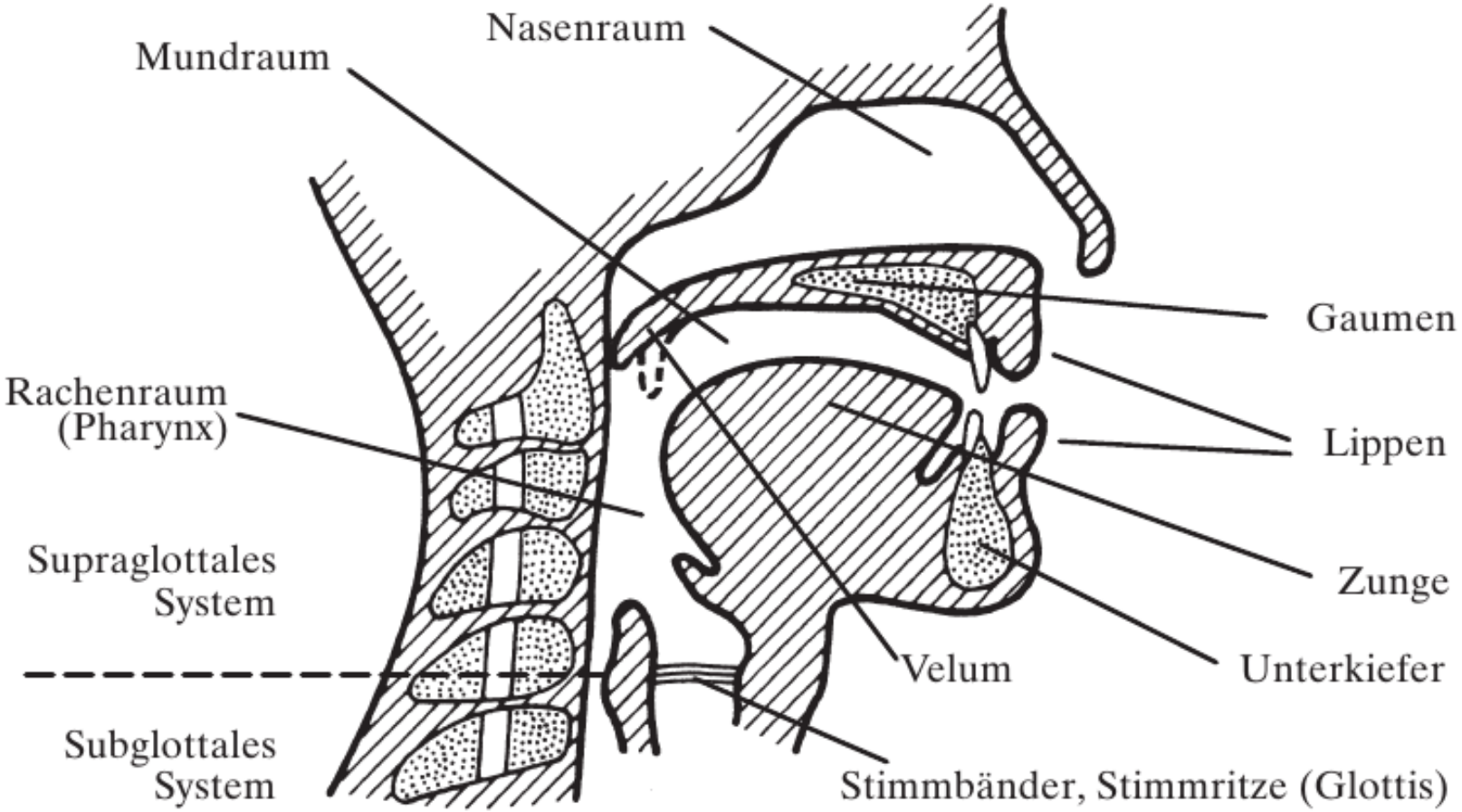
Speech Production I

- Lungs produce air flow
- In the larynx (*Kehlkopf*) the vocal cords start vibrating and produce sound
- in the vocal tract, the sound is formed to produce a speech sound.



Quelle: Vary, Martin (2006): Digital Speech Transmission, Wiley

Speech Production II



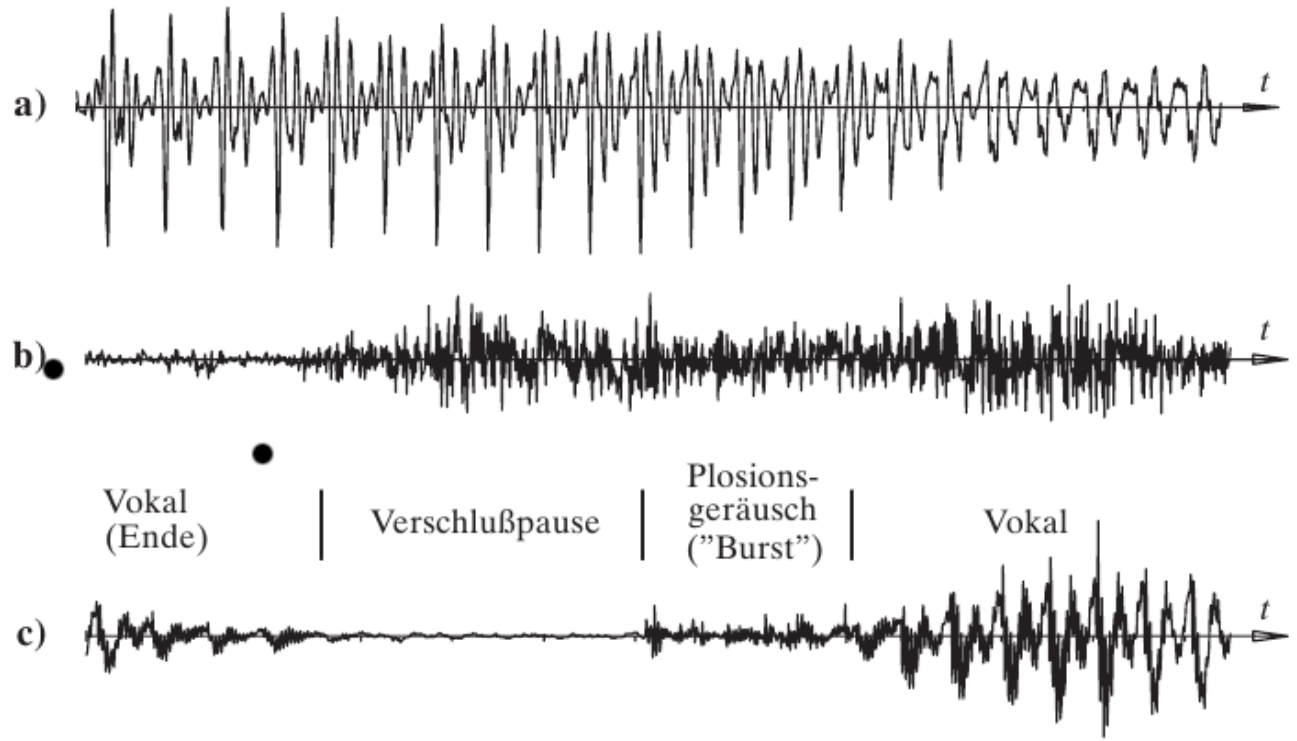
Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

Speech Sounds I

The most important speech sounds are

- voiced sounds
  - vowels (a,e,i,o,u)
  - sounds with mixed excitation (/v/)
- unvoiced sounds
  - fricative (/s/,/th/,/sh/)
  - plosive (/k/,/p/,/t/)

Speech Sounds II



a) stimmhaft b) stimmlos c) Übergang Vokal-Plosiv-Vokal

Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

a) voiced b) unvoiced c) transition vowel-plosive-vowel

Illustration of the Word “Dish”

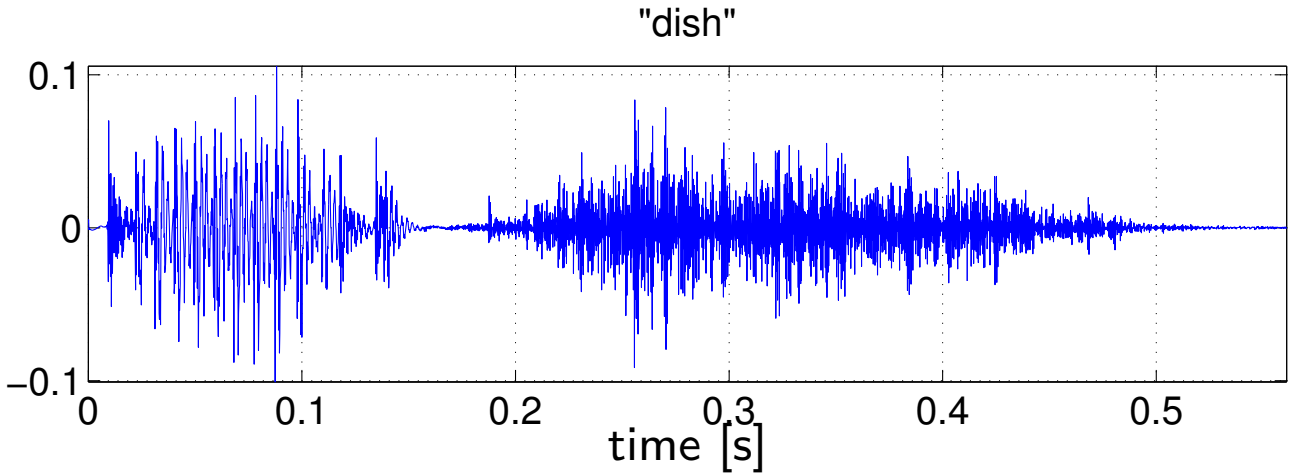
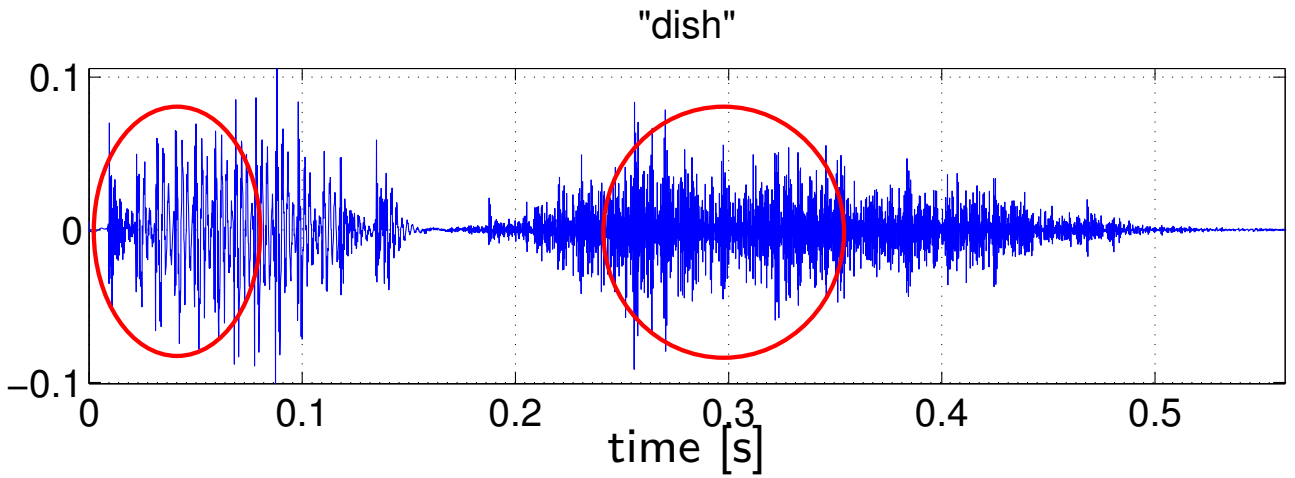
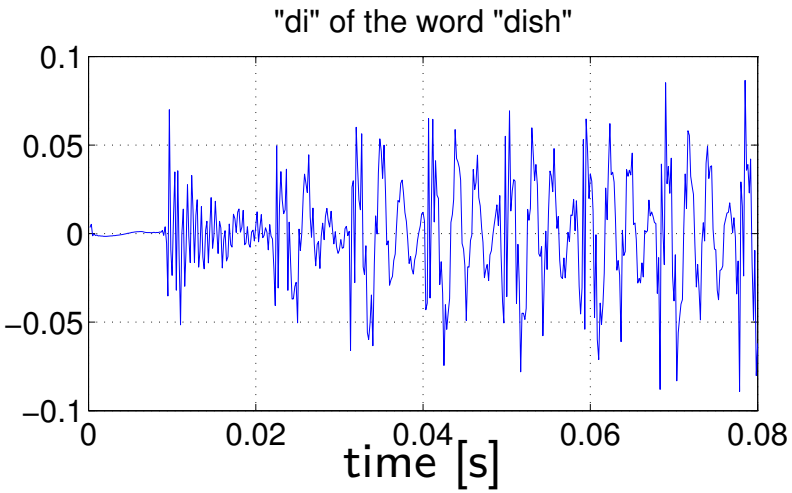
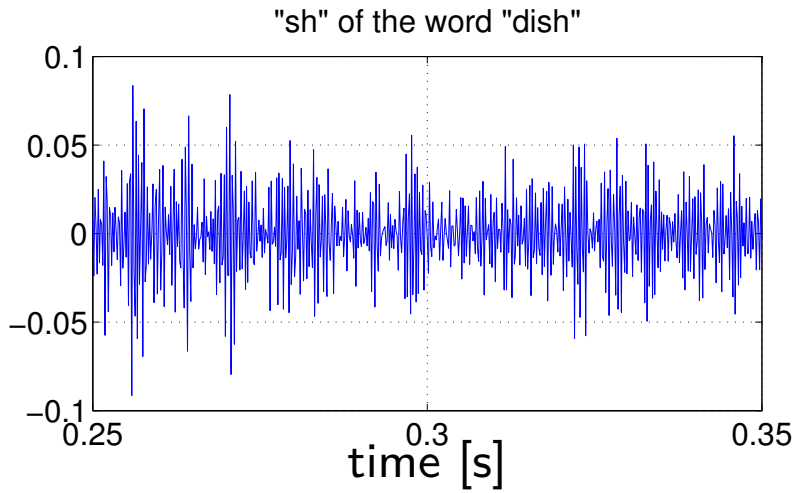


Illustration of the Word “Dish”







Phonetic Terms

Phone: Smallest speech segment with distinct physical or perceptual properties

Phoneme: The smallest contrastive linguistic unit which may bring about a change of meaning. One phoneme consists of a set of phones that are thought of as the same element within the phonology of a particular language ($\rightarrow$ (allophones)).

Allophone: one phone of the many that constitute a phoneme

Examples:

- kiss vs. kill; different in phonemes /I/ and /s/
- cat, kit, school, skill: they all contain the phoneme /k/, but are pronounced differently
- German: Das gerollte und nicht gerollte 'r' sind sind unterschiedliche Phone des gleichen Phonems /r/, und somit Allophone.

Phonemes

- Natural human languages have between 10 and 80 phonemes
- The German language has about 40 Phonemes (20 vowel phonemes, 20 consonant phonemes)
- English: 24 consonant phonemes, 20 vowel phonemes

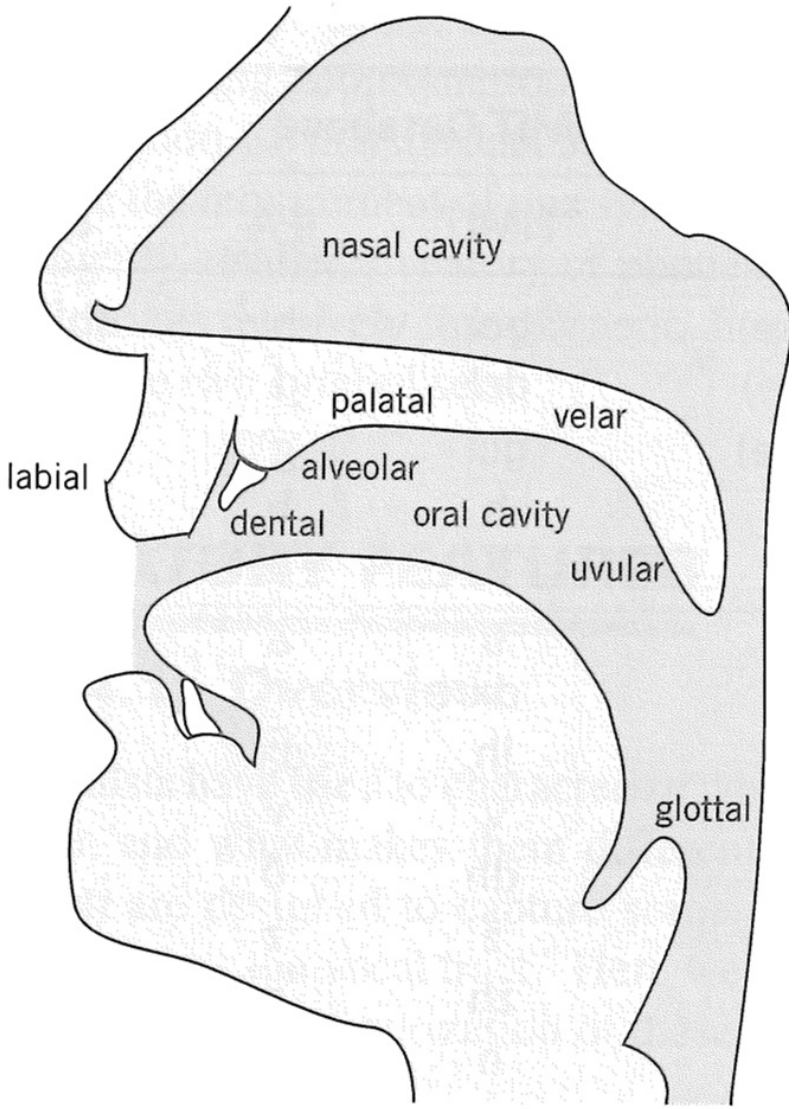
Phonemes are characterized by

- The way of articulation
  - Vowel
  - Nasal
  - Fricative
  - Plosive
  - ...
- Excitation signal
  - Voiced / unvoiced (noise-like by constrictions of the vocal tract)
- Place of articulation

Place of Articulation

Labial:	Lips
Bilabial:	upper and lower lip, e.g. /b/
Dental:	teeth
Alveolar:	socket of the superior teeth (German: oberer Zahndamm), e.g. /t/
Retroflex:	tongue between alveolar (*Zahndamm*) and the hard palate (*Gaumen*); American 'r' in "shore"
Palatal:	hard palate (*vorderer harte Gaumen*); German "ich"
Velar:	soft palate, e.g. /g/
Uvular:	back of the tongue against or near the uvula (*Gaumenzäpchen*); German: allophone of /r/ in "Rübe"
Pharyngeal:	root of the tongue against pharynx (*Rachen*); e.g. arabic pressed "h"
Glottal:	articulated with glottis; "h" in "hat"; German verreisen vs. vereisen (glottal stop, *Glottisschlag*));

Place of Articulation



Int. Phonetic Alphabet: Consonants

Stelle / Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ	n̪	n		ɳ	ɲ	ŋ	ɴ		
Affrikate			t͡s d͡z		t͡ʃ d͡ʒ						
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateraler Frikativ				ɬ ɮ							
Trill	ʙ			r					ʀ		
Flap				ɾ		ɽ					
Approximant	ʍ	ʋ			ɹ	ɻ	j	(w)			
Lateral approximant				l		ɭ	ʎ				

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Demo for speech sounds: `http://soundsofspeech.uiowa.edu/index.html`

Int. Phonetic Alphabet: Consonants – German

Stelle / Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ	n̪	n		ɳ	ɲ	ŋ	ɴ		
Affrikate			t͡s d͡z		t͡ʃ d͡ʒ						
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateraler Frikativ				ɬ ɮ							
Trill	ʙ			r					ʀ		
Flap				ɾ		ɽ					
Approximant	ʍ	ʋ			ɹ	ɻ	j	(w)			
Lateral approximant				l		ɭ	ʎ				

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Demo for speech sounds: `http://soundsofspeech.uiowa.edu/index.html`

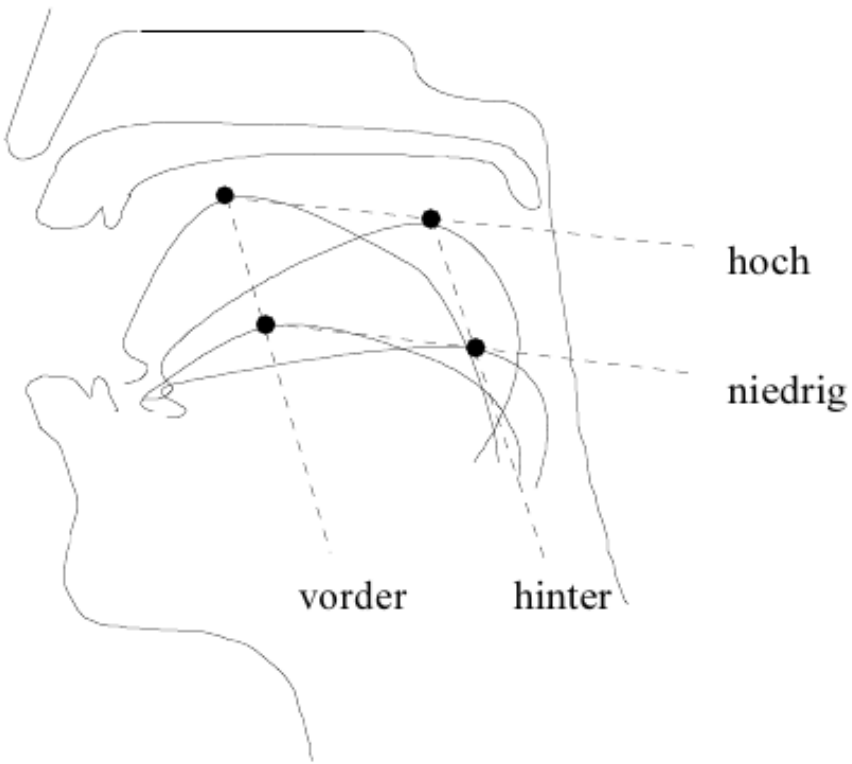
Int. Phonetic Alphabet: Consonants – English

Stelle / Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ	n̪	n		ɳ	ɲ	ŋ	ɴ		
Affrikate			t̪s̪ d̪z̪		t͡ʃ d͡ʒ						
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateraler Frikativ				ɬ ɮ							
Trill	ʙ			r					ʀ		
Flap				ɾ		ɽ					
Approximant	w	ʋ			ɹ	ɻ	j	(w)			
Lateral approximant				l		ɭ	ʎ				

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Demo for speech sounds: `http://soundsofspeech.uiowa.edu/index.html`

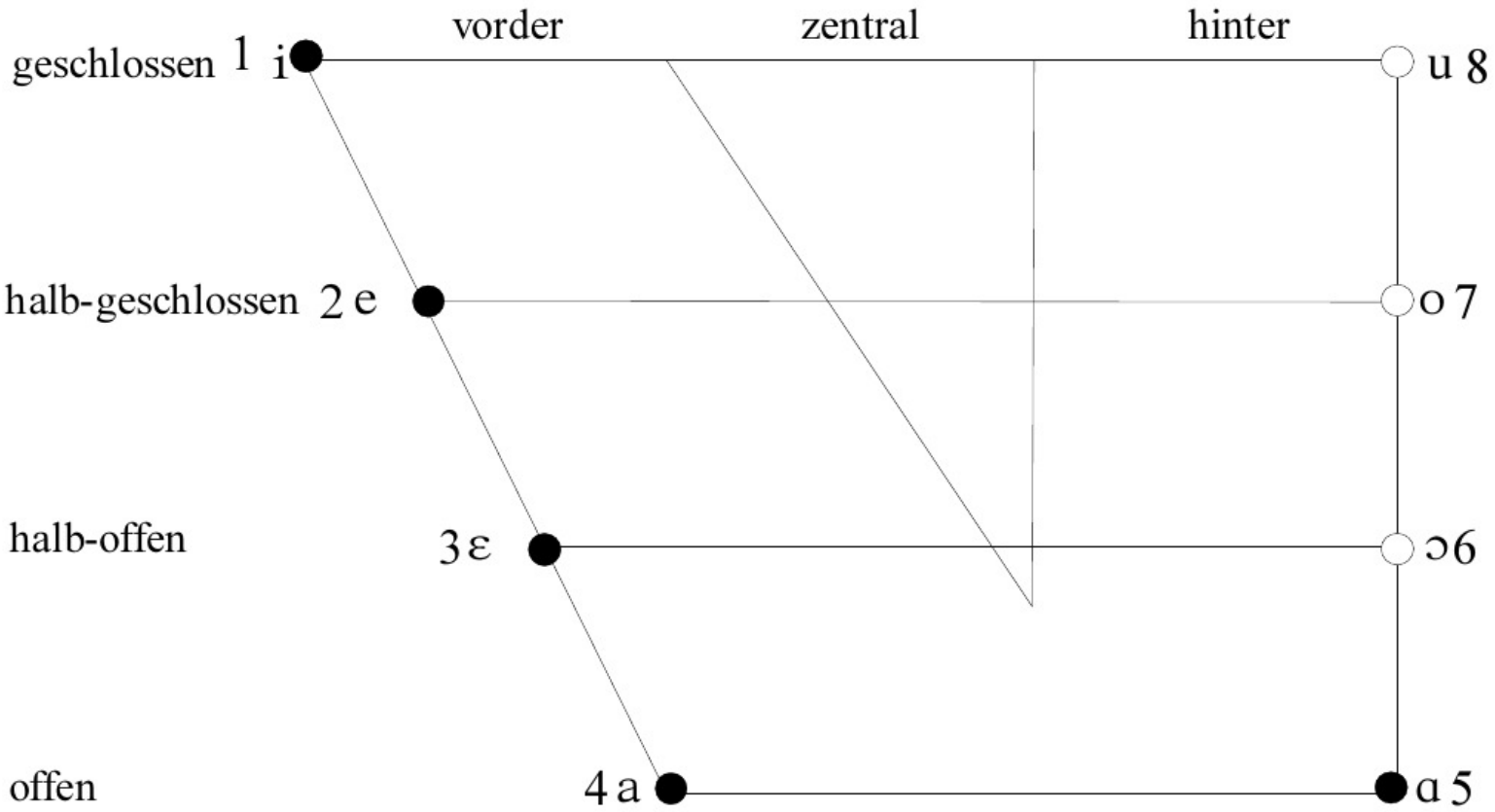
Cardinal Vowels I



- Cardinal vowels are used to describe the position of the tongue in the oral cavity
- Cardinal vowels describe extreme positions of the tongue. In this form they do not necessarily appear in natural speech.
- Two dimensions for tongue position
  - horizontal (front, back)
  - vertical (high, low)

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Primary Cardinal Vowels

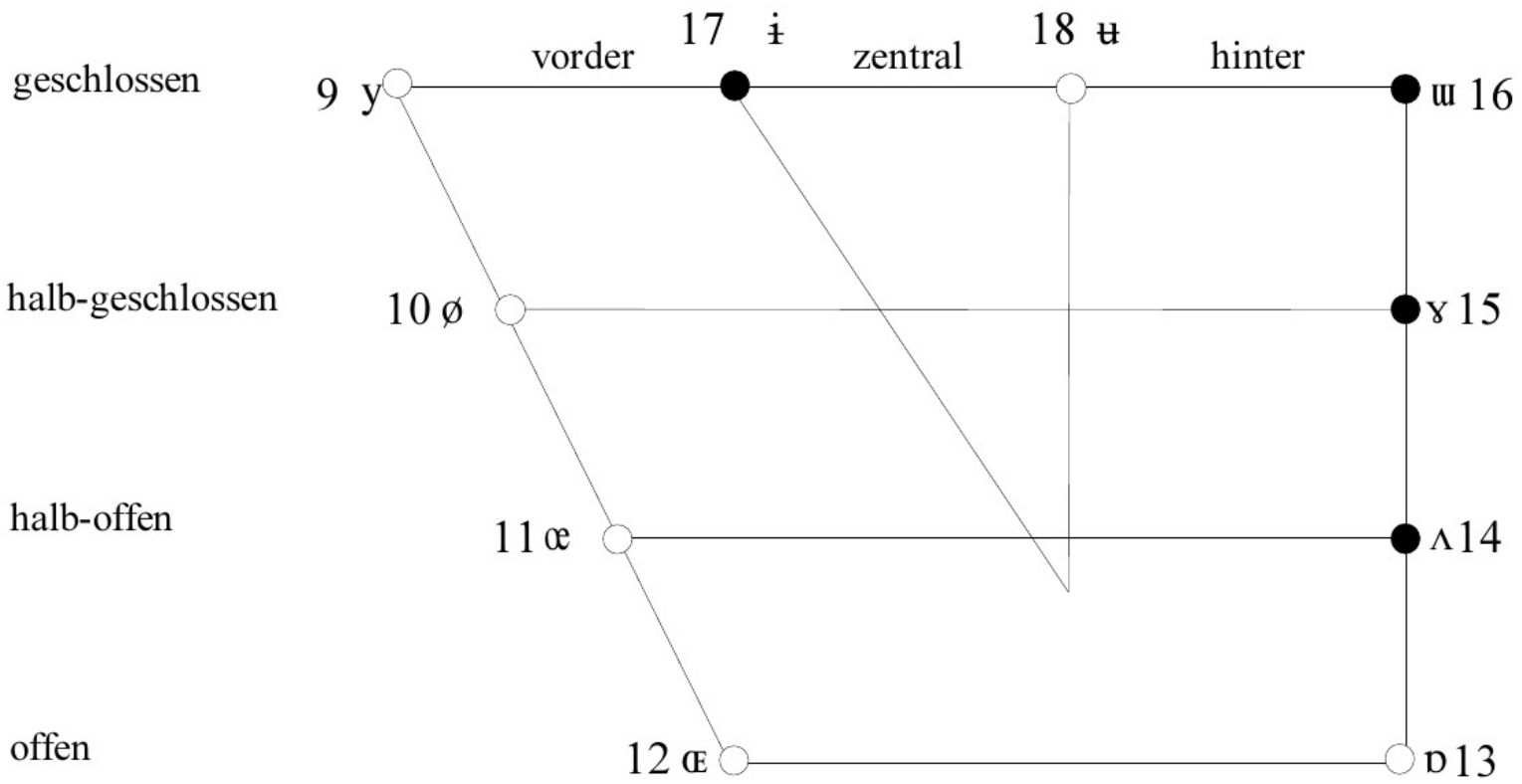


○: round lips

●: open lips

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Secondary Cardinal Vowel



○: round lips

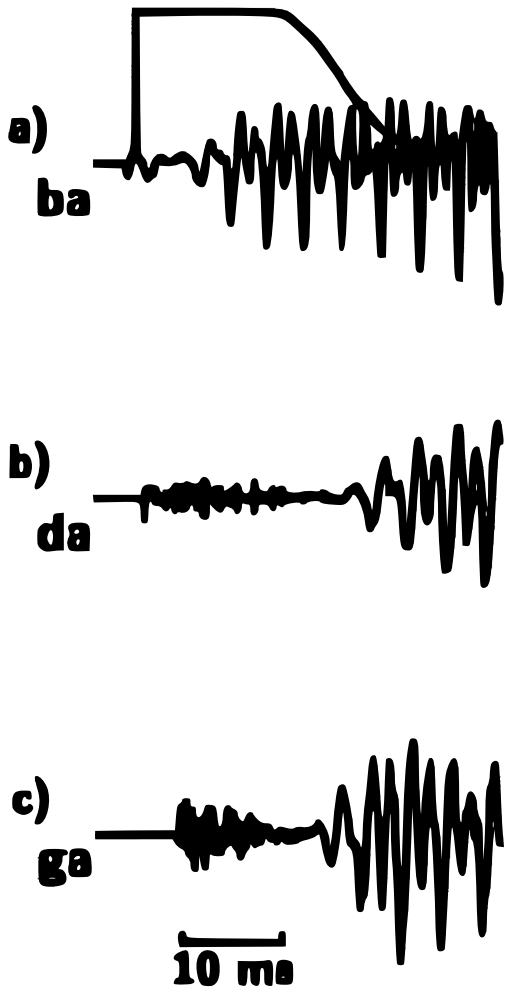
●: open lips

Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

Original recordings D. Jones

Co-Articulation

- The shape of the vocal tract can not change instantly
- depending on the context, phonemes are pronounced differently
- Examples:
  - ”n“ in ”tenth“
    - ”n“ usually alveolar (Zahndamm)
    - for ”tenth“ rather dental
  - ”s“ in ”seat“ vs ”suit“
  - ”ku“, ”ki“
    - since ”u“ requires round lips, while ”i“ requires open lips, the ”k“ sounds differently
  - ”ba“, ”da“, ”ga“
    - for ”b“,”d“,”g“ the place of articulation moves towards the



Prosody

- ➡ rhythm, stress, and intonation of speech
  - reflects
    - emotional state of the speaker
    - form of the utterance (statement, question, or command)
    - irony or sarcasm
    - emphasis, contrast and focus

Remark: Often, only the intonation is meant when we say 'prosody'. However, intonation is strictly speaking only part of the prosody.

Example: TIMIT Database

TABLE 23.2 TIMIT Phone Types

		Phones in the TIMIT Database			
TIMIT	**IPA**	**Example**	**TIMIT**	**IPA**	**Example**
pcl	p˺	(p closure)	bcl	b˺	(b closure)
tcl	t˺	(t closure)	dcl	d˺	(d closure)
kcl	k˺	(k closure)	gcl	g˺	(g closure)
p	p	**p**ea	b	b	**b**ee
t	t	**t**ea	d	d	**d**ay
k	k	**k**ey	g	g	**g**ay
q	ʔ	bat	dx	ɾ	dirty
ch	t͡ʃ	**ch**oke	jh	d͡ʒ	**j**oke
f	f	**f**ish	v	v	**v**ote
th	θ	**th**in	dh	ð	**th**en
s	s	**s**ound	z	z	**z**oo
sh	ʃ	**sh**out	zh	ʒ	azure
m	m	**m**oon	n	n	**n**oon
em	m̩	bott**om**	en	n̩	butt**on**
ng	ŋ	si**ng**	eng	ŋ̍	Washington
nx	ɾ̃	wi**nn**er	el	l̩	bott**le**
l	l	**l**ike	r	r	**r**ight
w	w	**w**ire	y	j	**y**es
hh	h	**h**ay	hv	ɦ	a**h**ead
er	ɝ	b**ir**d	axr	ɚ	butt**er**
iy	i	b**ee**t	ih	ɪ	b**i**t
ey	e	b**ai**t	eh	ɛ	b**e**t
ae	æ	b**a**t	aa	ɑ	f**a**ther
ao	ɔ	b**ough**t	ah	ʌ	b**u**t
ow	o	b**oa**t	uh	ʊ	b**oo**k
uw	u	b**oo**t	ux	ʉ	t**oo**t

Outline

1. Introduction
   - Speech Production
   - Source-Filter Model
   - Hearing
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Automatic Speech Recognition

Source-Filter Model

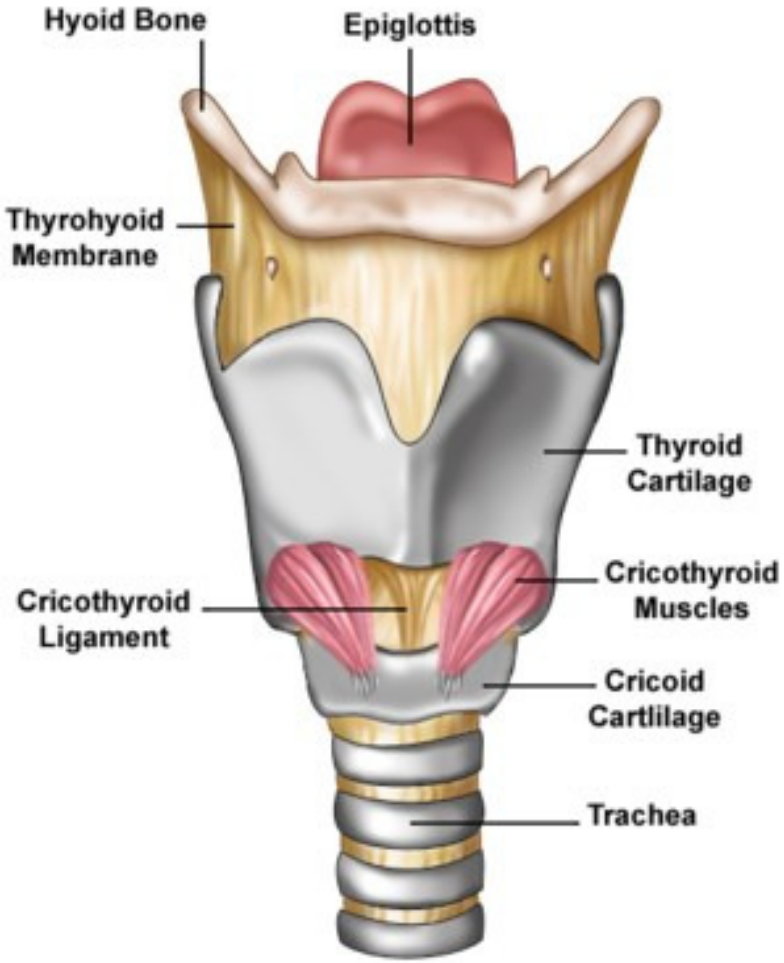
- The production of a speech signal can be described using a source-filter model.

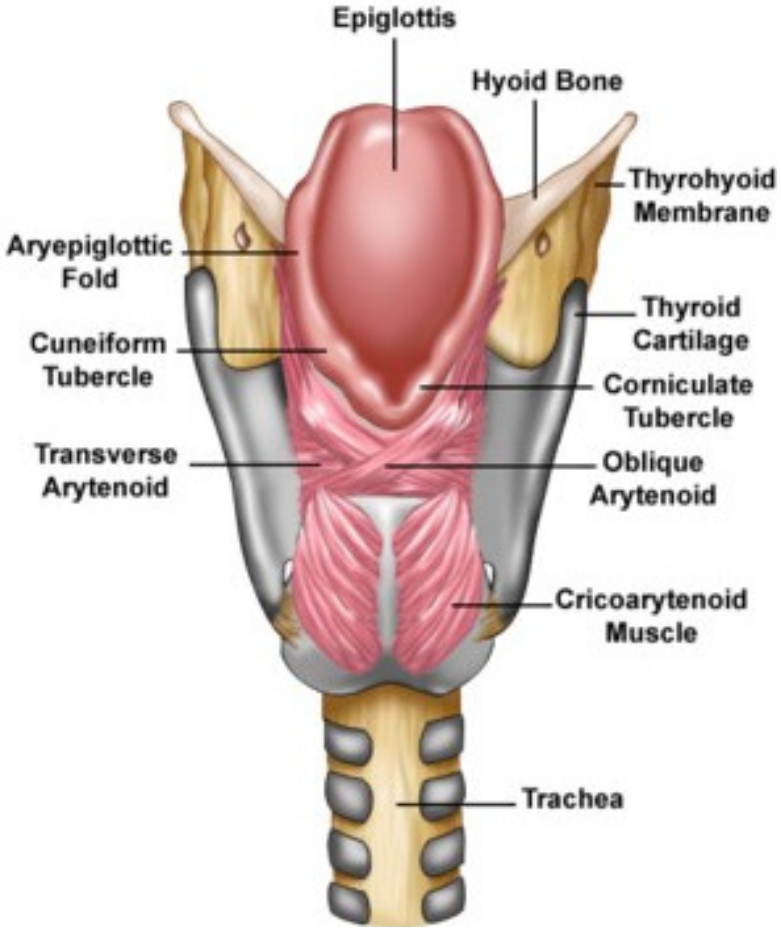
- Simplifying assumption: source and filter are mutually independent

  Source: air flow, vibration of vocal cords

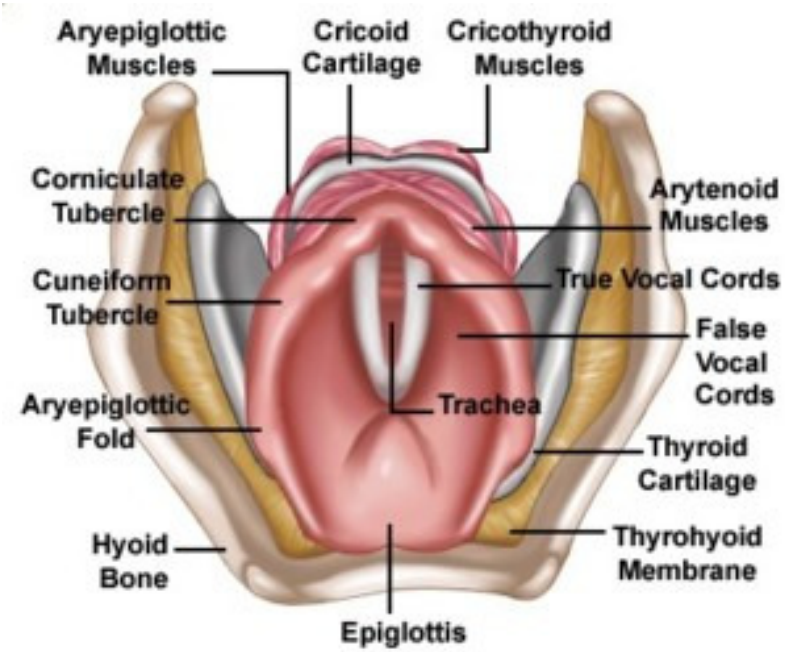
  Filter: Shape of the vocal tract: Position of tongue, lips, palate, ...

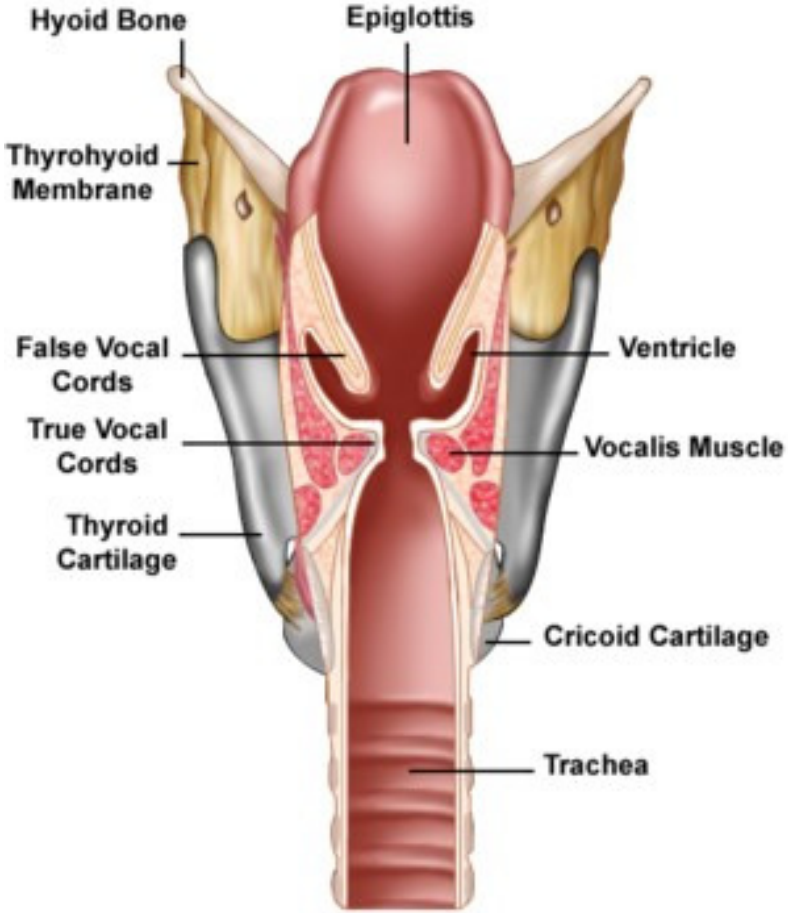
Larynx (*Kehlkopf*) I





Larynx (*Kehlkopf*) II

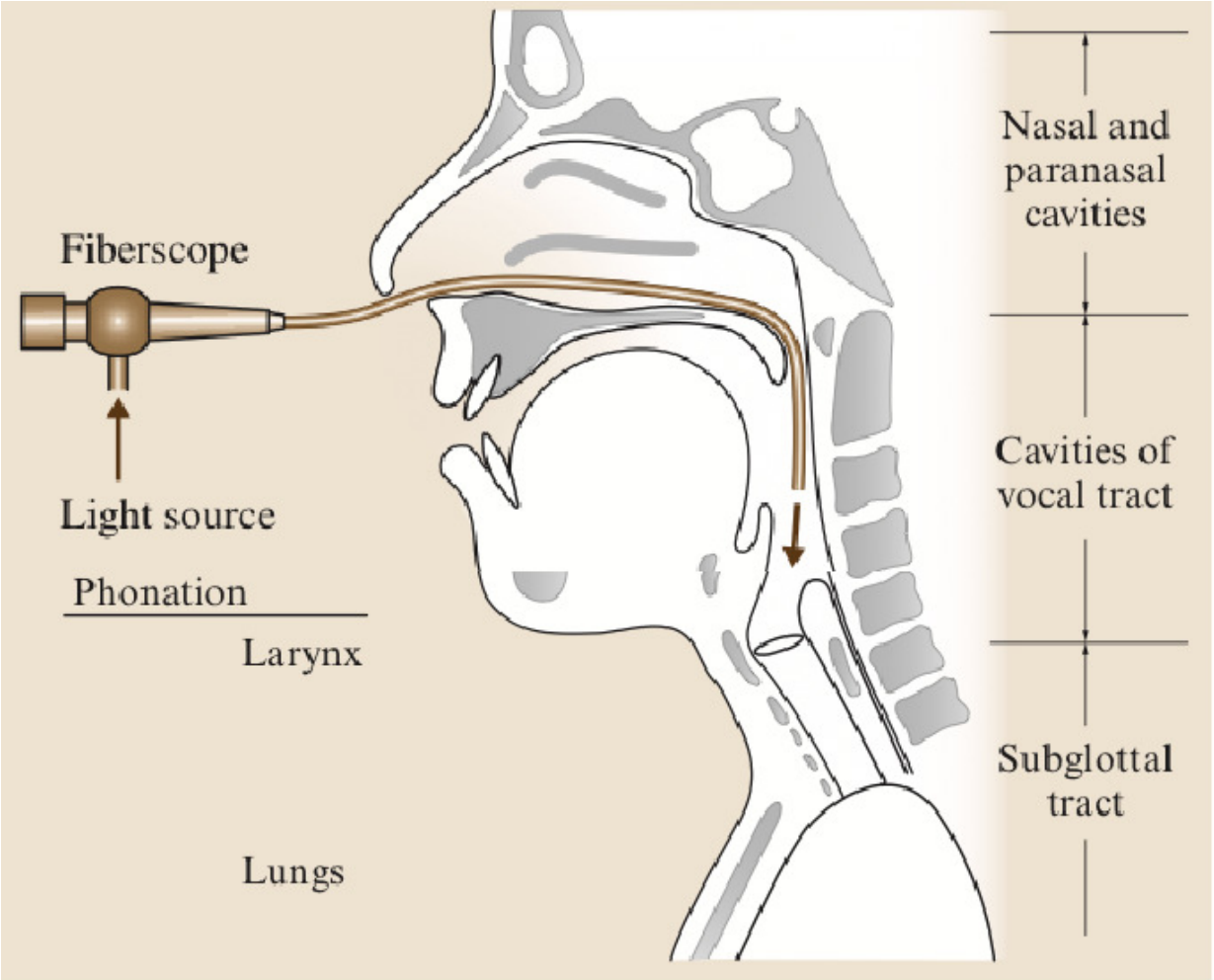




Larynx (*Kehlkopf*) III

Video

Larynx (*Kehlkopf*) IV



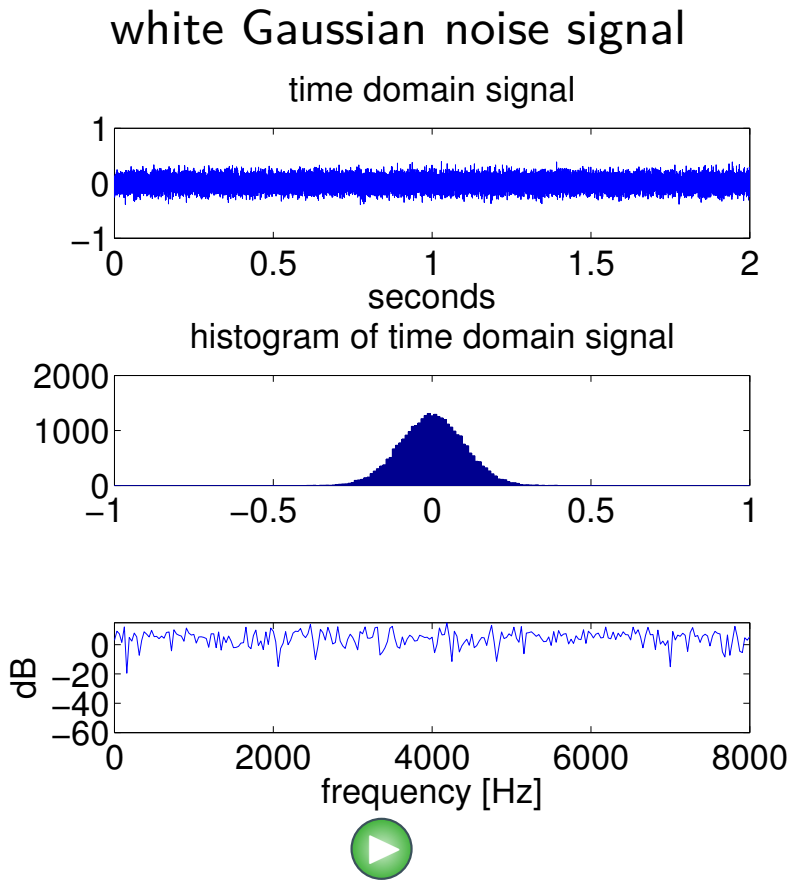
Springer Handbook of Speech Processing, Benesty, Sondhi, Huang (Eds.), Springer, Berlin.

Fiberscope (*Endoskop*) recording

Video

Unvoiced excitation

- The unvoiced excitation is noise-like. Can be well described using Gaussian noise.

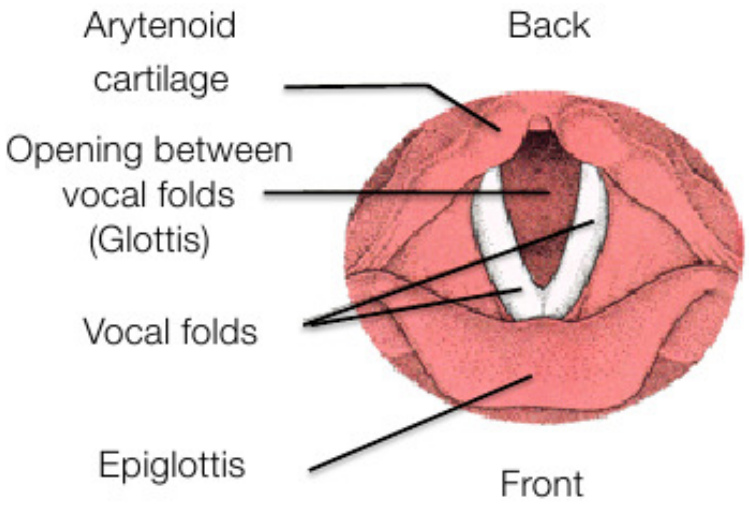


Voiced Excitation

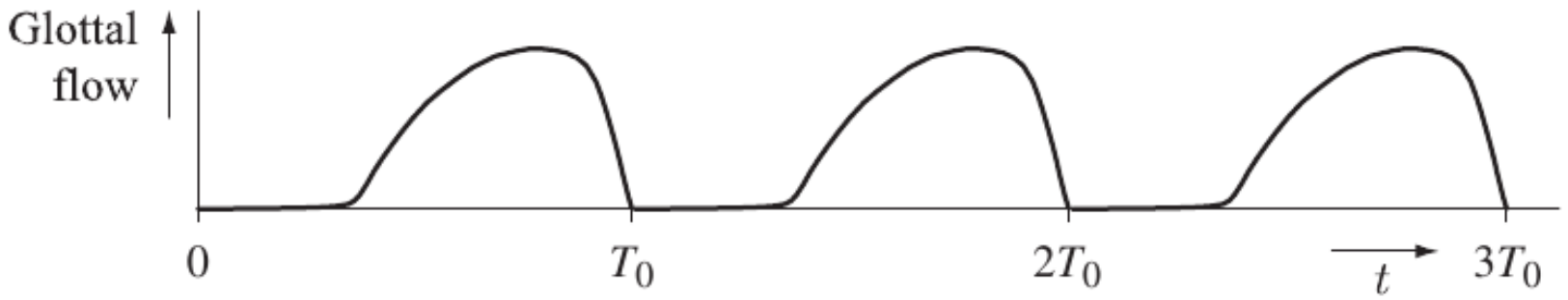
The excitation signal

voiced: Vocal folds vibrate, the frequency depends on physiological parameters

unvoiced: vocal folds are open. Constrictions in the vocal tract cause a turbulent air flow.



Air Flow for Voiced Excitation



Vocal cords produce a pulsating air flow through the vocal cords.

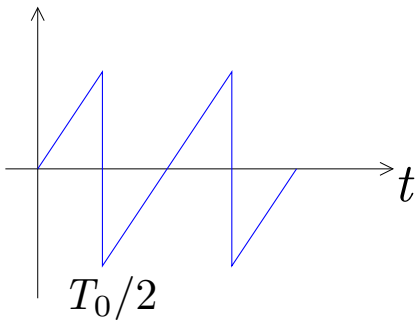
- Opening of glottis due to the increased pressure
- Air flows through the glottis, vocal cords are under tension
- Because of the constriction of the glottis, the flow velocity increases while the pressure decreases (Bernoulli-effect)
- The vocal cords snap together, the air flow is interrupted
- the pressure increases, the glottis opens up

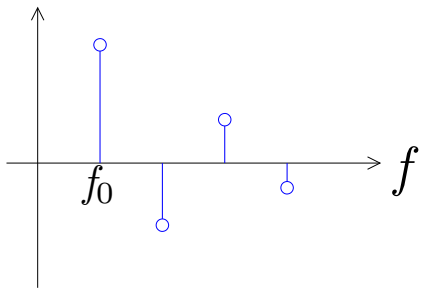
Fourier series: Every periodic function $g(t)$ with period T_0 can be represented by a series of sine and cosine functions, whose frequencies are integer multiples of the fundamental frequency $f_0 = 1/T_0$:

$$g(t) = \frac{a_0}{2} + \sum_{h=1}^{\infty} \left(a_h \cos(2\pi h f_0 t) + b_h \sin(2\pi h f_0 t) \right)$$

➡ The glottis signal consists of the fundamental oscillation and its harmonics.

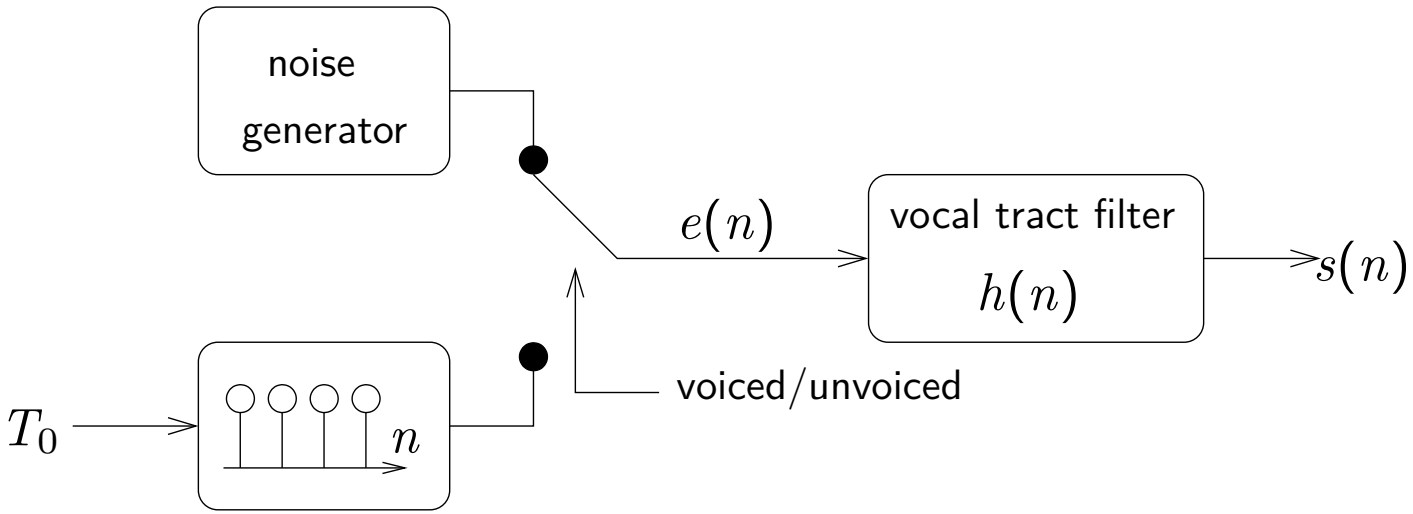
Example:



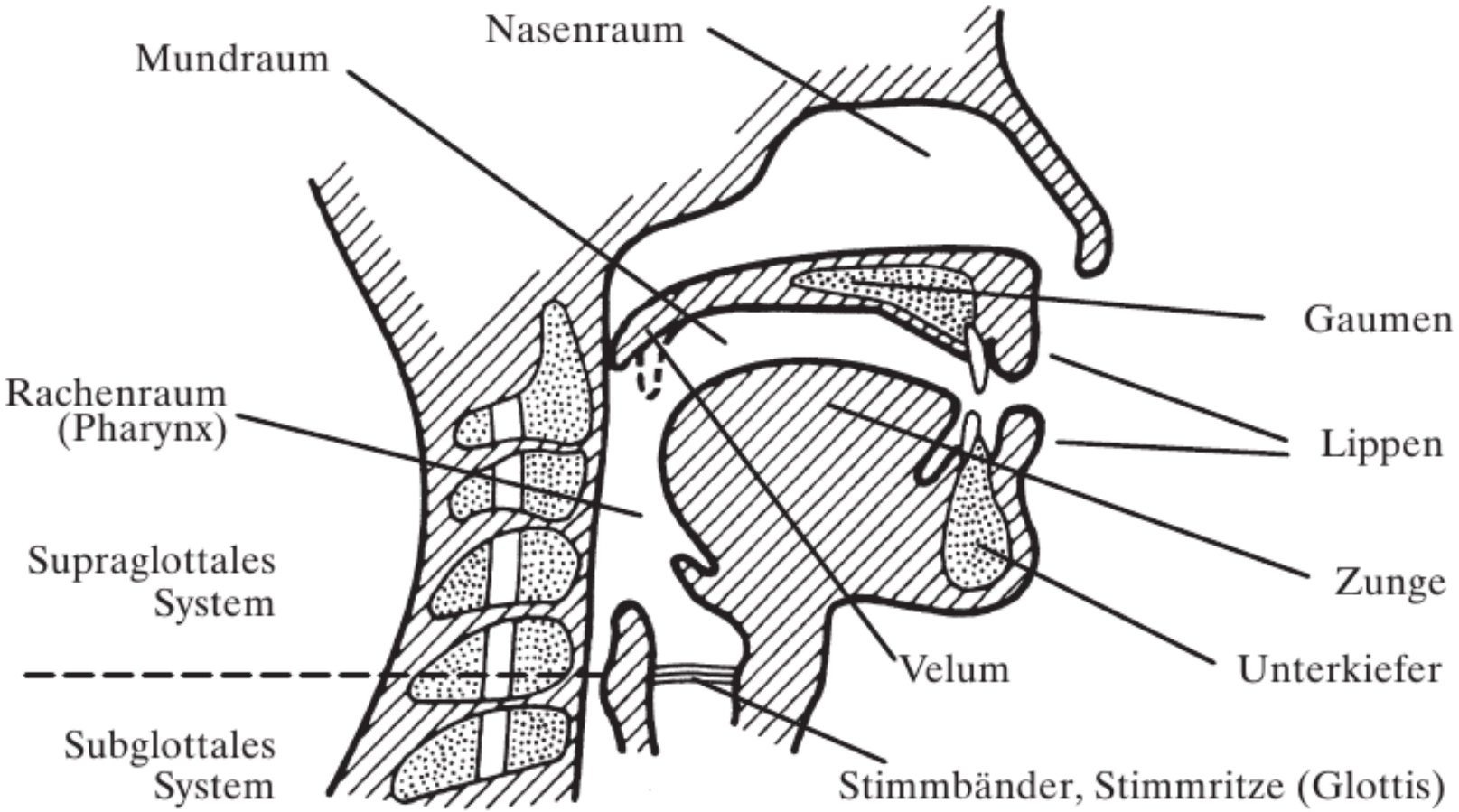


$$g(t) = \frac{1}{\pi f_0} \sum_{h=1}^{\infty} \frac{(-1)^{h-1}}{h} \times \sin(2\pi h f_0 t) \quad (1)$$

Simple Model of Speech Production

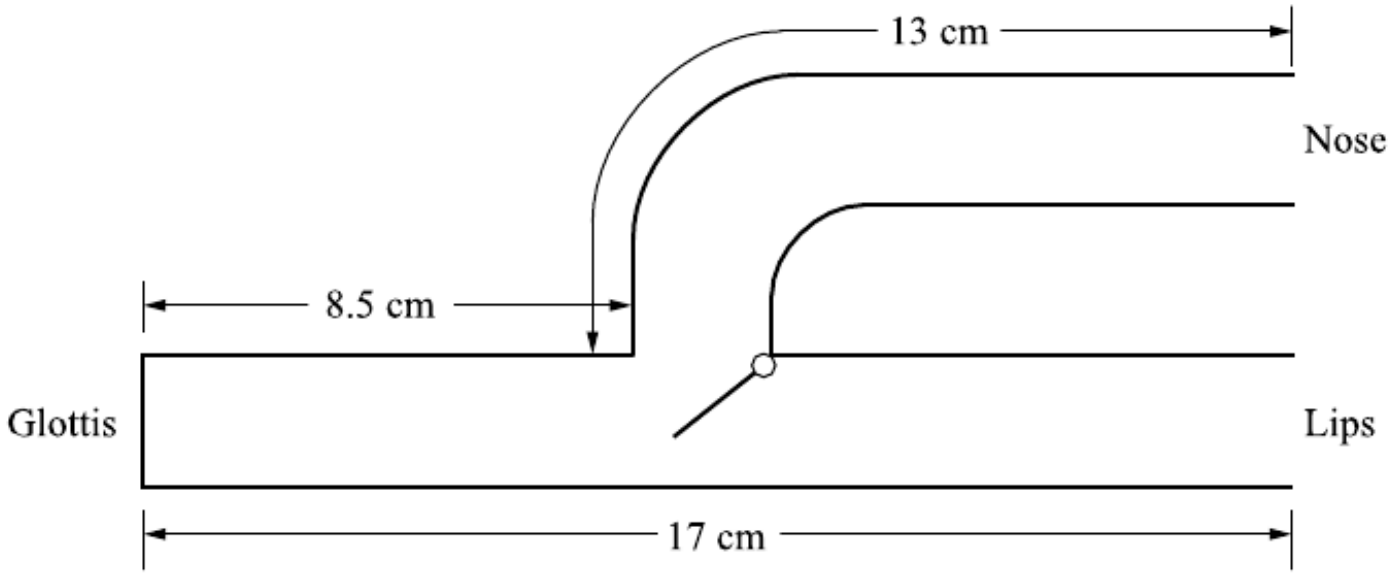


Sketch of the Vocal Tract



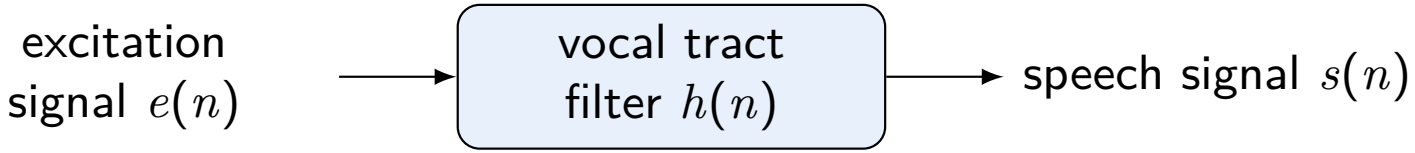
Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

Simple Model of the Vocal Tract



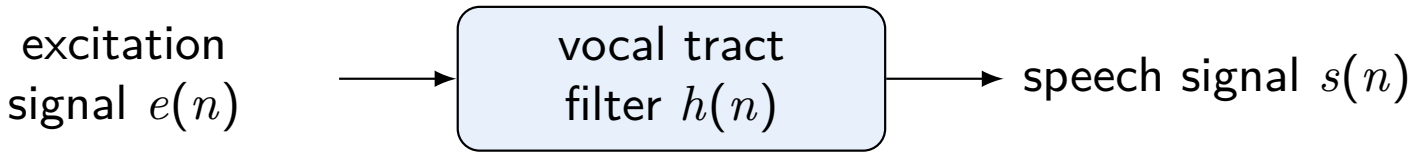
Model of Speech Production

- The vocal tract is modeled by the filter $h(n)$.



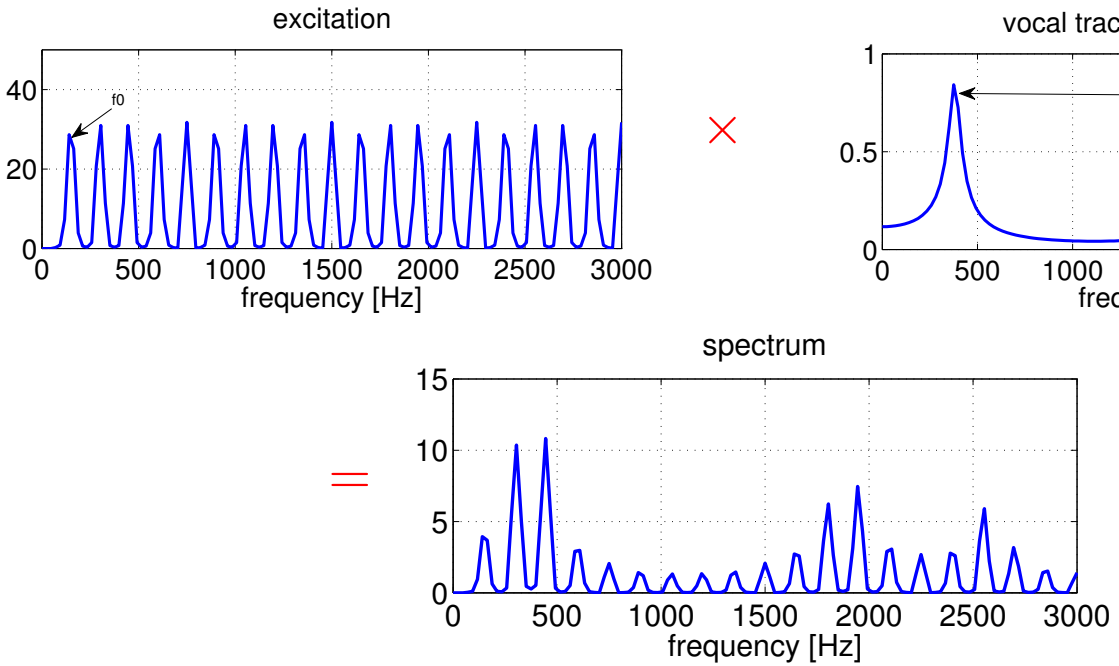
Model of Speech Production

- The vocal tract is modeled by the filter $h(n)$.



- $s(n) = e(n) * h(n) \quad \circ\!\!-\!\!\bullet \quad S(f) = E(f) \cdot H(f)$

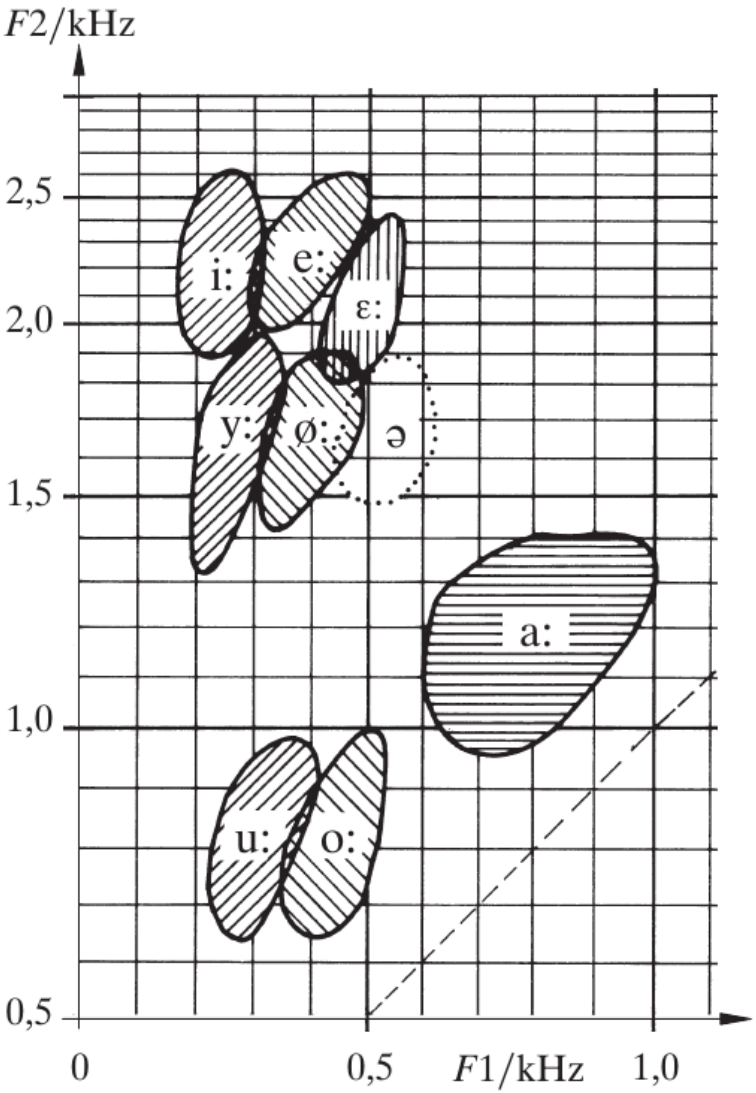
- Spectral decomposition for the utterance “i” in “dish”:

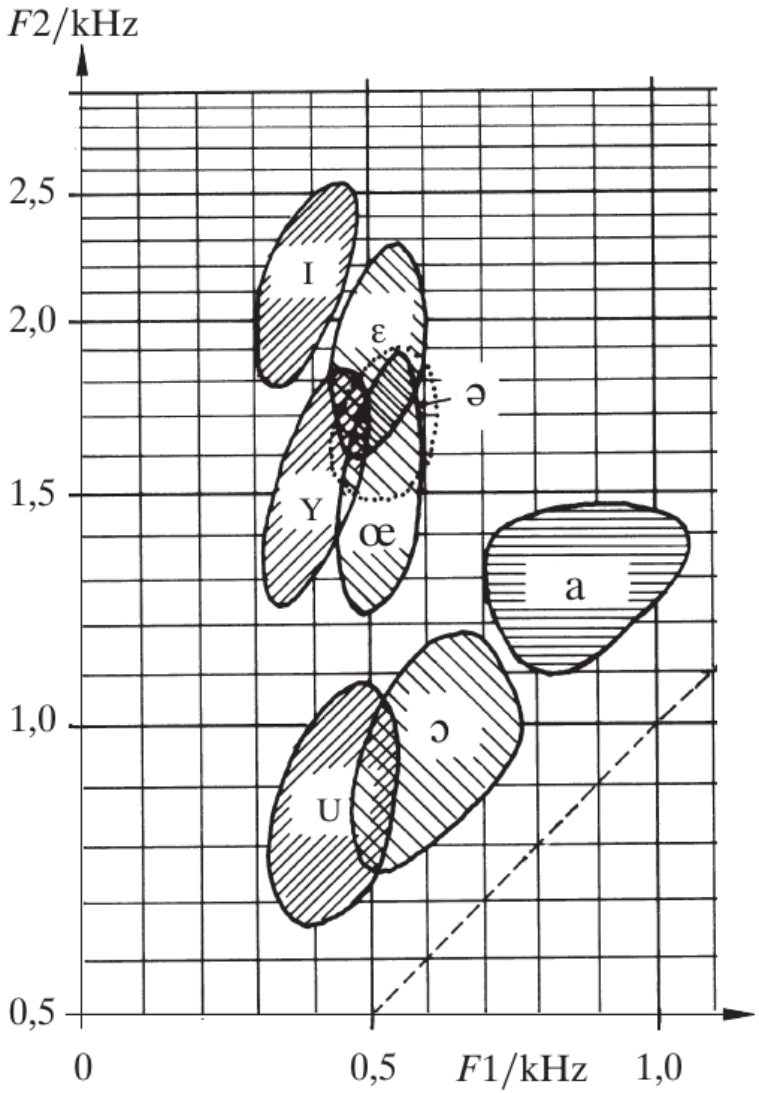


Formant frequency vs. fundamental frequency

- In the source-filter model, excitation (source) and filter (vocal tract) are treated as being independent.
  - Formants: Peaks of the spectral envelope, resonances of the vocal tract
    - ➡ defines the meaning of a phone
  - fundamental frequency: first peak of the spectral fine structure, and distance between spectral harmonics.

Formant Map for German Vowels



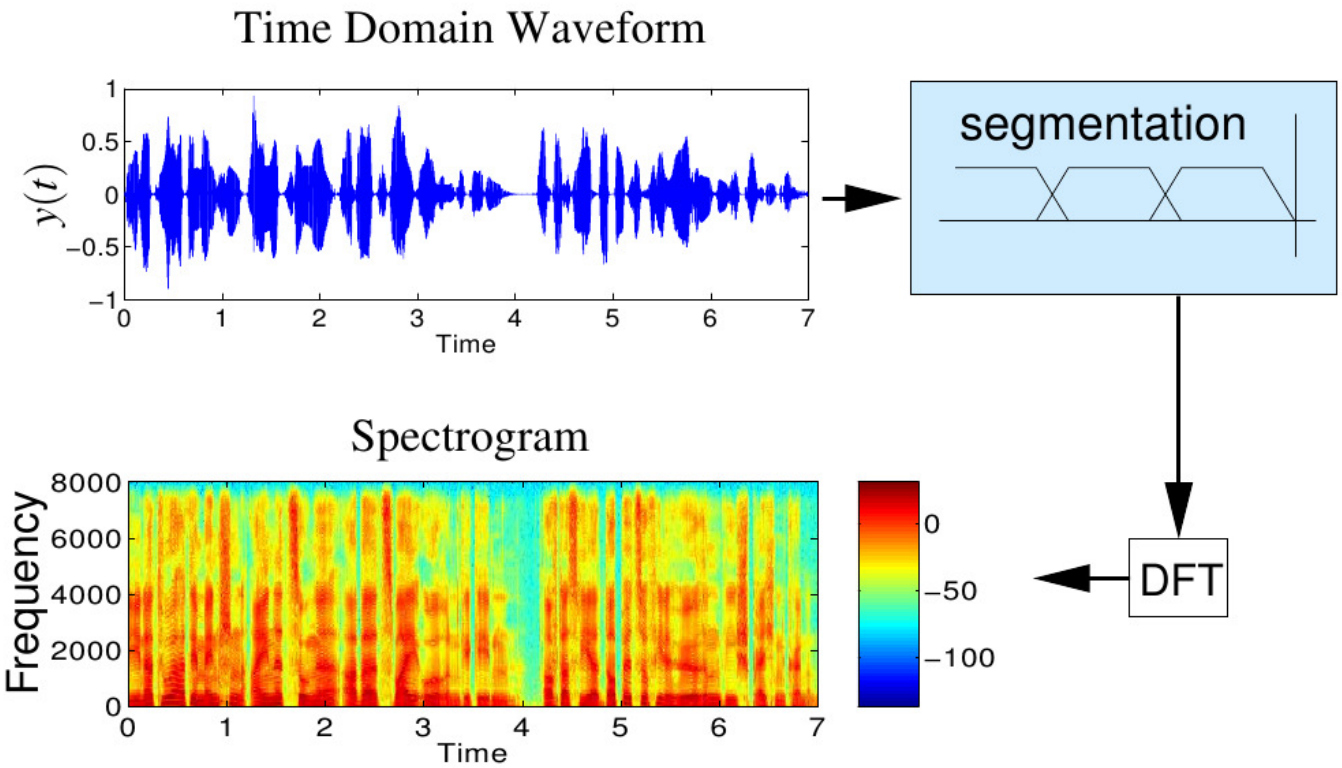


Quelle: Vary, Heute, Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

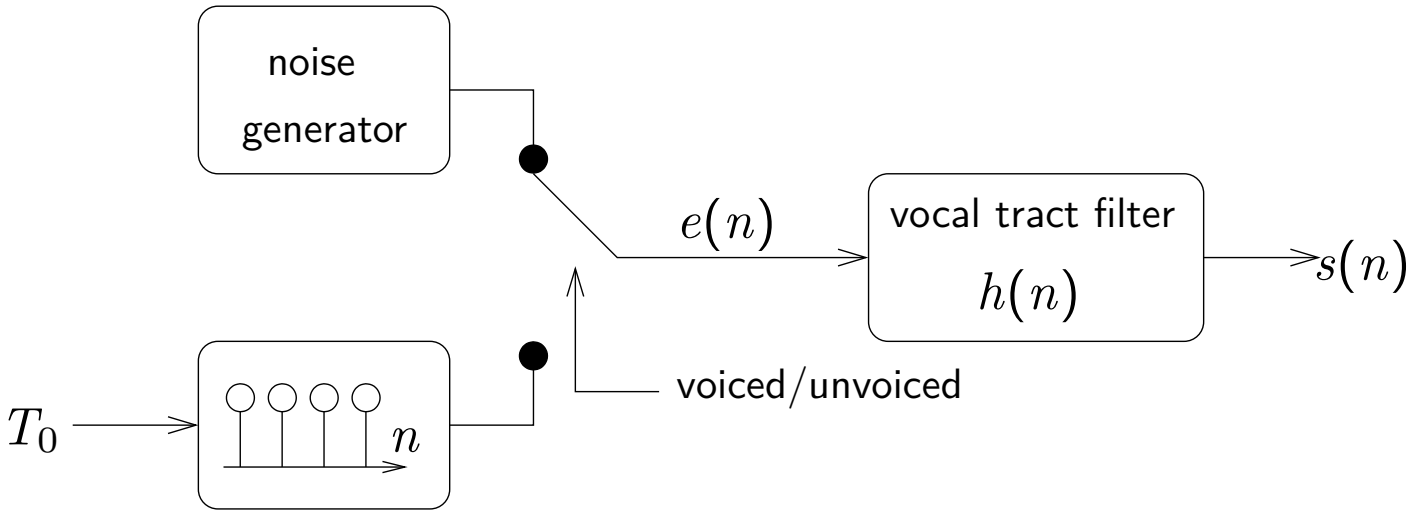
Example

Speech analysis in wavesurfer:

- recorded vowels,
- natural speech



Simple Model of Speech Production



Required Parameters:

- voiced/unvoiced classification
- fundamental period T_0
- vocal tract filter

Homer Dudleys Voder (1939) I

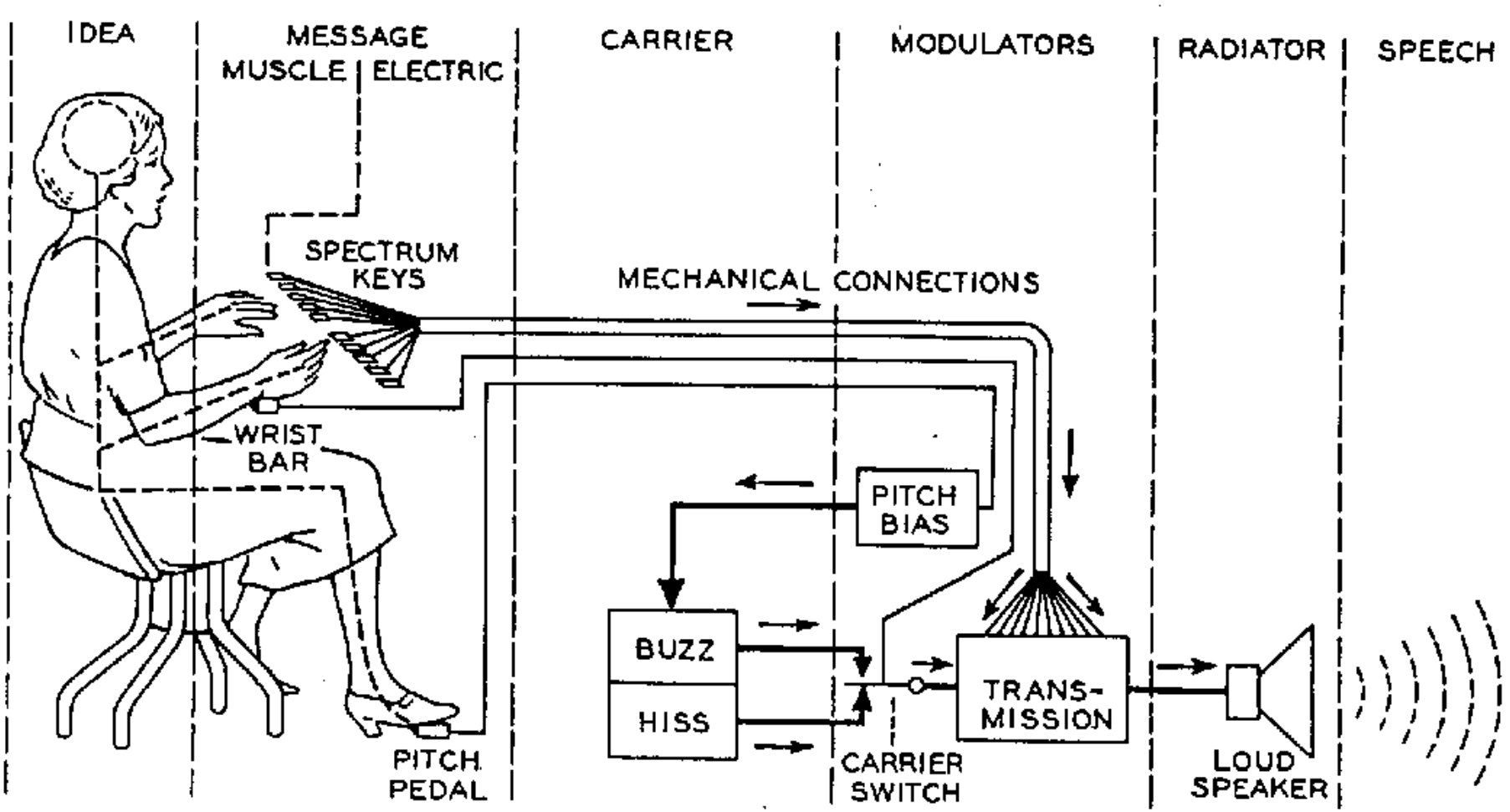


Fig. 8—Schematic circuit of the voder.

Homer Dudleys Voder (1939) II

Voder Video (long)

- for the word “concentration” 13 different sounds in succession
- one year of practice needed
- from 320 trained persons, only 28 people suceeded becoming “expert operators”

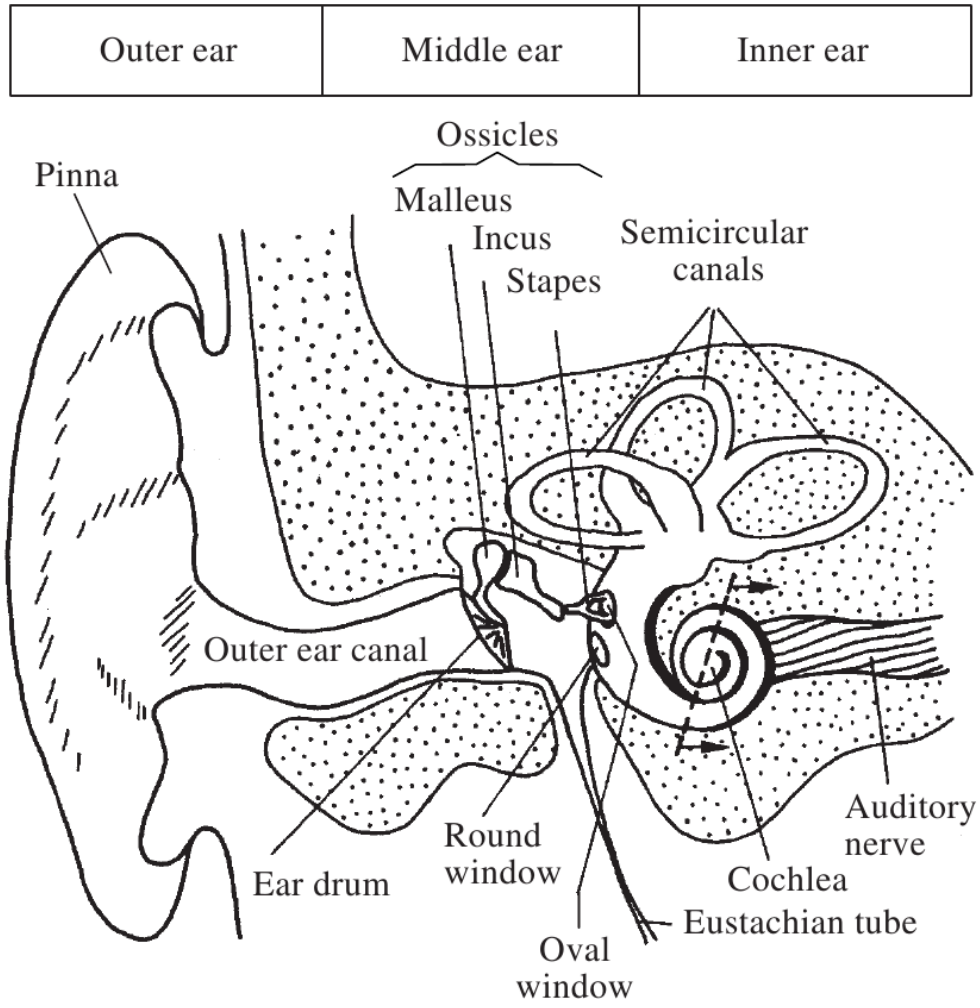
Nice Demo for Speech Production

https://dood.al/pinktrombone/

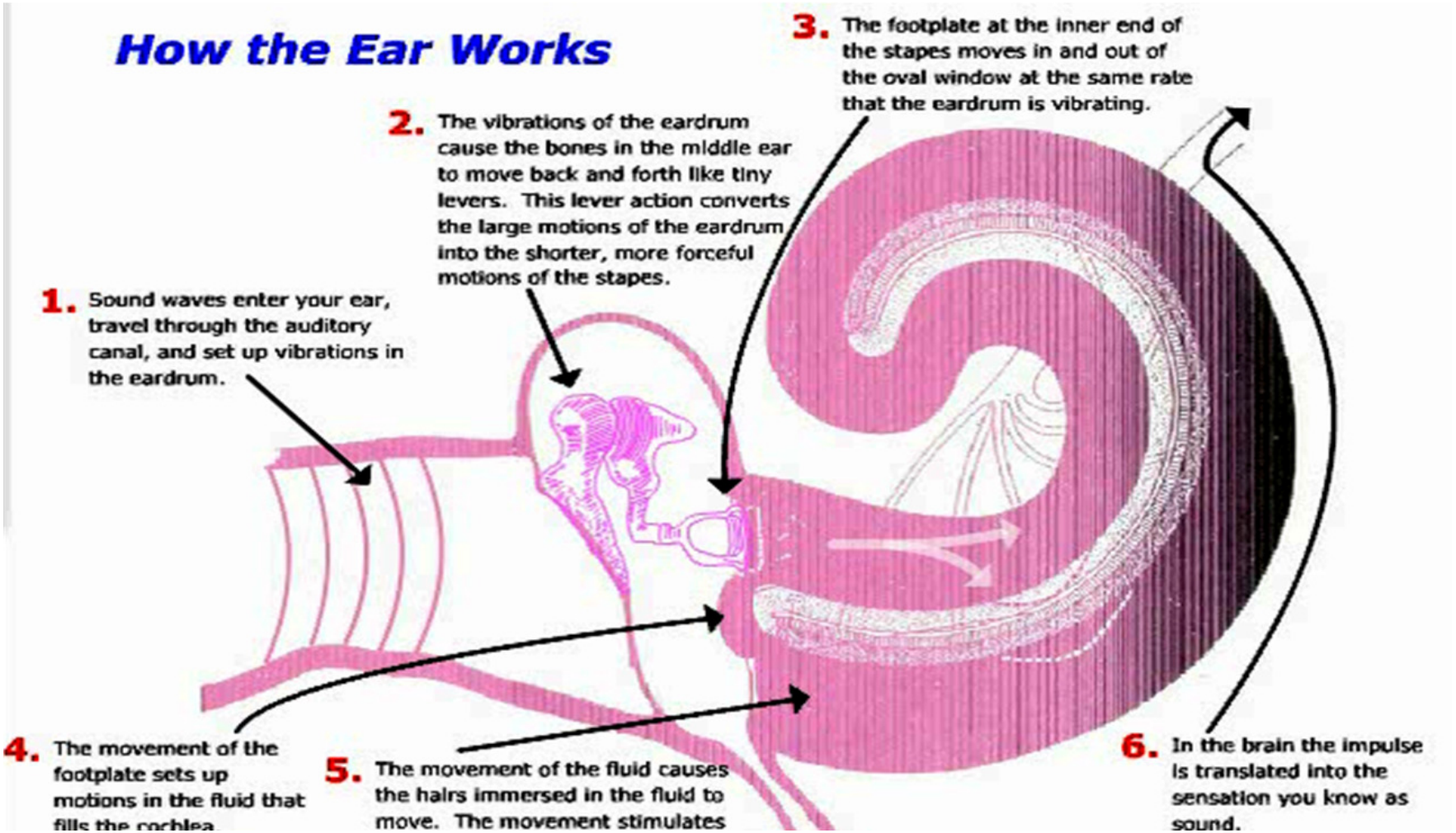
Outline

1. Introduction
   - Speech Production
   - Source-Filter Model
   - Hearing
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Automatic Speech Recognition

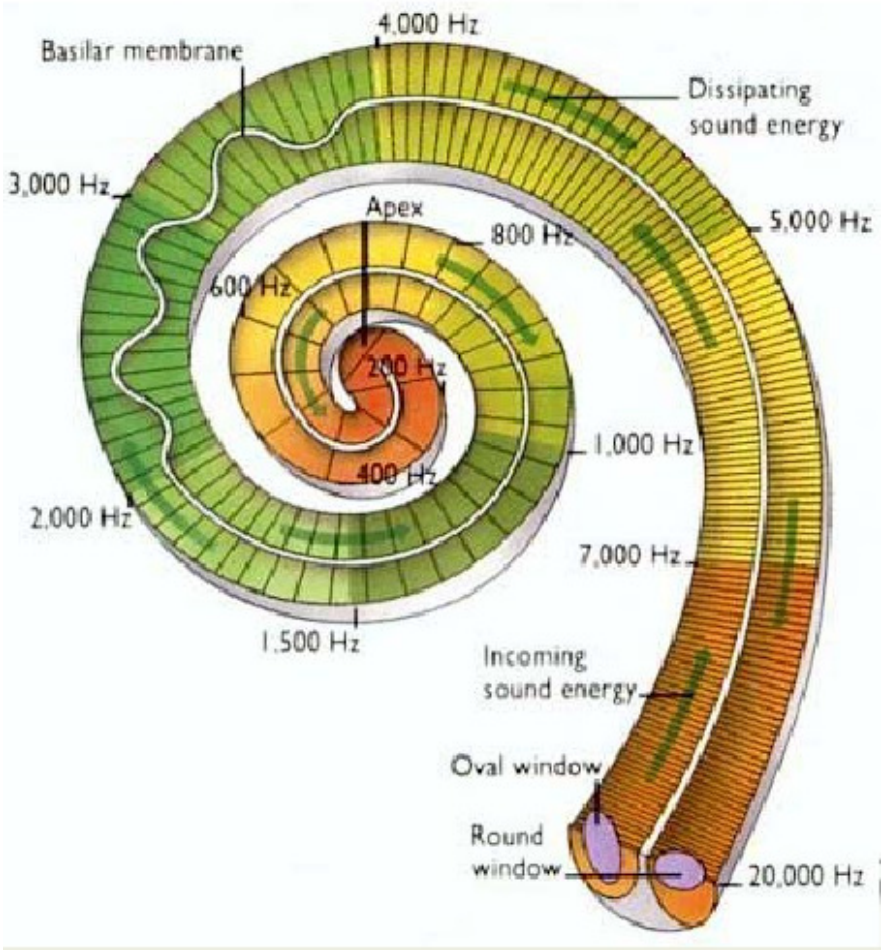
Outer Ear, Middle Ear, Inner Ear



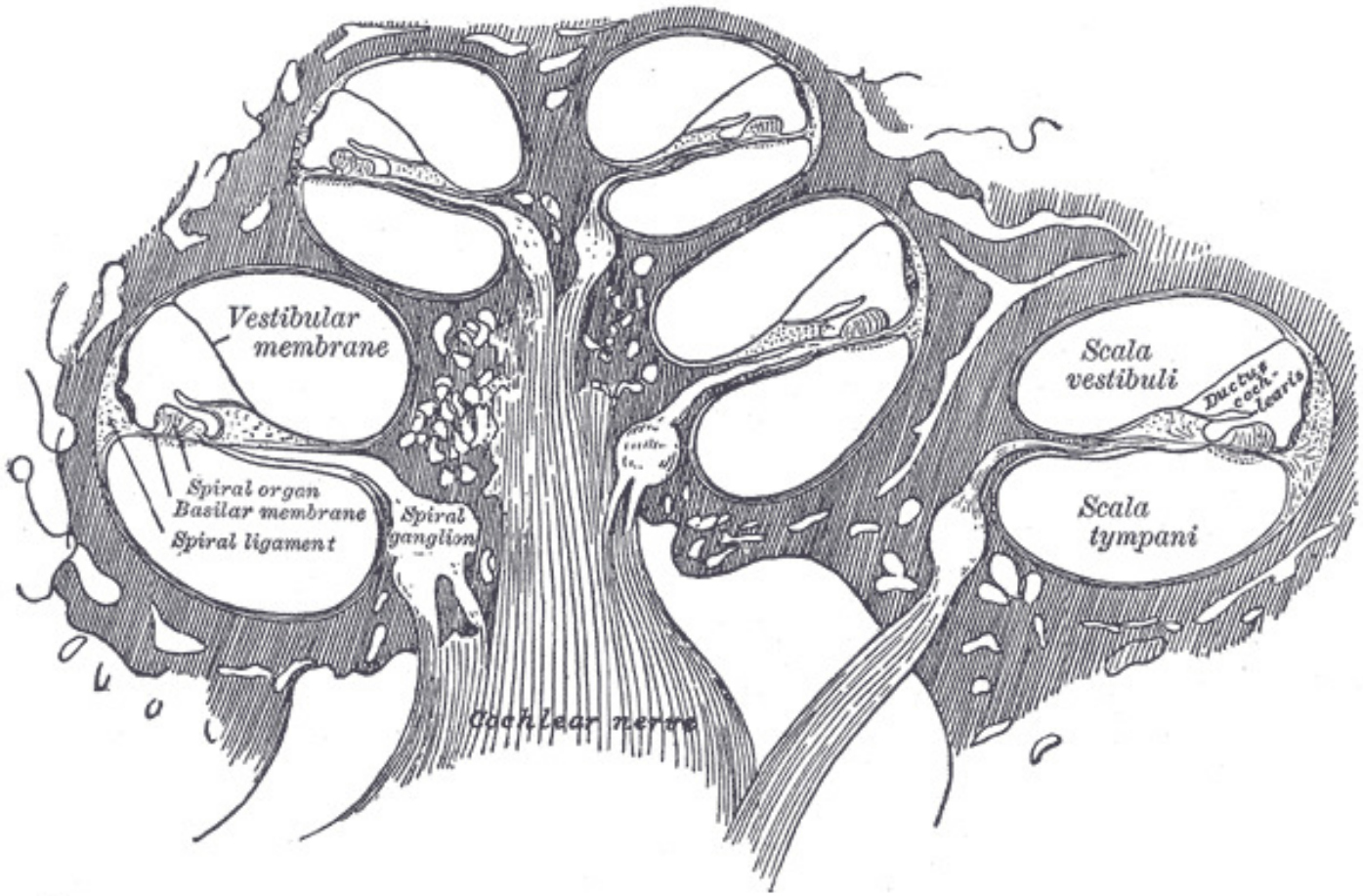
Quelle: Zwicker, Fastl, 1999


How the Ear Works
1. Sound waves enter your ear, travel through the auditory canal, and set up vibrations in the eardrum.
2. The vibrations of the eardrum cause the bones in the middle ear to move back and forth like tiny levers. This lever action converts the large motions of the eardrum into the shorter, more forceful motions of the stapes.
3. The footplate at the inner end of the stapes moves in and out of the oval window at the same rate that the eardrum is vibrating.
4. The movement of the footplate sets up motions in the fluid that fills the cochlea.
5. The movement of the fluid causes the hairs immersed in the fluid to move. The movement stimulates
6. In the brain the impulse is translated into the sensation you know as sound.

Frequency/Place Transformation on the Basilar Membrane

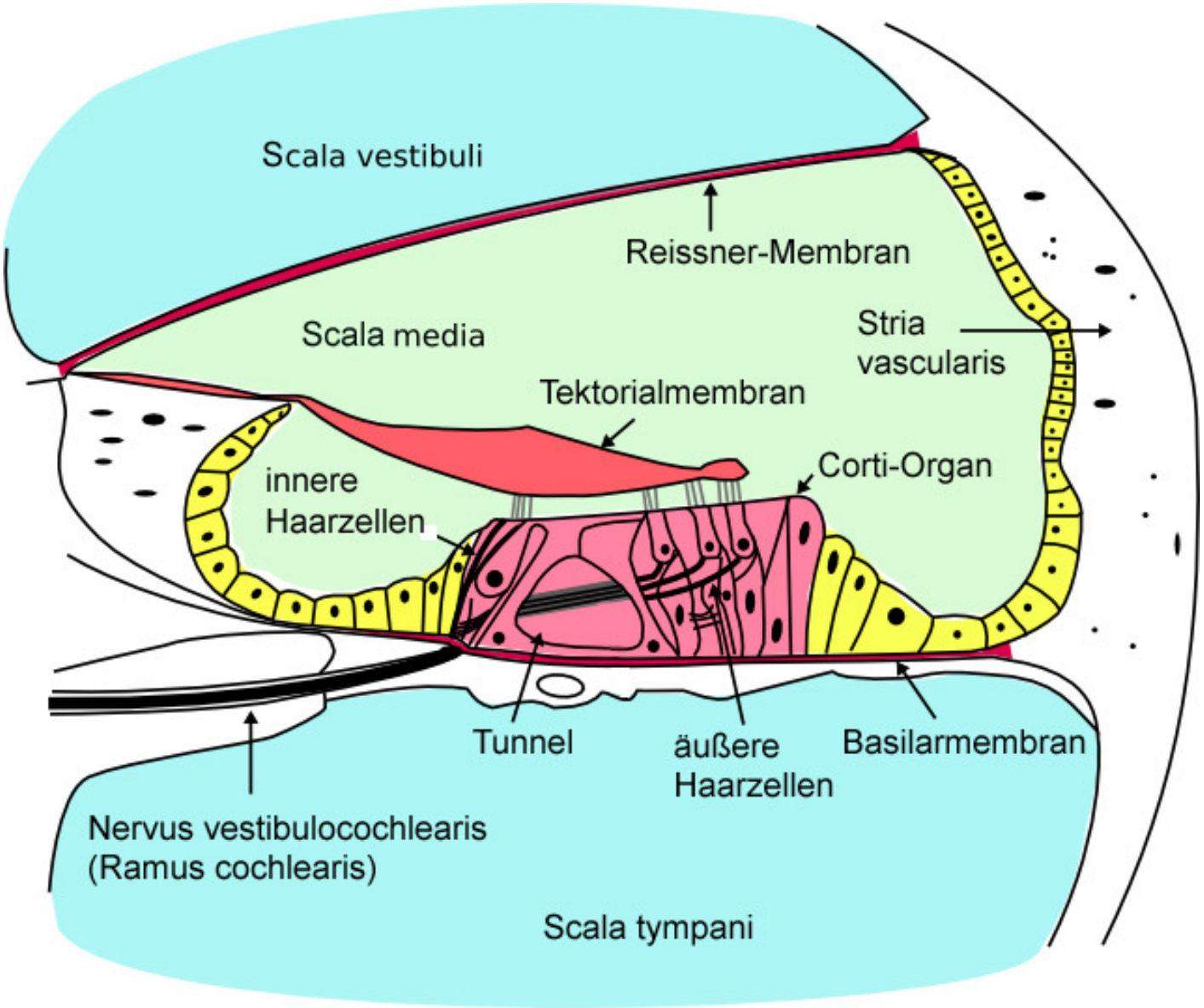


Profile of the Cochlea



Quelle: Henry Gray, Gray's Anatomy, 2008

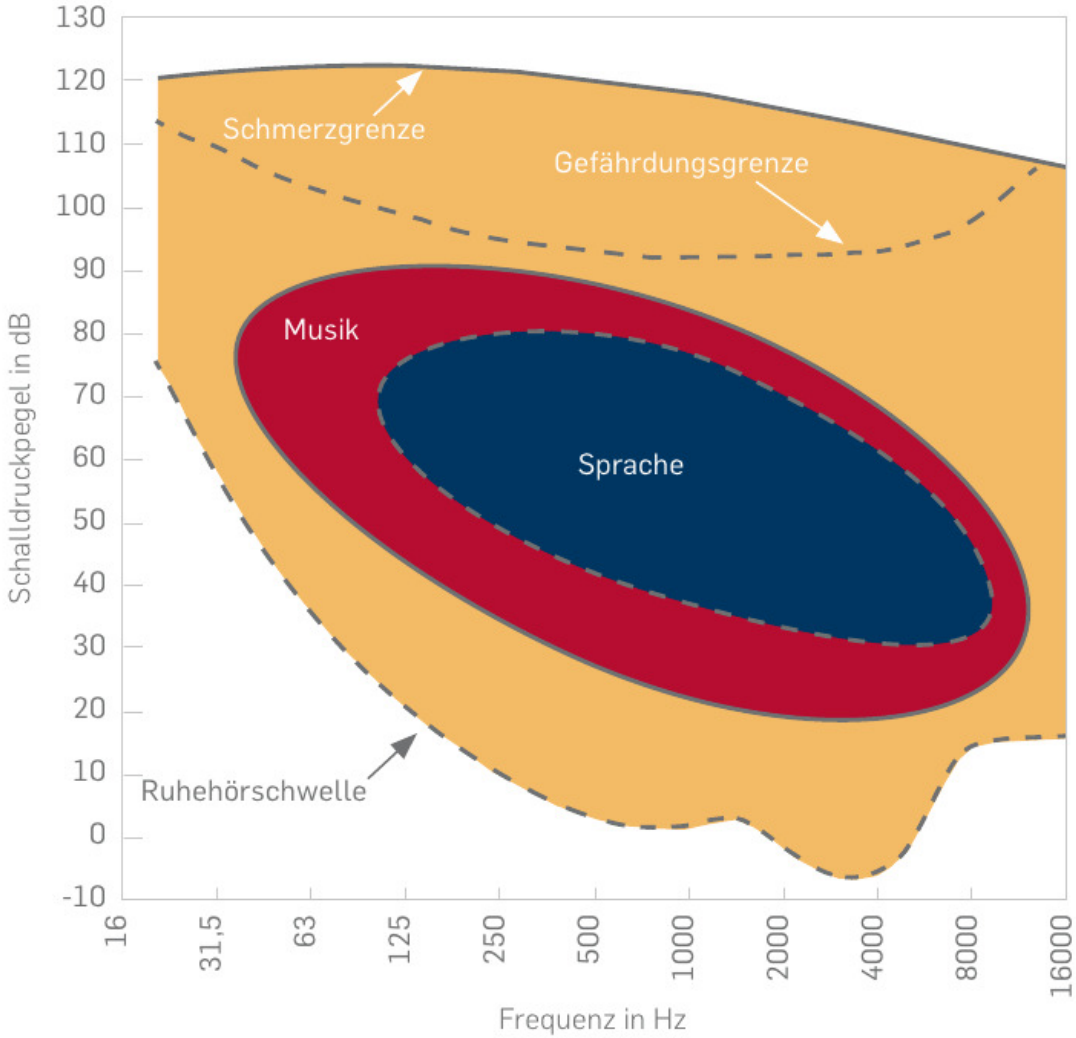
Corti-Organ


Scala vestibuli
Reissner-Membran
Scala media
Stria vascularis
Tektorialmembran
Corti-Organ
innere Haarzellen
Tunnel
äußere Haarzellen
Basilarmembran
Nervus vestibulocochlearis (Ramus cochlearis)
Scala tympani

Short movie

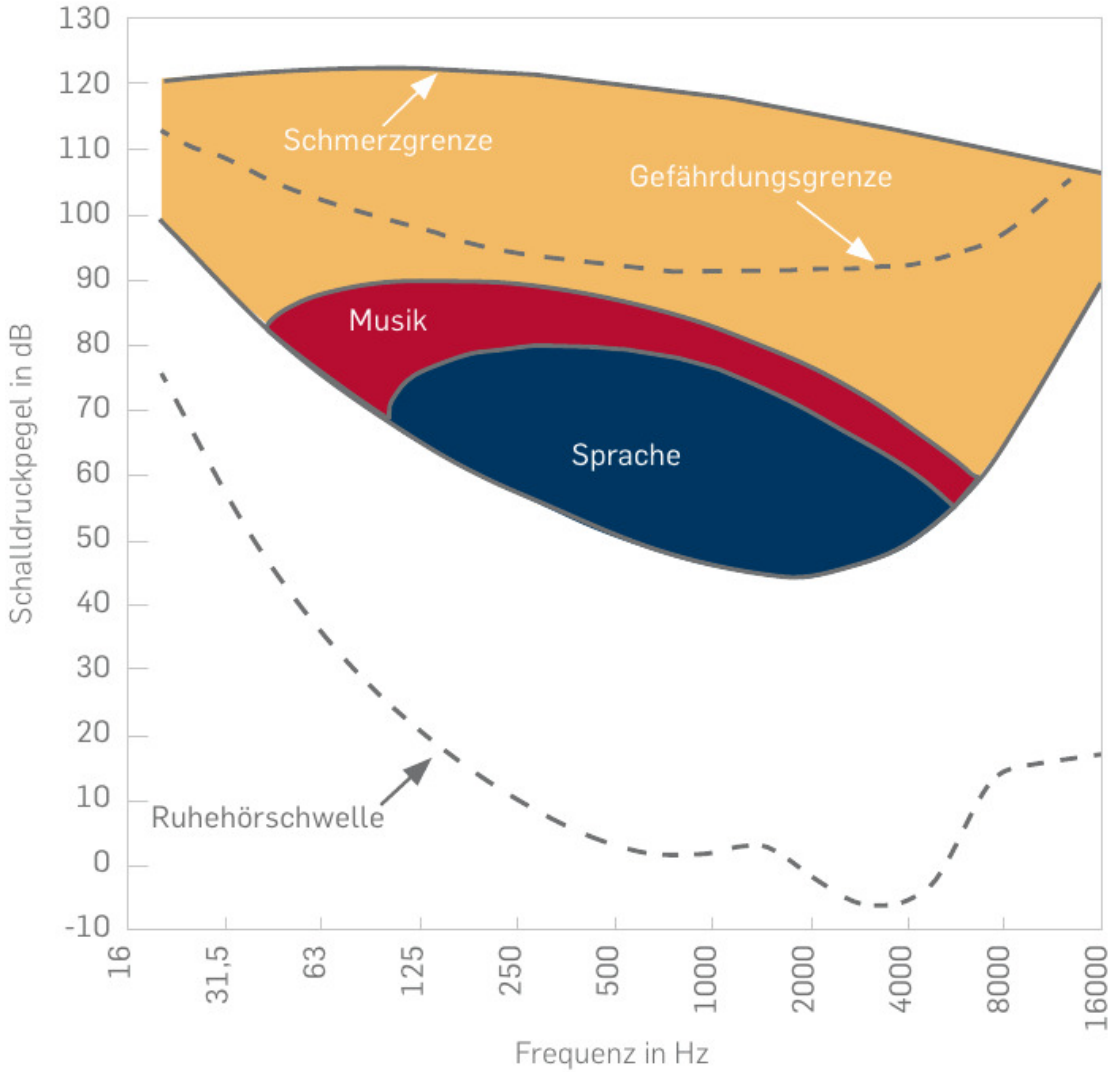
Short Movie

Auditory Sensation Area for the Normal Hearing



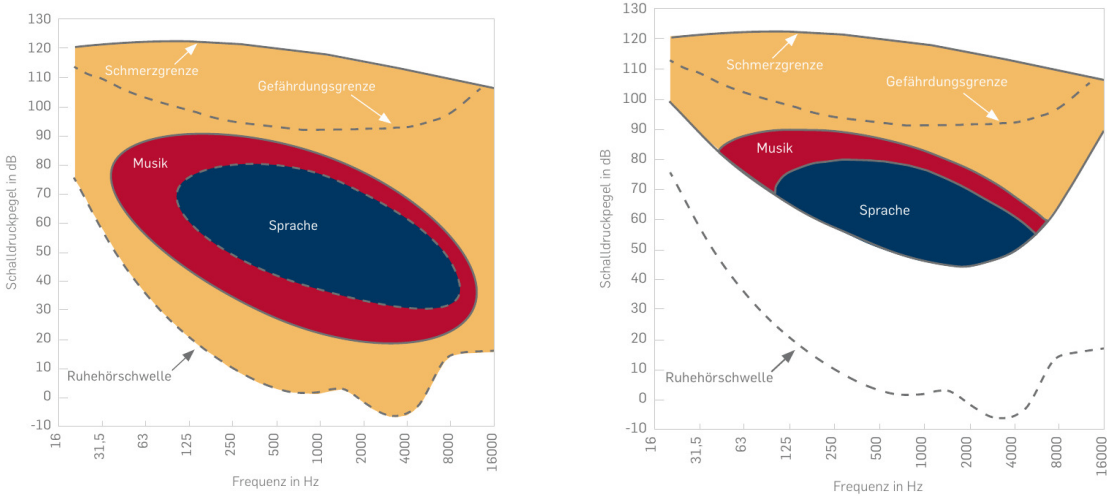
Reduction of the Auditory Sensation Area

for the Hearing Impaired



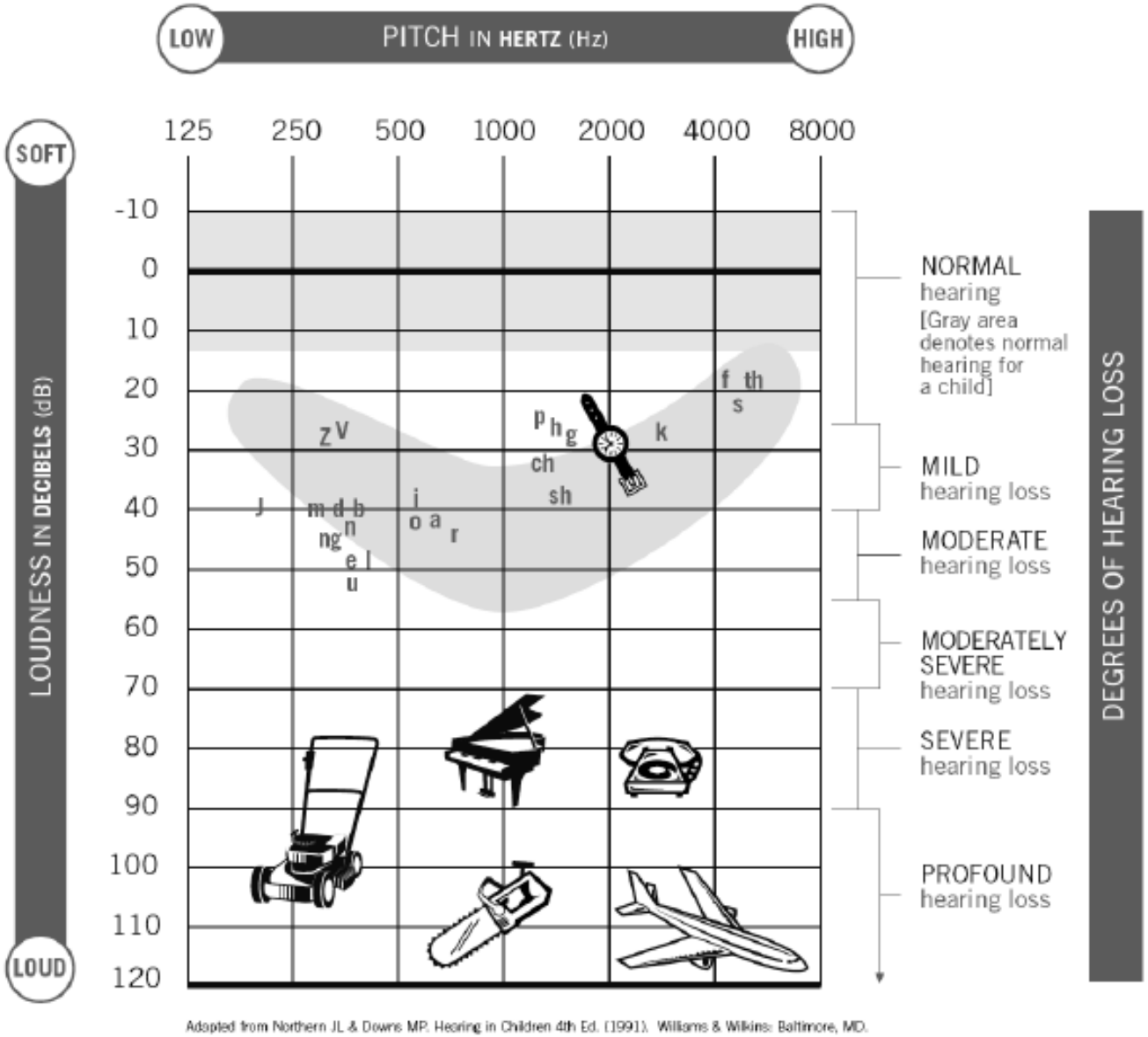
Reduction of the Auditory Sensation Area

for the Hearing Impaired



- For hearing impaired the sensation area is reduced.
- Hearing aids can not simply amplify sound.
- Instead, soft sounds are amplified more than louder sounds.
- This decreases the signal-to-noise-ratio (SNR)

➡ Noise reduction is required!



Adapted from Northern JL & Downs MP. Hearing in Children 4th Ed. (1991). Williams & Wilkins: Baltimore, MD.

Different Kinds of Defective Hearing

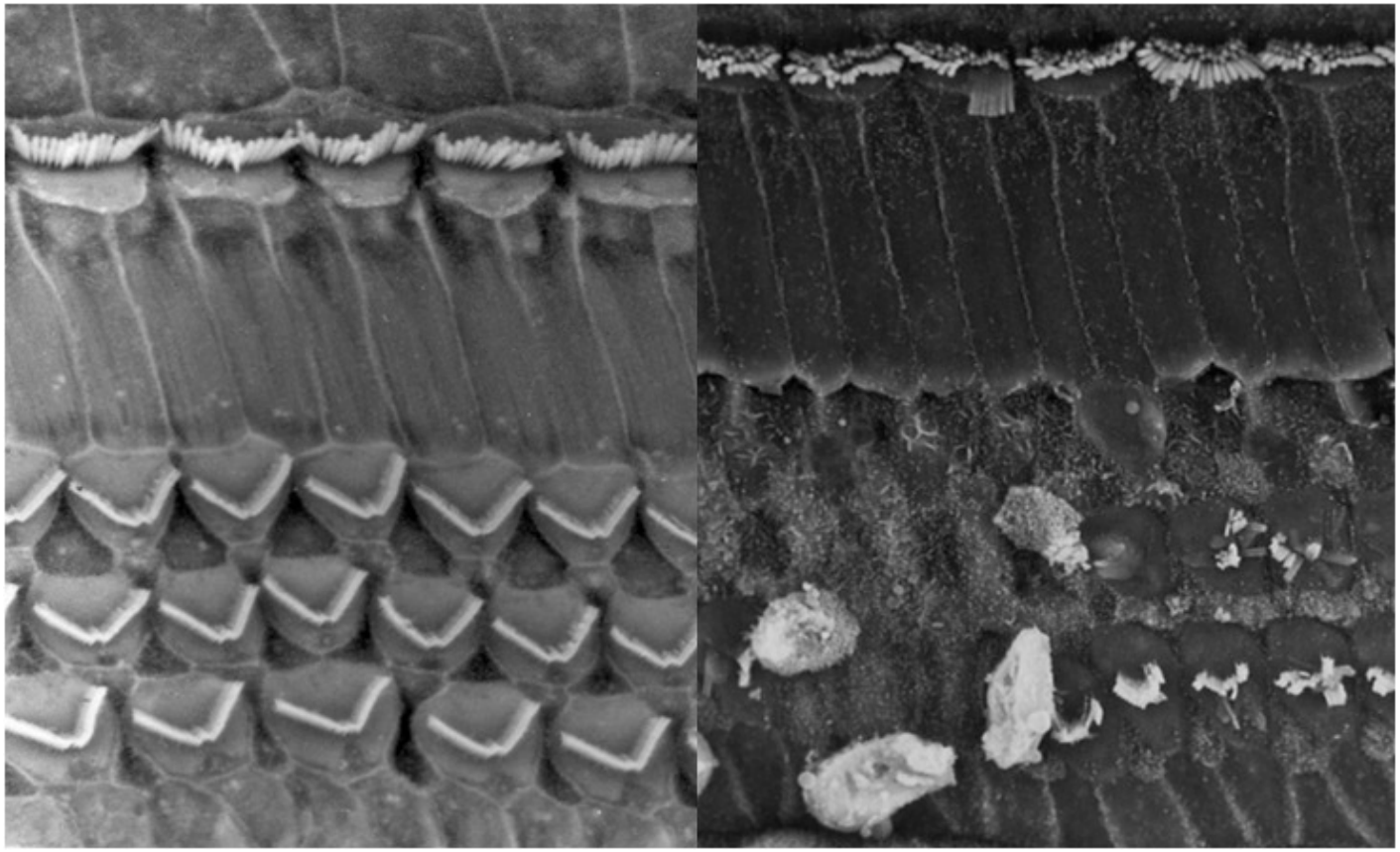
- Conductive hearing loss
  - Sounds is not properly conducted through the outer and/or middle ear
  - Sounds is perceived but attenuated.
  - Can often be healed and/or well treated with hearing aids
- Sensorineural hearing loss
  - defective inner ear, for instance dead or damaged hair cells
  - often co-occurence of ringing in the ears (tinnitus).
  - soft sounds too soft, loud sounds too loud
  - Decreased frequency resolution

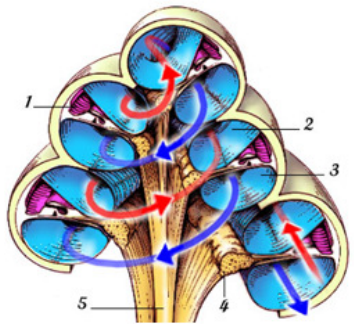
  ➡ decreased speech understanding in noise.

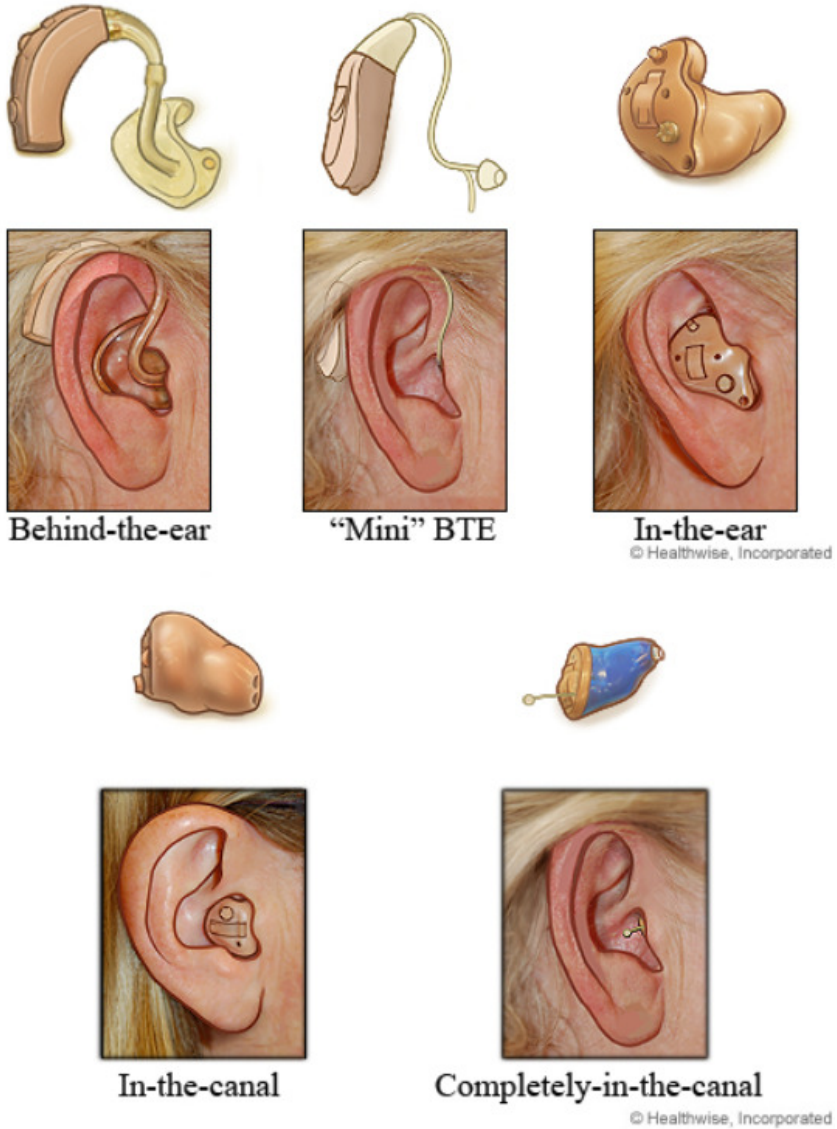
Healthy and Damaged Hair Cells

healthy hair cells (from above)

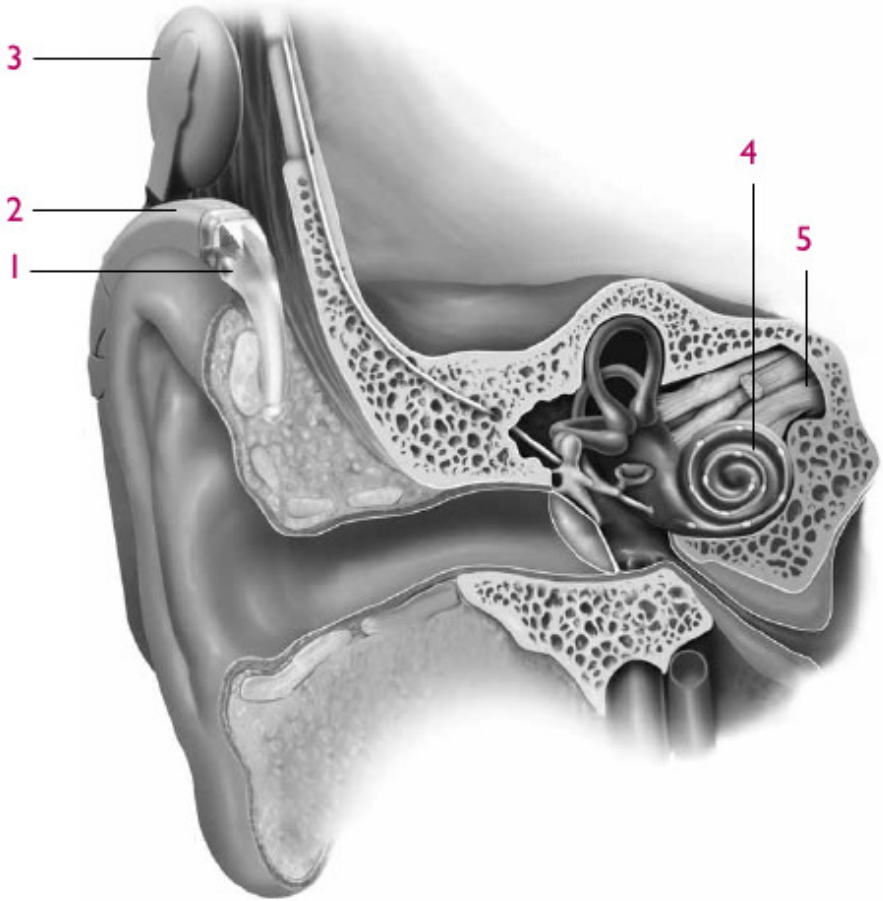
hair cells for sensorineural deafness






Behind-the-ear
"Mini" BTE
In-the-ear
© Healthwise, Incorporated
In-the-canal
Completely-in-the-canal
© Healthwise, Incorporated

Cochlear Implants



1 Sounds are picked up by the microphone.

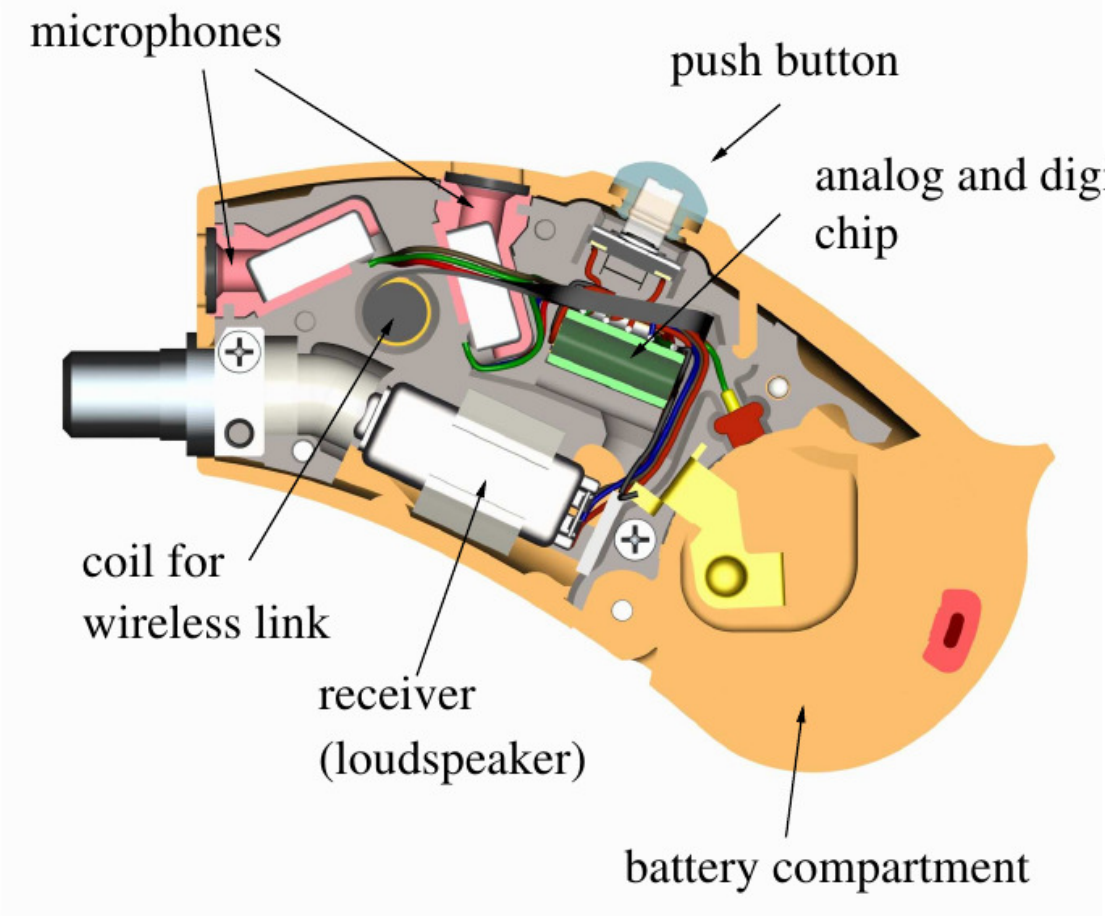
2 The signal is then "coded" (turned into a special pattern of electrical pulses).

3 These pulses are sent to the coil and are then transmitted across the skin to the implant.

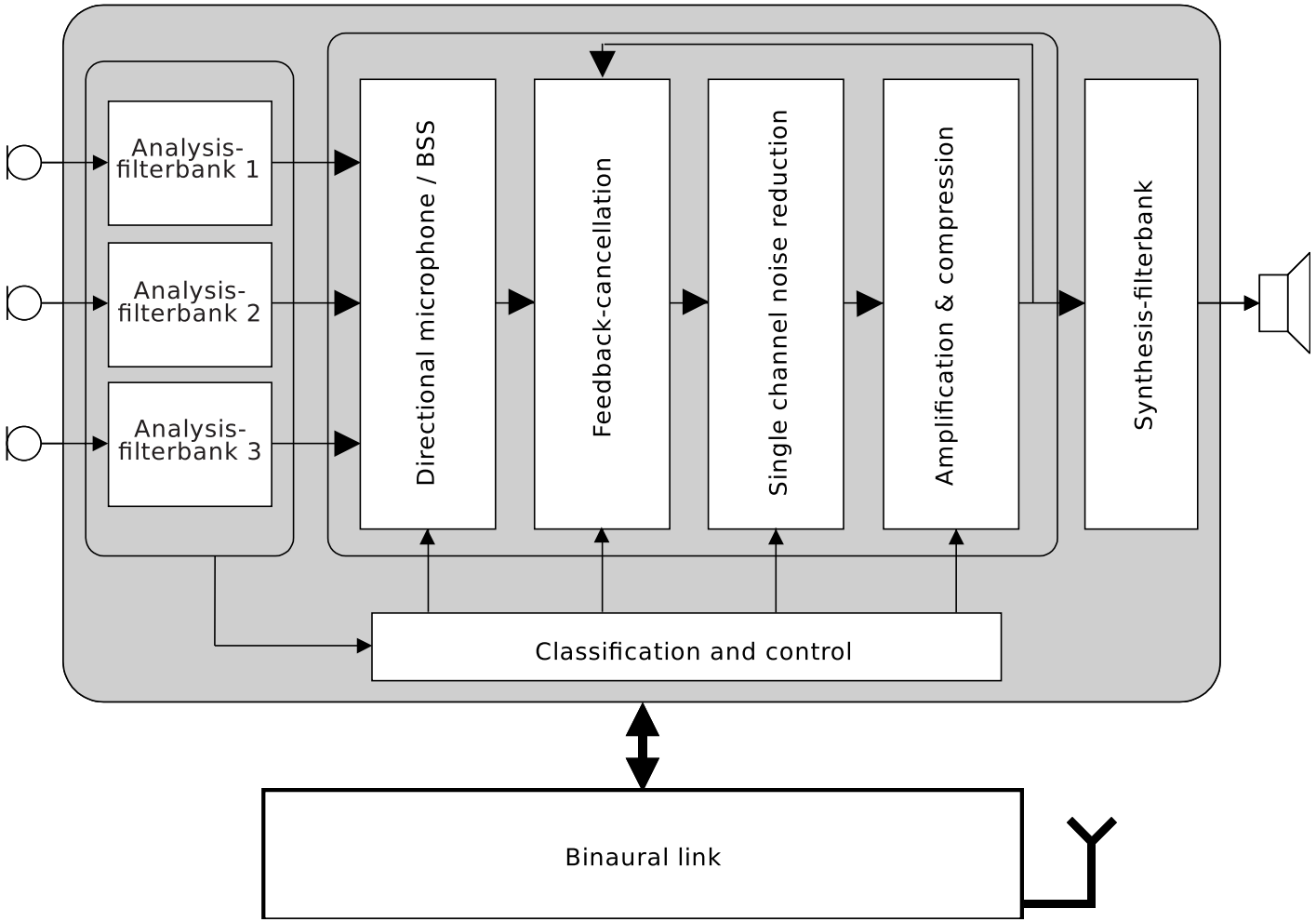
4 The implant sends a pattern of electrical pulses to the electrodes in the cochlea.

5 The auditory nerve picks up these electrical pulses and sends them to the brain. The brain recognizes these signals as sound.

Quelle: Handbook for Educators, Med-El



Quelle: Siemens Audiologische Technik



Hamacher et al. in: Martin et al. (eds.), Wiley 2008

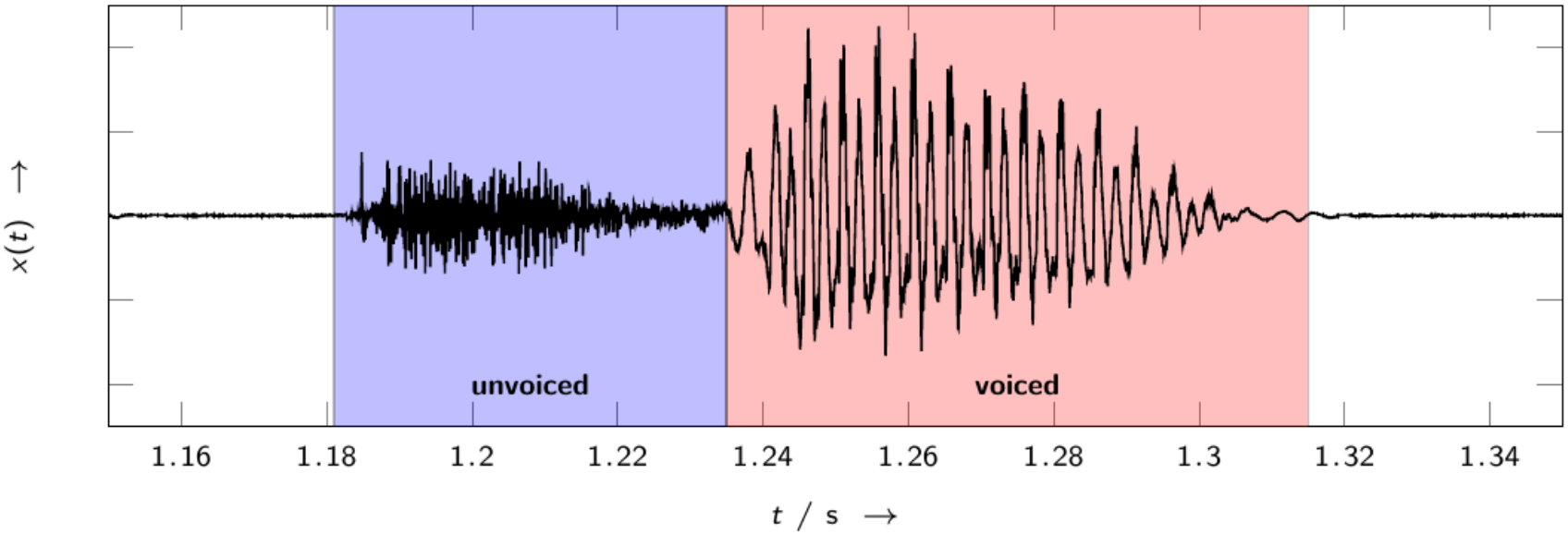
Outline

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP Signal Processing

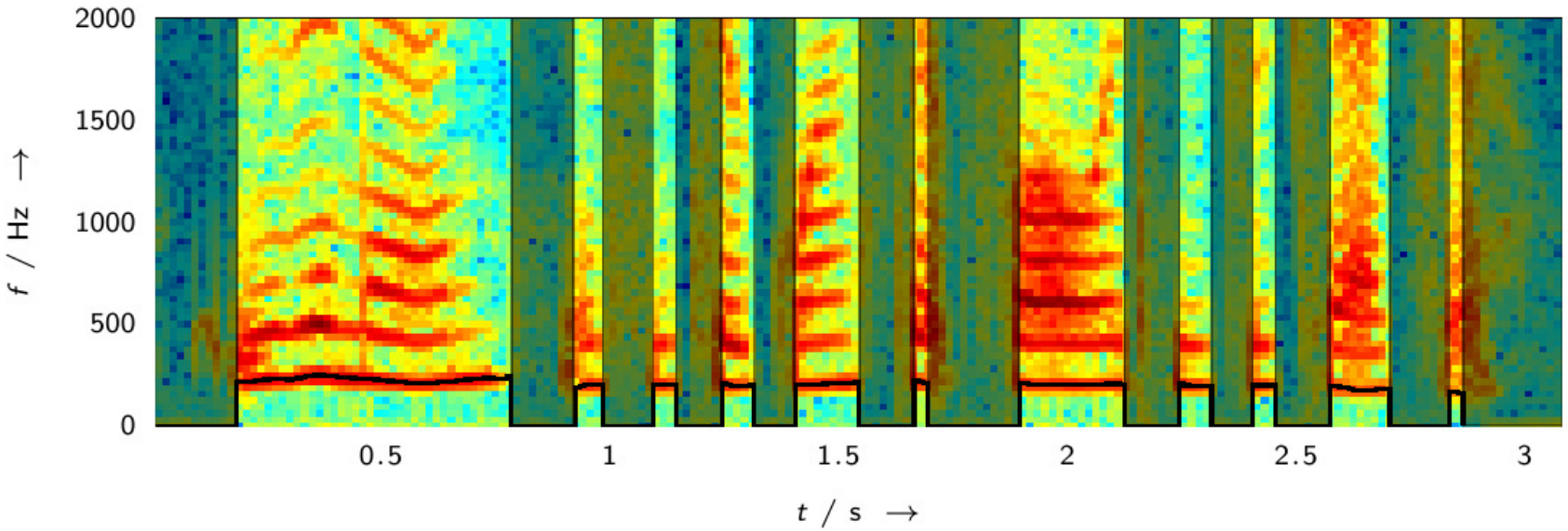
2. Fundamental Frequency Estimation

Periodicity in speech signals



- **Unvoiced speech:**
  - noisy excitation
  - pitch not available
- **Voiced speech:**
  - periodic glottis excitation
  - pitch available

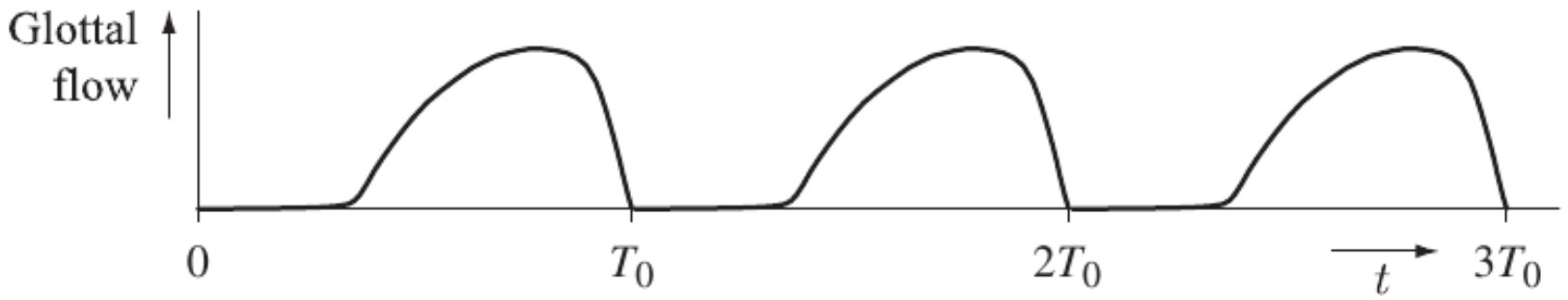
Periodicity in speech signals



Speech Fundamental Frequency Estimation

- The speech fundamental frequency f_0 is an important parameter in speech signal processing, e.g. for
  - speech coding
  - speech enhancement
  - speech modeling
  - speaker recognition
- Often *pitch* is synonymiously used. Althought, strictly speaking, pitch is a perceptual quantity.
- The perceived pitch is influenced by the loudness and length of a tone.
- Here, we refer to the physical quantity given by the inverse of the fundamental period.
- Range of the fundamental frequency: 40 Hz – 600 Hz (600 Hz for children)
- male speakers: around 100Hz; female speakers: around 200Hz

Air Flow for Voiced Excitation



Vocal cords produce a pulsating air flow through the vocal cords.

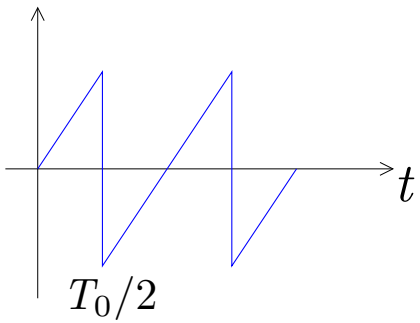
- Opening of glottis due to the increased pressure
- Air flows through the glottis, vocal cords are under tension
- Because of the constriction of the glottis, the flow velocity increases while the pressure decreases (Bernoulli-effect)
- The vocal cords snap together, the air flow is interrupted
- the pressure increases, the glottis opens up

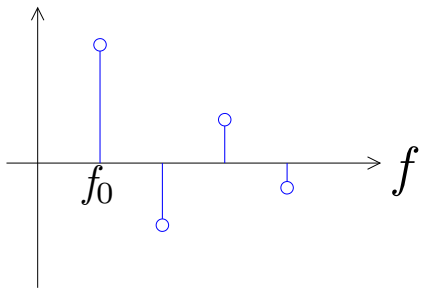
Fourier series: Every periodic function $g(t)$ with period T_0 can be represented by a series of sine and cosine functions, whose frequencies are integer multiples of the fundamental frequency $f_0 = 1/T_0$:

$$g(t) = \frac{a_0}{2} + \sum_{h=1}^{\infty} \left(a_h \cos(2\pi h f_0 t) + b_h \sin(2\pi h f_0 t) \right)$$

➡ The glottis signal consists of the fundamental oscillation and its harmonics.

Example:





$$g(t) = \frac{1}{\pi f_0} \sum_{h=1}^{\infty} \frac{(-1)^{h-1}}{h} \times \sin(2\pi h f_0 t) \quad (2)$$

Residual Effect

- Telephone speech is bandpass filtered between 300 Hz and 3400 Hz ("telephone voice")
- The lowest harmonic (fundamental frequency) is often not present in the signal. Still we can distinguish between male and female speakers.
- Example
  - 200 Hz Ton
  - 300 Hz Ton
  - ?

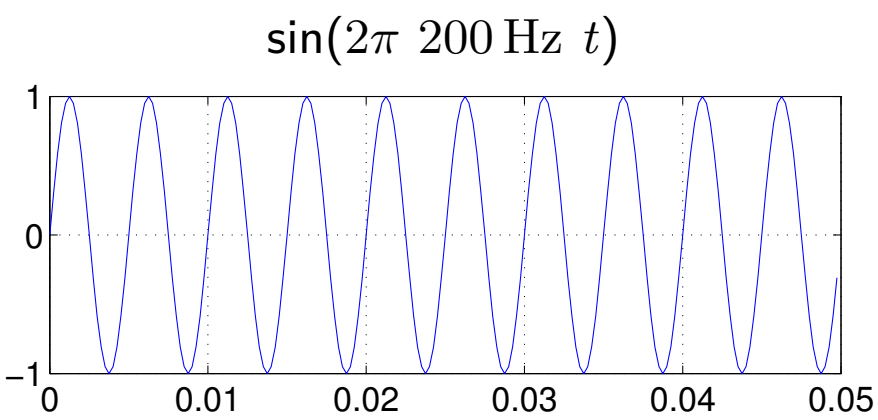
Residual Effect

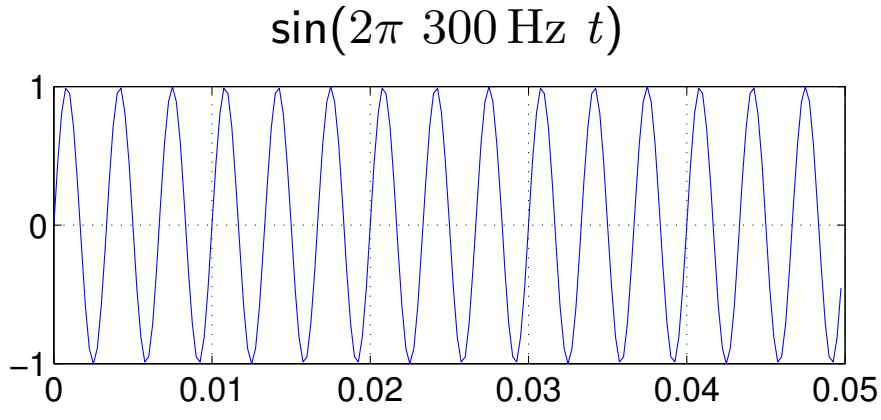
- Telephone speech is bandpass filtered between 300 Hz and 3400 Hz ("telephone voice")
- The lowest harmonic (fundamental frequency) is often not present in the signal. Still we can distinguish between male and female speakers.
- Example
  - 200 Hz Ton
  - 300 Hz Ton
  - ?

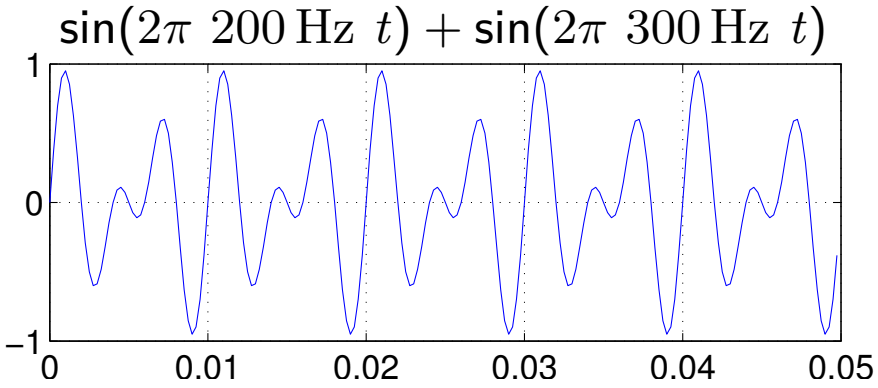
➡ Adding a 200 Hz sinusoid and a 300 Hz sinusoid, the resulting tone-complex has a fundamental period of 1/100Hz.

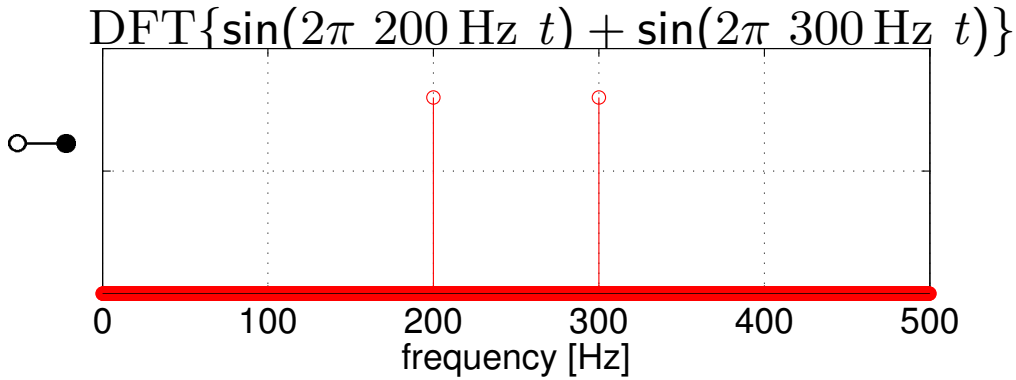
➡ Distance between harmonics equals perceived fundamental frequency

Residual Effect I





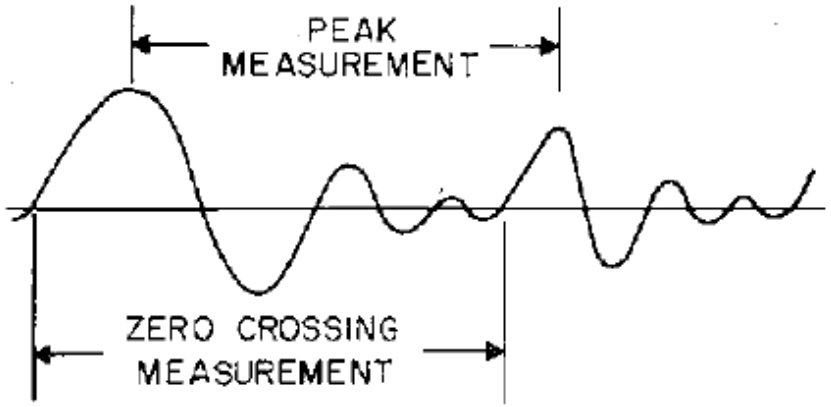




Example: Telephone Voice

- While typical average speech fundamental frequencies are betwen 50-300 Hz, frequencies below 300 Hz are not transmitted over the telephone channel.
- Telephone-speech
- Wideband-speech
- Even though the speech fundamental frequency is not transmitted, we can still determine the pitch.
- See background read: J. Hecht (2014): "Why Mobile Voice Quality Still Stinks—and How to Fix it," IEEE Spectrum.

- Simple solution: Distance between peaks (or zero-crossing before the peaks) in the time-domain



Quelle: Rabiner et al., IEEE TASSP, Oct. 1976

✔ Simple way for a fast assessment of the fundamental frequency (e.g. using *Wavesurfer*)

✖ Not applicable for an automatic pitch estimation algorithm: large error rate for natural speech and in noise.

➡ Better: autocorrelation-base

Autocorrelation and power spectral density (PSD)

Let $x(n)$ denote a realization of a random process

- Autocorrelation function

$$\varphi_{\mathrm{xx}}(\lambda) = \mathrm{E}(x(n)x^*(n+\lambda)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u\, v\, p_{x(n)x^*(n+\lambda)}(u,v)\mathrm{d}u\mathrm{d}v$$

- The signal is shifted against itself $\rightarrow$ measure of self-similarity
- Estimation for a quasi-stationary segment of length N for lag $\lambda > 0$

$$\widehat{\varphi}_{\mathrm{xx}}(\lambda) = \frac{1}{N-|\lambda|}\sum_{n=0}^{N-|\lambda|-1} x(n)x^*(n+\lambda).$$

- The Fourier transform of the autocorrelation function is called *power spectral density* (PSD)

$$\Phi_X(f) = \sum_{\lambda=-\infty}^{\infty} \varphi_{\mathrm{xx}}(\lambda)\mathrm{e}^{-\mathrm{j}\Omega\lambda}$$

White Noise

- power spectral density constant over frequency ("white")

$$\Phi_X(f) = \sigma_X^2$$

- Autocorrelation function:

$$\varphi_{\mathrm{XX}}(\lambda) \propto \sigma_X^2 \delta(\lambda) = \begin{cases} \sigma_X^2 & \lambda = 0 \\ 0 & \lambda \neq 0 \end{cases}$$

➡ samples are mutually uncorrelated

White Noise

- power spectral density constant over frequency ("white")

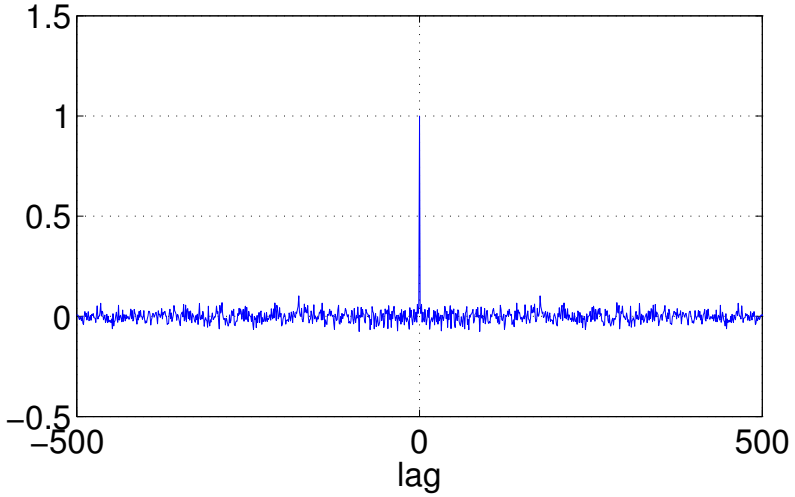
$$\Phi_X(f) = \sigma_X^2$$

- Autocorrelation function:

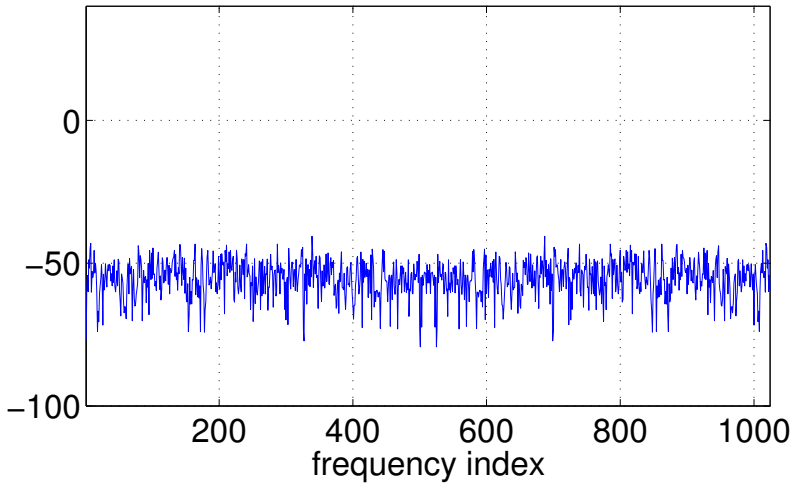
$$\varphi_{\mathrm{XX}}(\lambda) \propto \sigma_X^2 \delta(\lambda) = \begin{cases} \sigma_X^2 & \lambda = 0 \\ 0 & \lambda \neq 0 \end{cases}$$

➡ samples are mutually uncorrelated

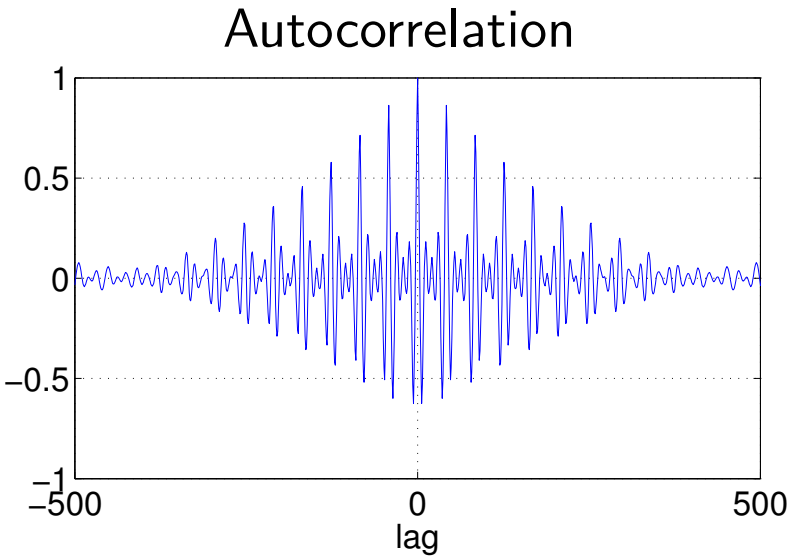
Autocorrelations function

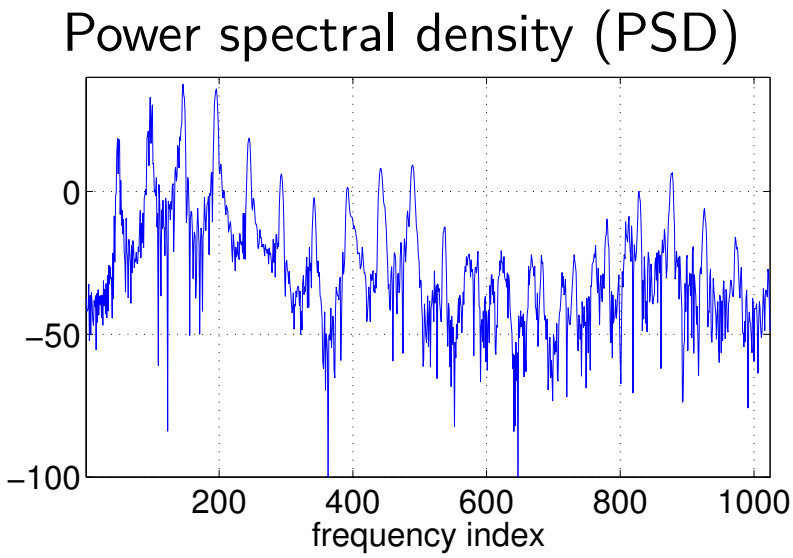


Power Spectral Density (PSD)



Autocorrelation for Voiced Speech





- $f_\mathrm{s} = 8\,\mathrm{kHz}$, Segment length: $128\,\mathrm{ms}$, DFT length: 2048 $(256\,\mathrm{ms})$.

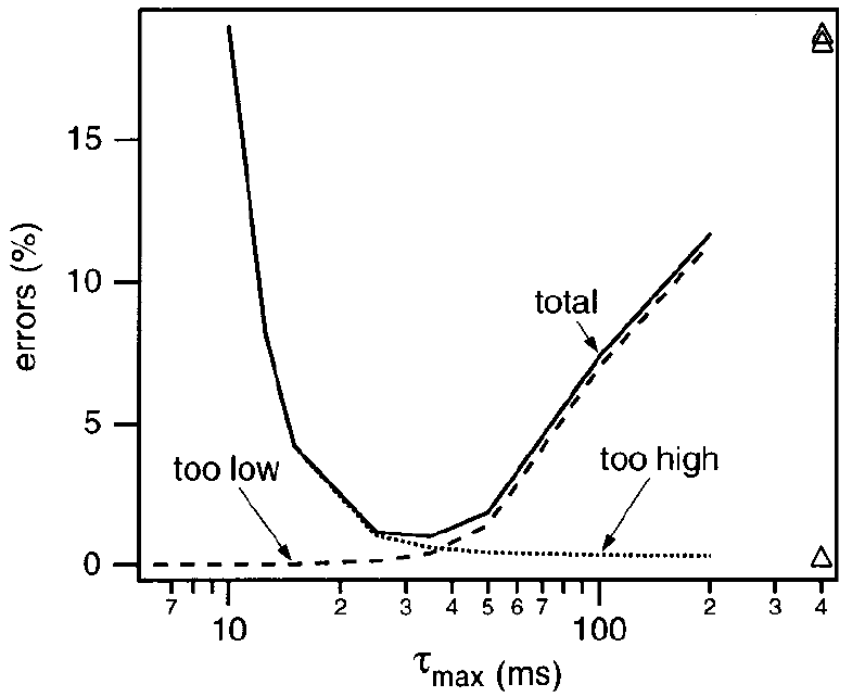
➡ Successive samples are correlated, i.e. statistically dependent

➡ colored, non-constant spectrum

➡ The peak next to the lag $\lambda = 0$ of the autocorrelation function corresponds to the fundamental period T_0.

➡ First peak in the fine structure of the spectrum corresponds to the speech fundamental period $f_0 = 1/T_0$.

- The window length must be carefully chosen
  - The more periods fit into a window, the more robust the estimation (the larger the window, the better)
  - The speech fundamental frequency changes over time (the shorter the window, the better)
- $\approx 30\,\mathrm{ms}$ is a good compromise (3 periods at $f_0 = 100\,\mathrm{Hz}$).



Relation between difference approach and ACF I

- Difference approach: for a signal that is periodic in T_0, we have

$$x(t) - x(t + T_0) = 0, \quad \forall t$$

- the same holds for the square of the difference

$$d_{T_0}(t) = \frac{1}{N - |\lambda|} \sum_{n=0}^{N-|\lambda|-1} (x(t) - x(t + T_0))^2$$

- Approach: find the T_0 that minimizes $d_{T_0}(t)$

➡ This is the same as computing

$$d_{T_0}(t) = \widehat{\varphi}_{\mathrm{x(t)}}(0) + \widehat{\varphi}_{\mathrm{x(t+T_0)}}(0) - 2\widehat{\varphi}_{\mathrm{x(t)}}(T_0)$$

- If $\widehat{\varphi}_{\mathrm{x(t+T_0)}}(0) = \widehat{\varphi}_{\mathrm{x(t)}}(0)$ the ACF and the difference approach are equivalent

Relation between difference approach and ACF II

- However:

Approach	Error (%)
ACF	10,00
Difference	1.95
YIN	0.50

➡ The ACF approach is sensitive for changing signal powers

- The difference approach is the basis of the well known YIN algorithm[1].

[1] A. d. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

Outline

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

3. Spectral Analysis of Audio Signals

Outline

4. Vocal Tract Model and Linear Prediction

Outline

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

5. Sampling, Quantization, and Speech Coding

Outline

6. Speech Enhancement

Outline

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SP
Signal Processing

7. Automatic Speech Recognition

References


A. d. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.