



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Prof. Timo Gerkmann

Speech Signal Processing

- 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
- K. Jung, R.M. Mersereau, "Medial and Radio Signal Processing for Mobile Communications", Cambridge University Press, 2018
- J. Benesty, I. Cohen, J. Chen, "Fundamentals of Signal Enhancement and Array Signal Processing", Wiley, 2018

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 Separation: How can we separate simultaneously talking speakers?

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. Machine Learning Basics

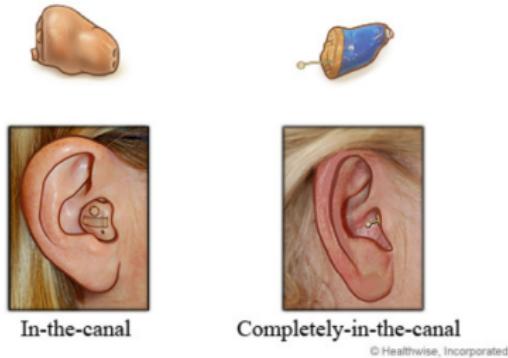
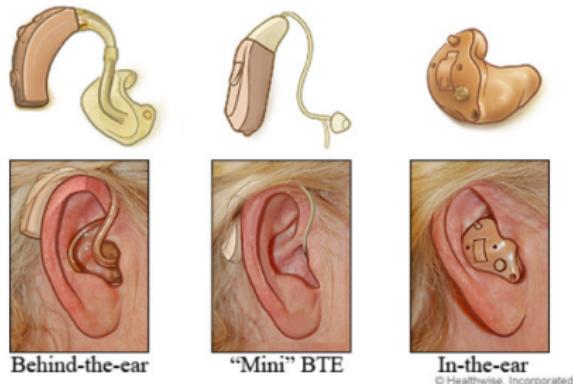
8. Source Separation

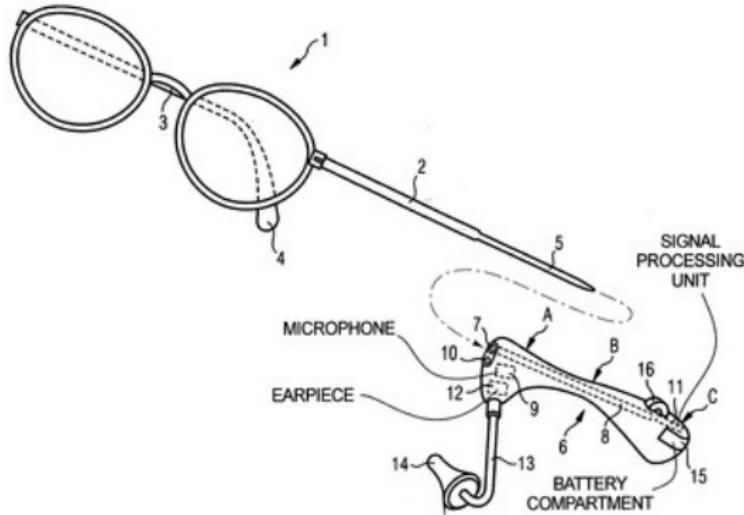


Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Introduction

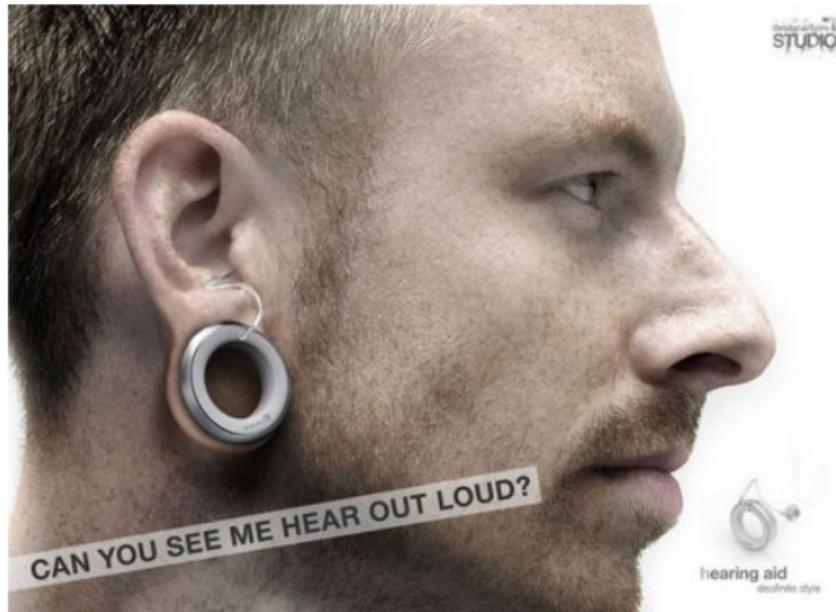




United States Patent 7103192

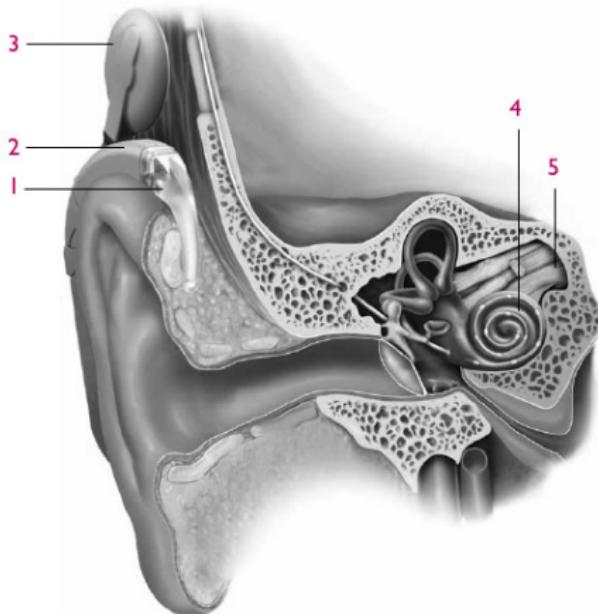


varibel.nl



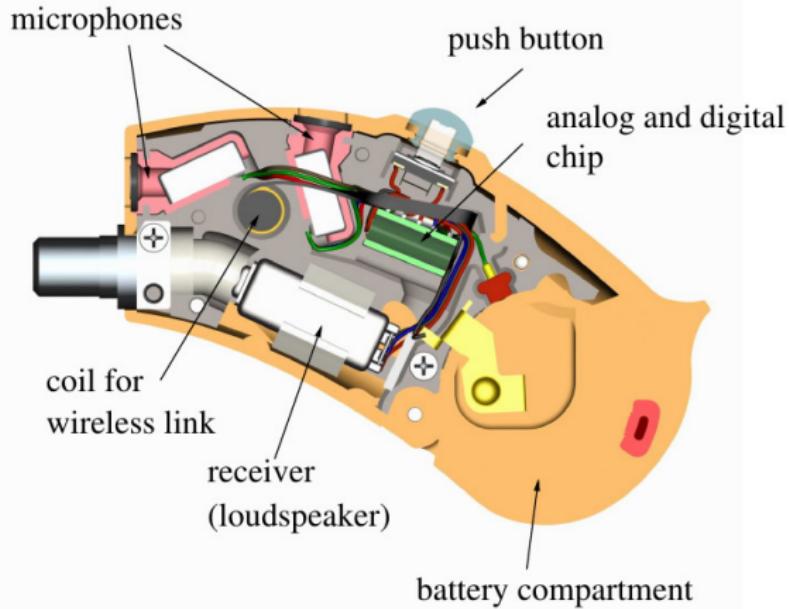
Quelle: designaffairs.com

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

- 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)



- Robust speech recognition required

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



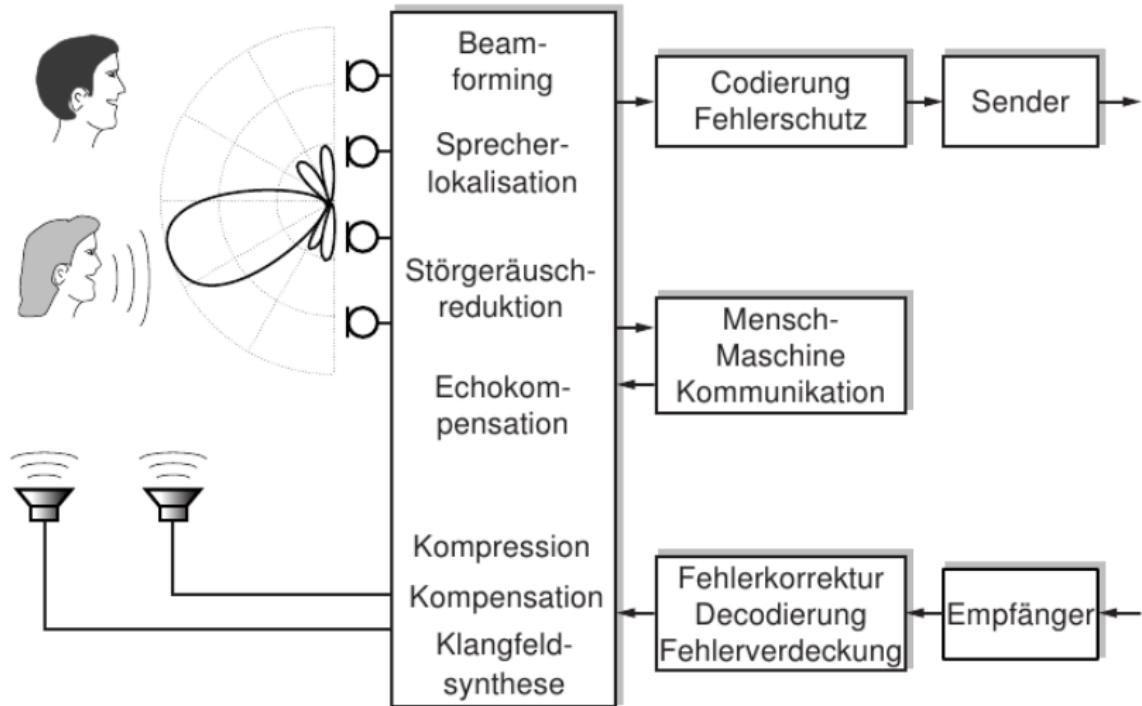
- Robust speech recognition required

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



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

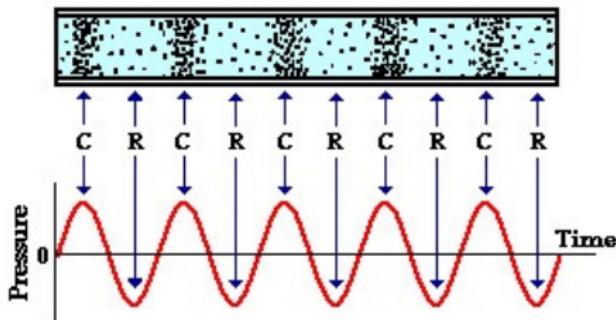




Apple Siri
Google Now / Google Glasses
Microsoft Cortana
Amazon Echo

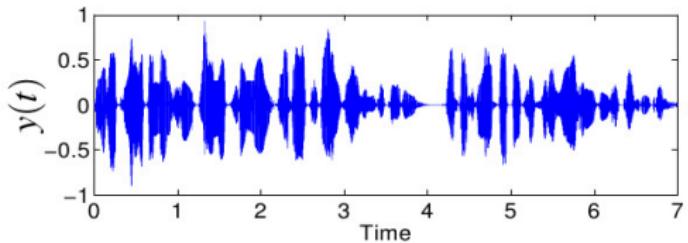


Sound is a Pressure Wave

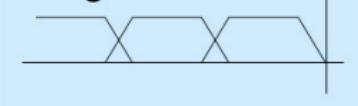


NOTE: "C" stands for compression and "R" stands for rarefaction

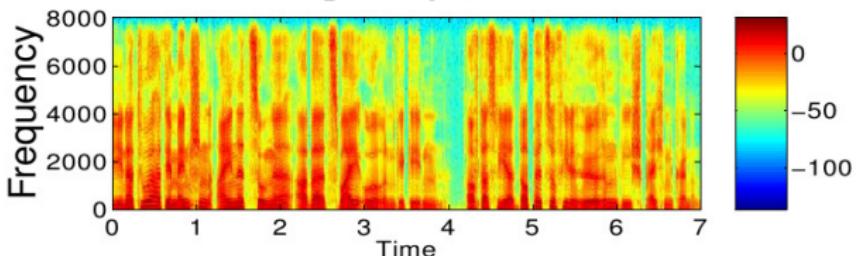
Time Domain Waveform



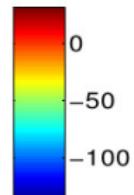
segmentation

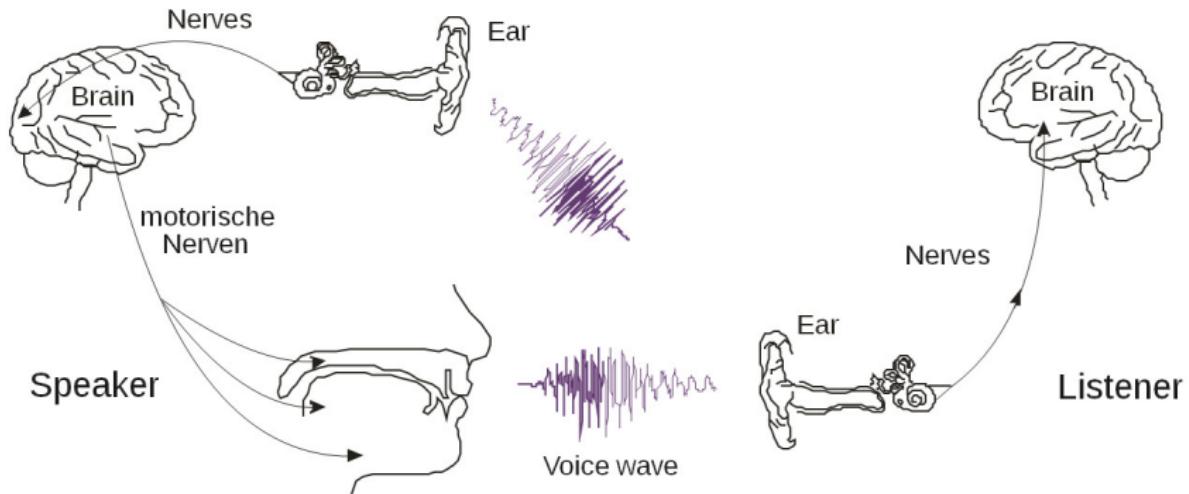


Spectrogram

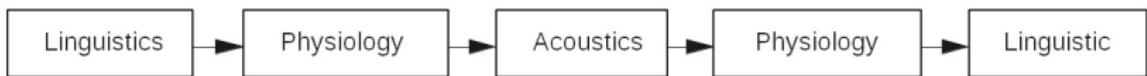


DFT

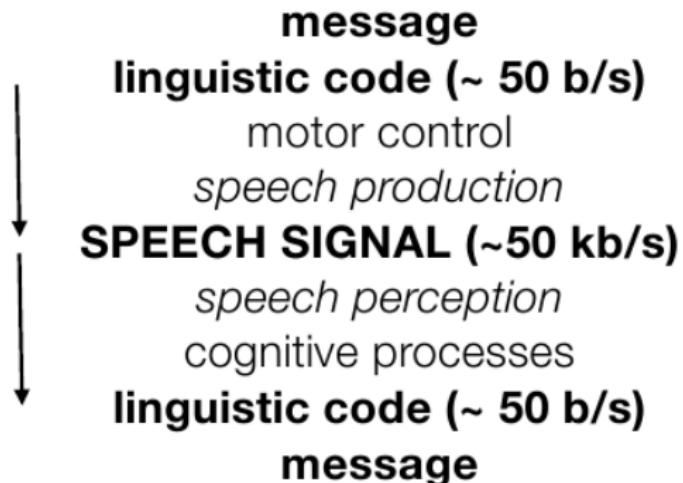




Chain:



Quelle: Vary, Heute, Hess, Digitale Sprachsignalverarbeitung

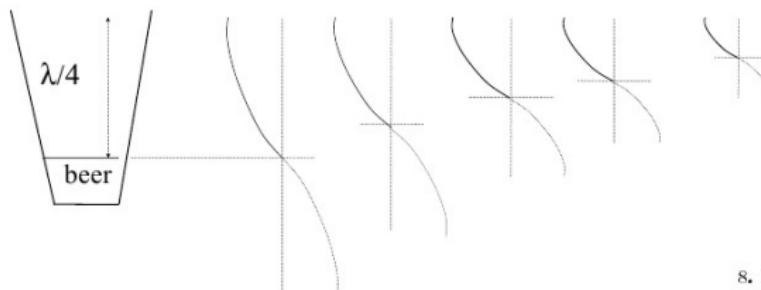


Quelle: Hermansky, lecture on feature extraction, 2005



- 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

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)

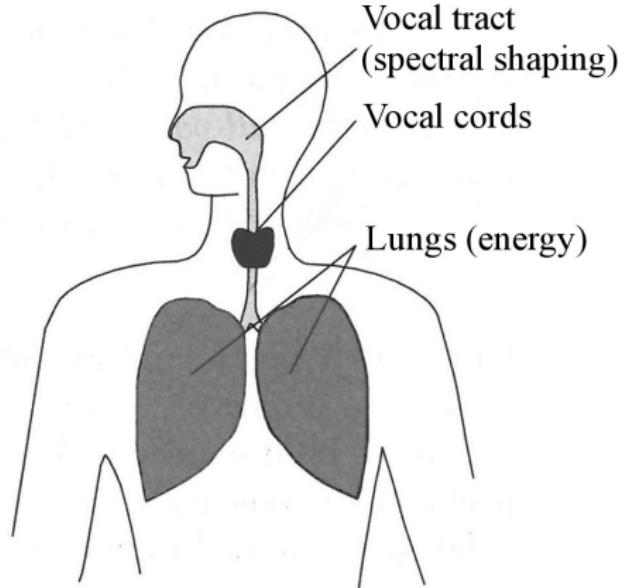
8. P. Ladefoged. *Three Areas of Experimental Phonetics*. Oxford University Press, 1967.

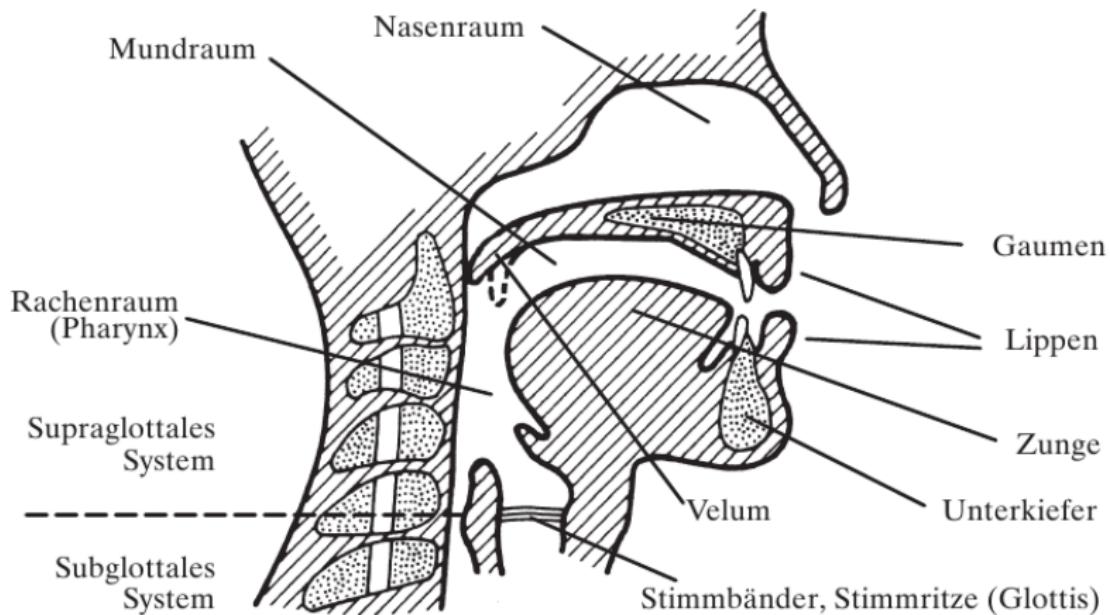


- Homer Dudley (1898 – 1981)
- Changes of sound pressure as a function of time

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. Machine Learning Basics
8. Source Separation

- 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.

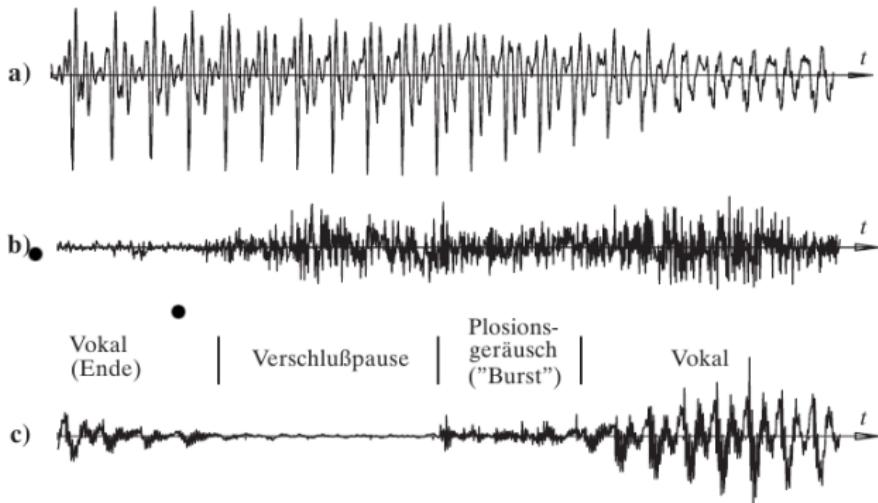




Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

The most important speech sounds are

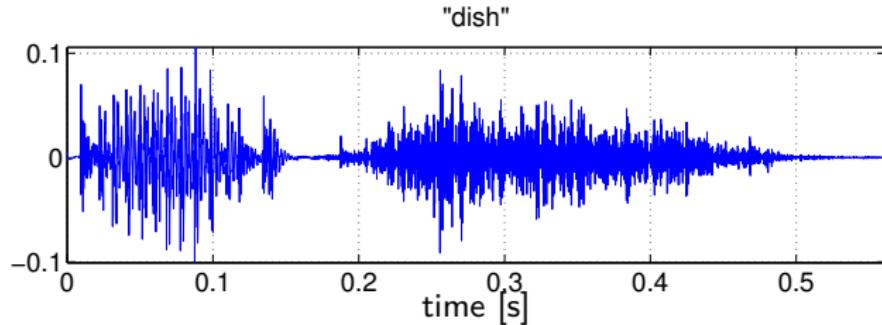
- voiced sounds
 - vowels (a,e,i,o,u)
 - sounds with mixed excitation (/v/)
- unvoiced sounds
 - fricative (/s/,/th/,/sh/)
 - plosive (/k/,/p/,/t/)

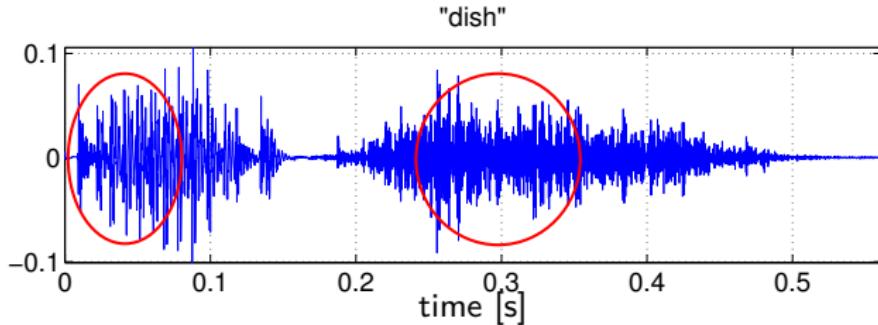


a) stimmhaft b) stimmlos c) Übergang Vokal-Plosiv-Vokal

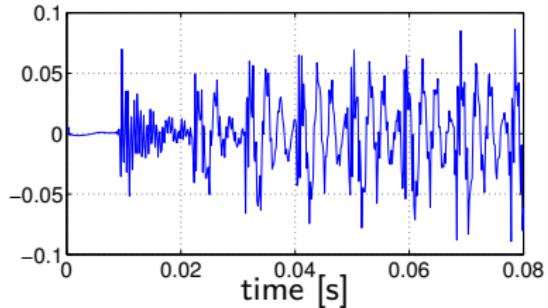
Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

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

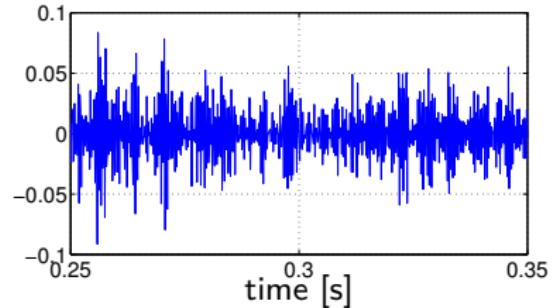




"di" of the word "dish"



"sh" of the word "dish"



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 (→ (allophones)).

Allophone: one phone of the many that constitute a phoneme

Examples:

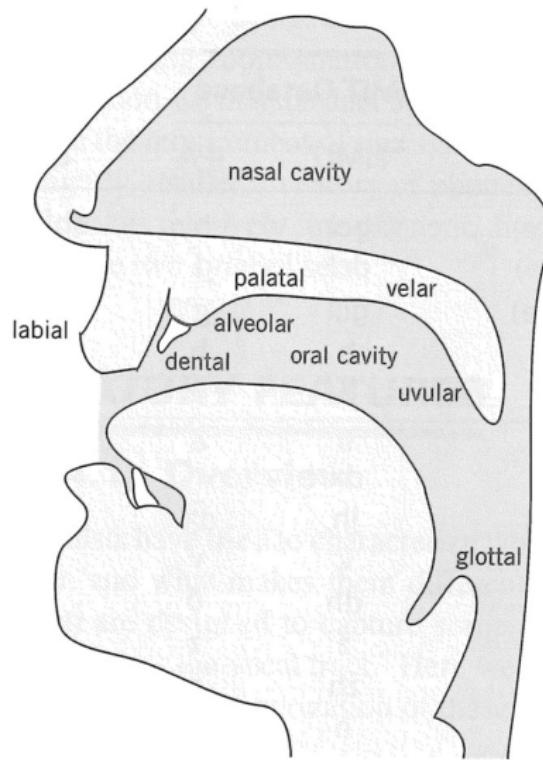
- kiss vs. kill; different in phonemes /ɪ/ and /s/
- cat, kit, school, skill: they all contain the phoneme /k/, but are pronounced differently
- German: Das gerollte und nicht gerollte 'r' sind sind unterschiedliche Phone des gleichen Phonems /r/, und somit Allophone.

- 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

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



Stelle Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t d	t d		t d	c j	k g	q g		?
Nasal	m	nj	n	n		n	n	nj	nj	n	
Affrikate			t̪s dz		tʃ dʒ						
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ɟ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateraler Frikativ				ɬ ɭ							
Trill	B			r					R		
Flap				r		t̪					
Approximant	w	v			ɹ		l̪	j	(w)		
Lateral approximant				l̪		l̪	ʎ				

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

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

Stelle Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t d	t d		t d	c ɟ	k g	q ɢ		?
Nasal	m	n̩	n̩	n		n̩	n̩	n̩	n̩	n	
Affrikate		ts dz		tʃ dʒ							
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ɟ	x ɣ	χ ʁ	h ɦ	h ɦ
Lateraler Frikativ				ɬ ɭ							
Trill	B			r					R		
Flap				r		t̚					
Approximant	w	v			x	-l	j	(w)			
Lateral approximant				l̚		l̚	ʎ				

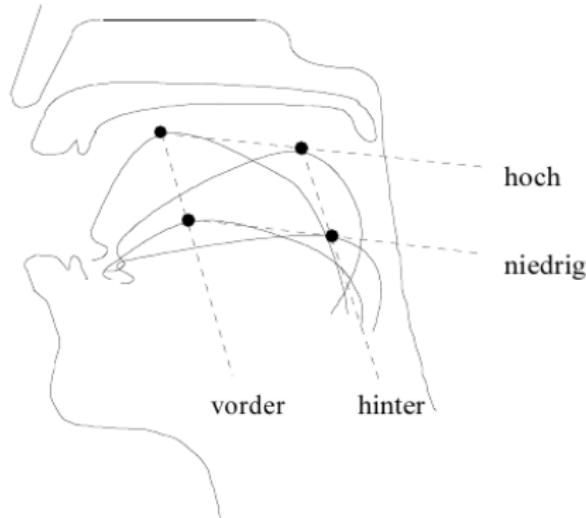
Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

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

Stelle Weise	Bi-Labial	Labio-Dental	Dental	Alveolar	Post-Alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosiv	p b		t d	t d		t d	c ɟ	k g	q ɢ		?
Nasal	m	n̪	n̪	n		n̪	n̪	n̪	n̪	n	
Affrikate			t̪s dz		tʃ dʒ						
Frikativ	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç j	xɣ	χʁ	h ɦ	h ɦ
Lateraler Frikativ				ɬ ɭ							
Trill	B			r					R		
Flap				r		ɾ					
Approximant	w	v			x	ɹ	j	j	(w)		
Lateral approximant				l̪		ɺ	ɻ				

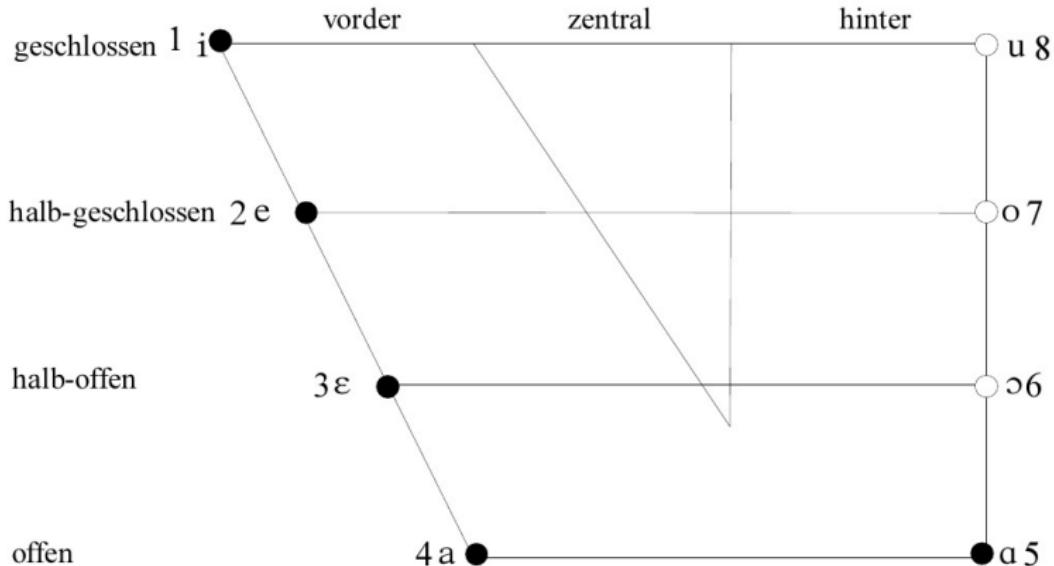
Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen

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



- 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)

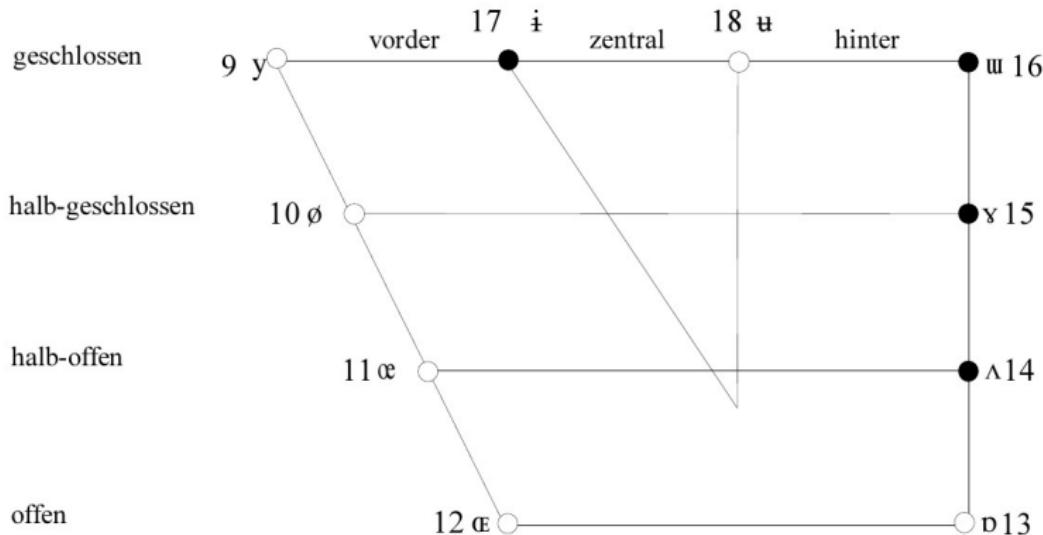
Quelle: Wagner: Skript zur Phonetik und Phonologie, Universität Bremen



○: round lips

●: open lips

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

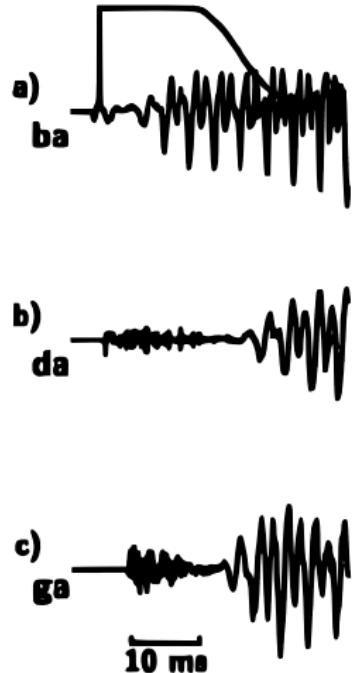


- : round lips
- : open lips

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

Original recordings D. Jones

- The shape of the vocal tract can not change instantly
- Depending on the context, phonemes are pronounced differently
- "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 back of the vocal tract.
 - Effect: Voice Onset Time



- 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.

TABLE 23.2 TIMIT Phone Types

Phones in the TIMIT Database					
TIMIT	IPA	Example	TIMIT	IPA	Example
pcl	p̚	(p closure)	bcl	b̚	(b closure)
tcl	t̚	(t closure)	dcl	d̚	(d closure)
kcl	k̚	(k closure)	gcl	g̚	(g closure)
p	p	pea	b	b	bee
t	t	tea	d	d	day
k	k	key	g	g	gay
q	?	bat	dx	r̚	dirty
ch	tʃ̚	choke	jh	dʒ̚	joke
f	f	fish	v	v	vote
th	θ̚	thin	dh	ð̚	then
s	s	sound	z	z	zoo
sh	ʃ̚	shout	zh	ʒ̚	azure
m	m	moon	n	n	noon
em	m̚	bottom	en	ə̚	button
ng	ŋ̚	sing	eng	ŋ̚	Washington
nx	ř̚	winner	el	l̚	bottle
l	l̚	like	r	r̚	right
w	w̚	wire	y	j̚	yes
hh	h̚	hay	hv	f̚	ahead
er	ə̚	bird	axr	ə̚	butter
iy	i̚	beet	ih	I̚	bit
ey	e̚	bait	eh	ɛ̚	bet
ae	æ̚	bat	aa	a̚	father
ao	ɔ̚	bought	ah	ʌ̚	but
ow	o̚	boat	uh	ʊ̚	book
uw	u̚	boot	ux	ü̚	toot



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. Machine Learning Basics

8. Source Separation

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

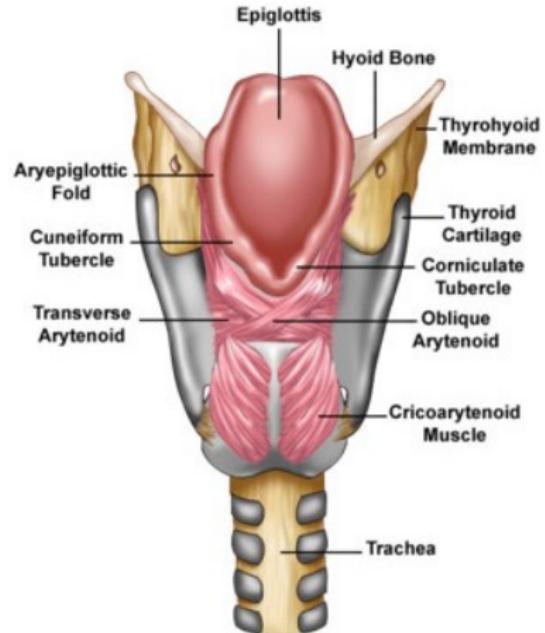
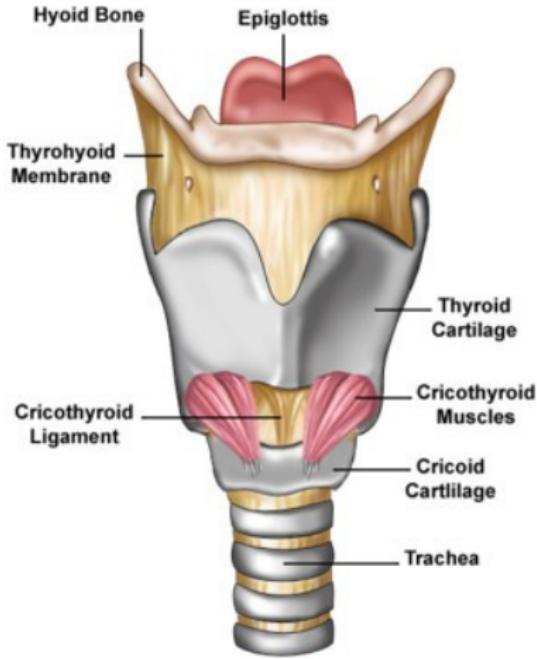


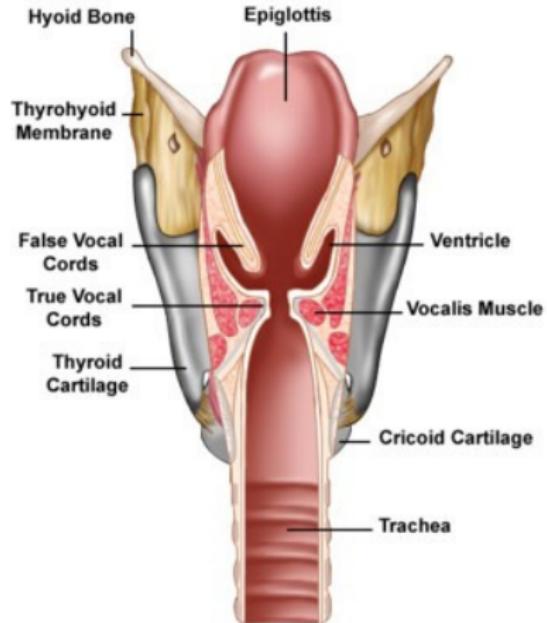
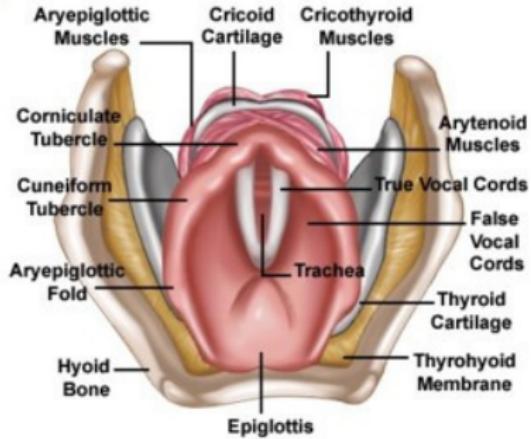
- Simplifying assumption: source and filter are mutually independent

Source: air flow, vibration of vocal cords

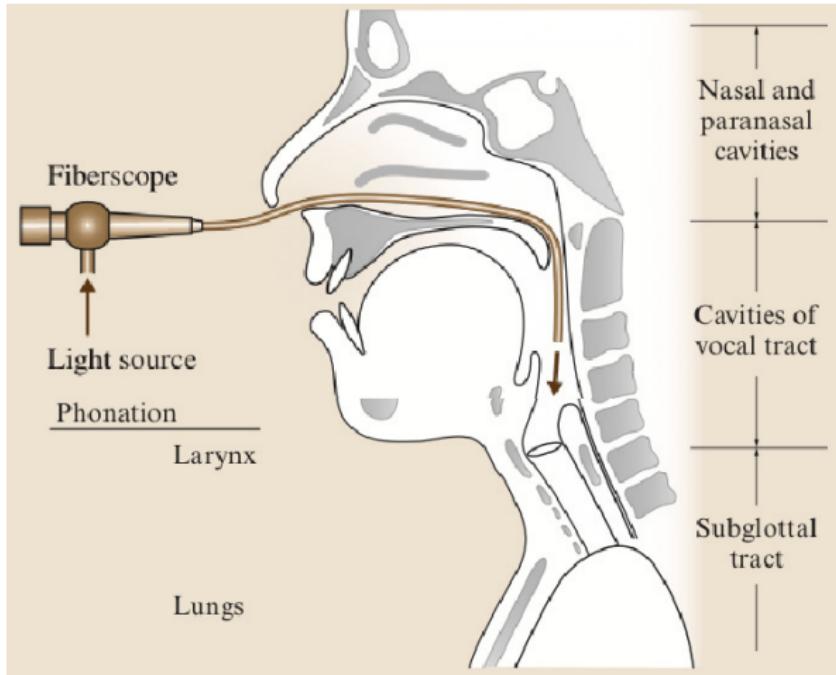
Filter: Shape of the vocal tract: Position of tongue, lips, palate,

...





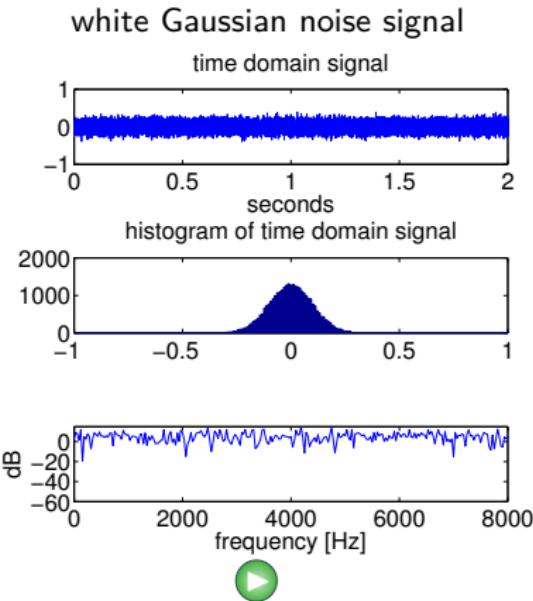
Video



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

Video

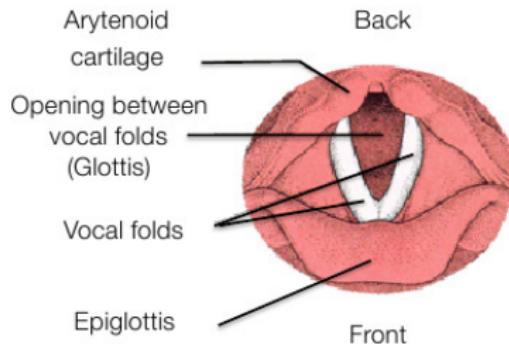
- The unvoiced excitation is noise-like. Can be well described using Gaussian noise.

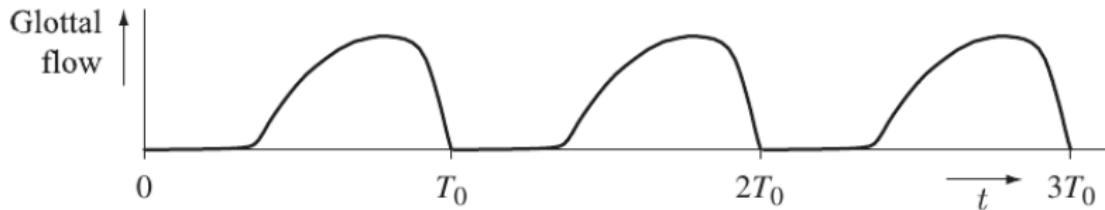


The excitation signal

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.





© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

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

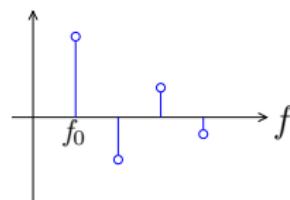
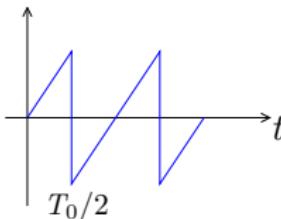
- Opening of glottis due to the increased pressure
- Air flows through the glottis, vocal cords are under tension
- Because of the opening of the glottis, the flow velocity increases while the pressure decreases (Bernoulli-effect)
- The vocal cords snap together, the air flow is interrupted
- The pressure 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 $f_0 = 1/T_0$:

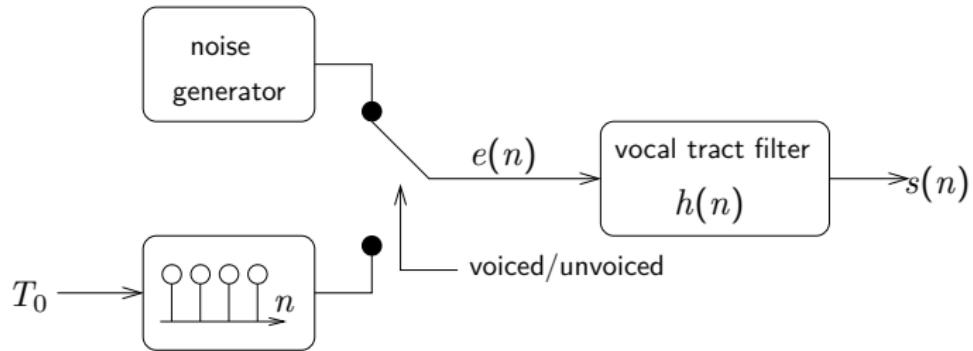
$$g(t) = \frac{a_0}{2} + \sum_{h=1}^{\infty} (a_h \cos(2\pi h f_0 t) + b_h \sin(2\pi h f_0 t))$$

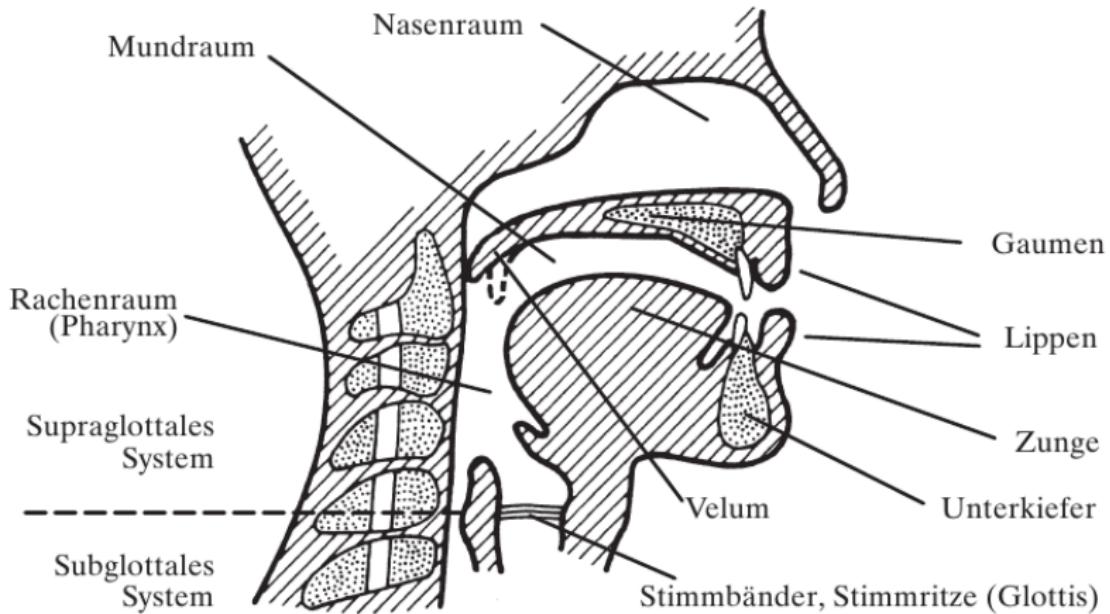
- The glottis signal consists of the fundamental oscillation and its harmonics.

Example:

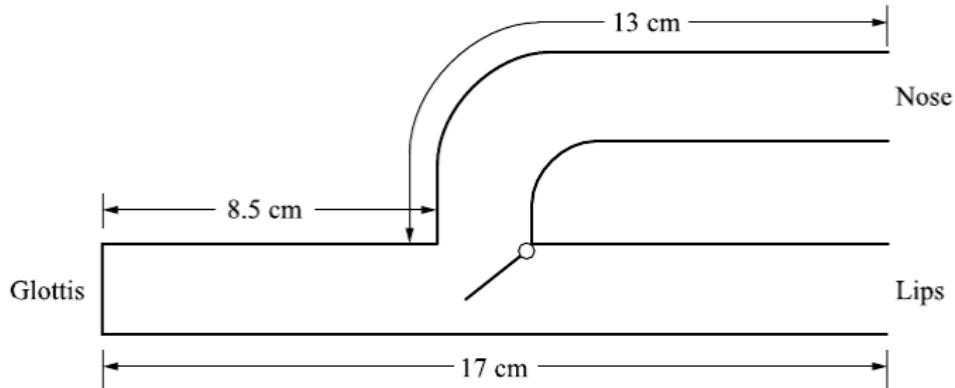


$$g(t) = \frac{1}{\pi f_0} \sum_{h=1}^{\infty} \frac{(-1)^{h-1}}{h} \times \sin(2\pi h f_0 t) \quad (1)$$



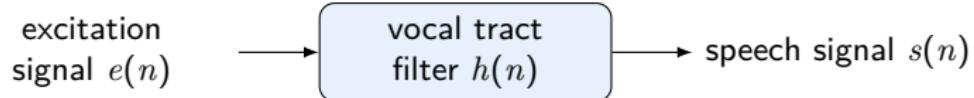


Quelle: Vary, Heute; Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

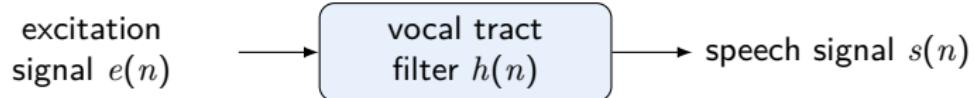


© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

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

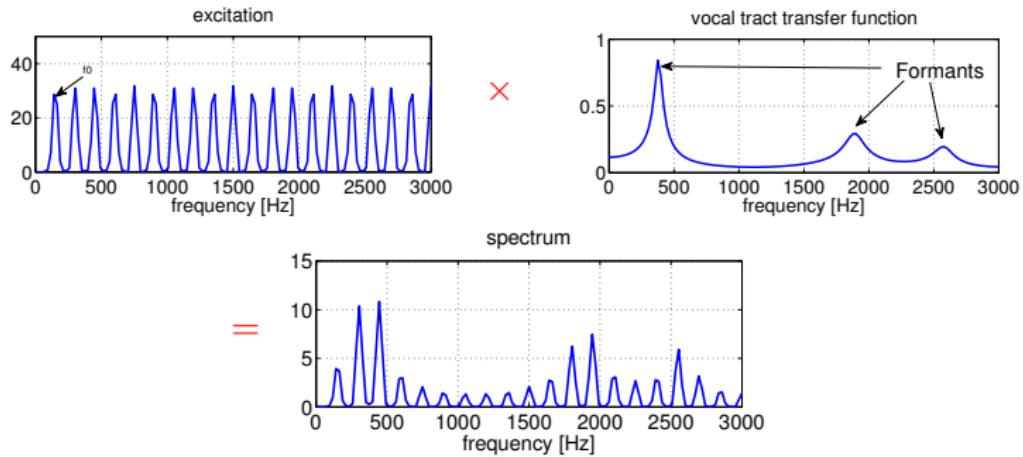


- The vocal tract is modeled by the filter $h(n)$.

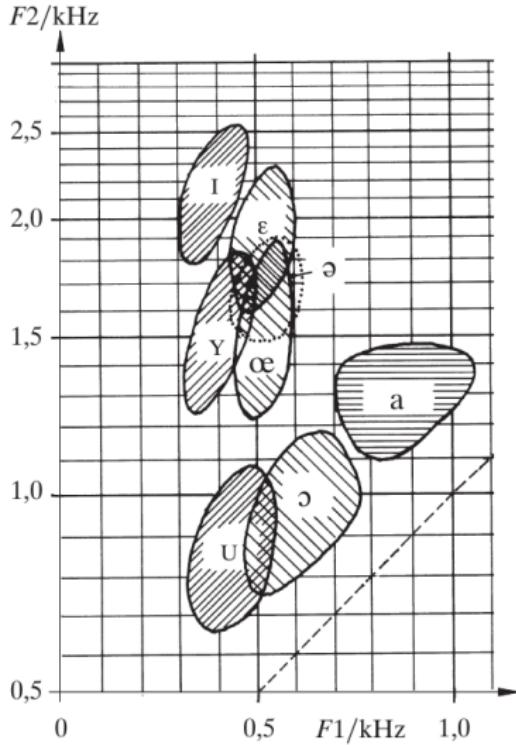
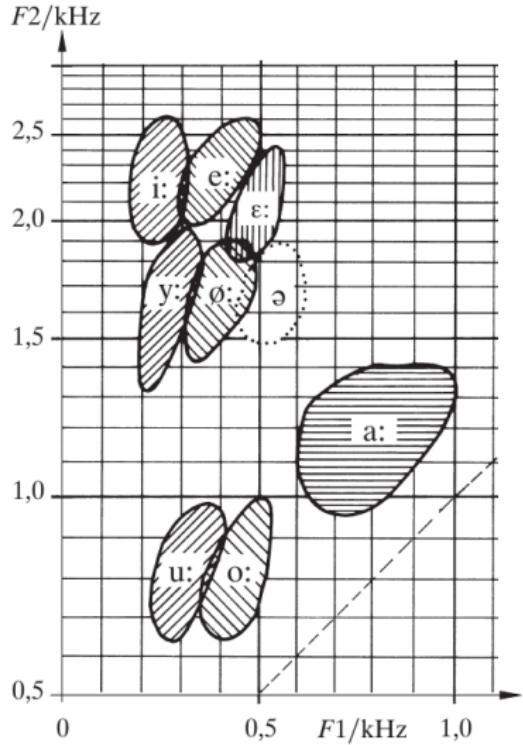


- $s(n) = e(n) * h(n)$ $\circ \bullet \quad S(f) = E(f) \cdot H(f)$

- Spectral decomposition for the utterance "i" in "dish":



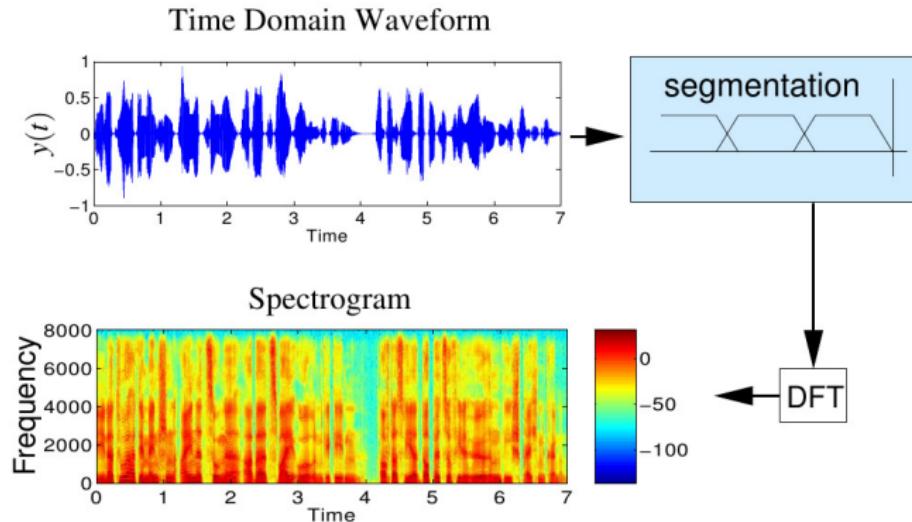
- In the source-filter model, excitation (source) and filter (vocal tract) are treated as being independent.
 - Formants: Peaks of the spectral envelope, resonances of the vocal tract
 - defines the meaning of a phone
 - Fundamental frequency: first peak of the spectral fine structure, and distance between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

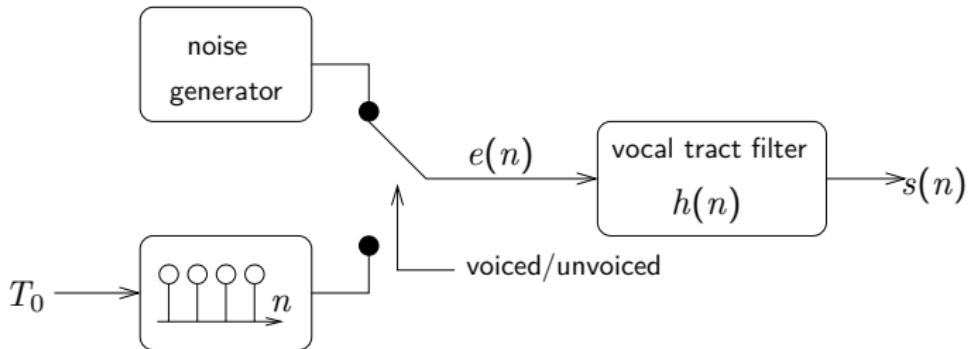


Quelle: Vary, Heute, Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

Speech analysis in wavesurfer:

- Recorded vowels,
- Natural speech





Required Parameters:

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

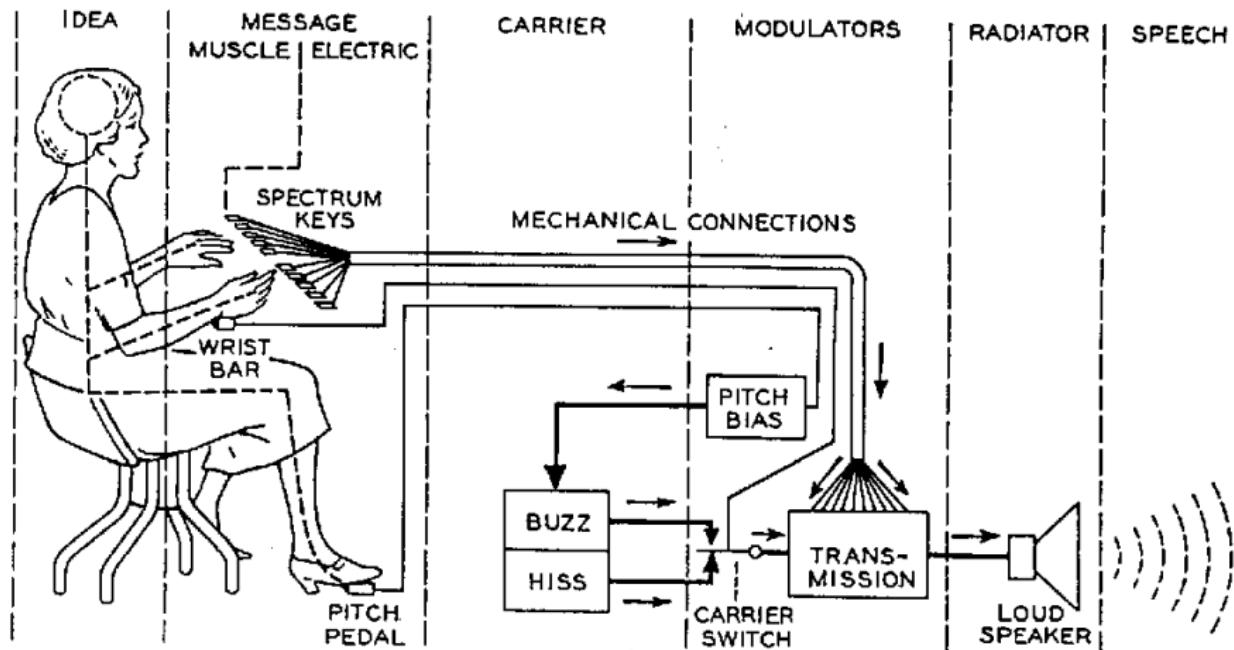


Fig. 8—Schematic circuit of the voder.



Voder Video (long)

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

<https://dood.al/pinktrombone/>

1. Introduction

- Speech Production
- Source-Filter Model
- Hearing

2. Fundamental Frequency Estimation

3. Spectral Analysis of Audio Signals

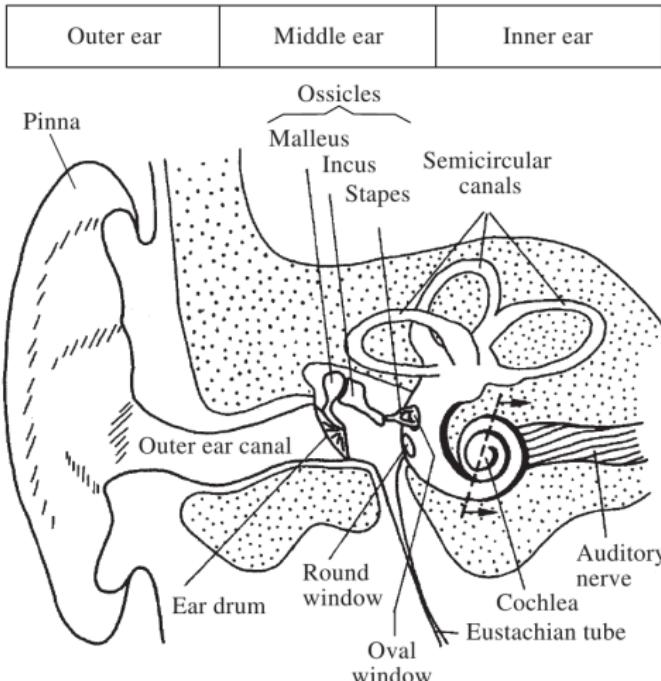
4. Vocal Tract Model and Linear Prediction

5. Sampling, Quantization, and Speech Coding

6. Speech Enhancement

7. Machine Learning Basics

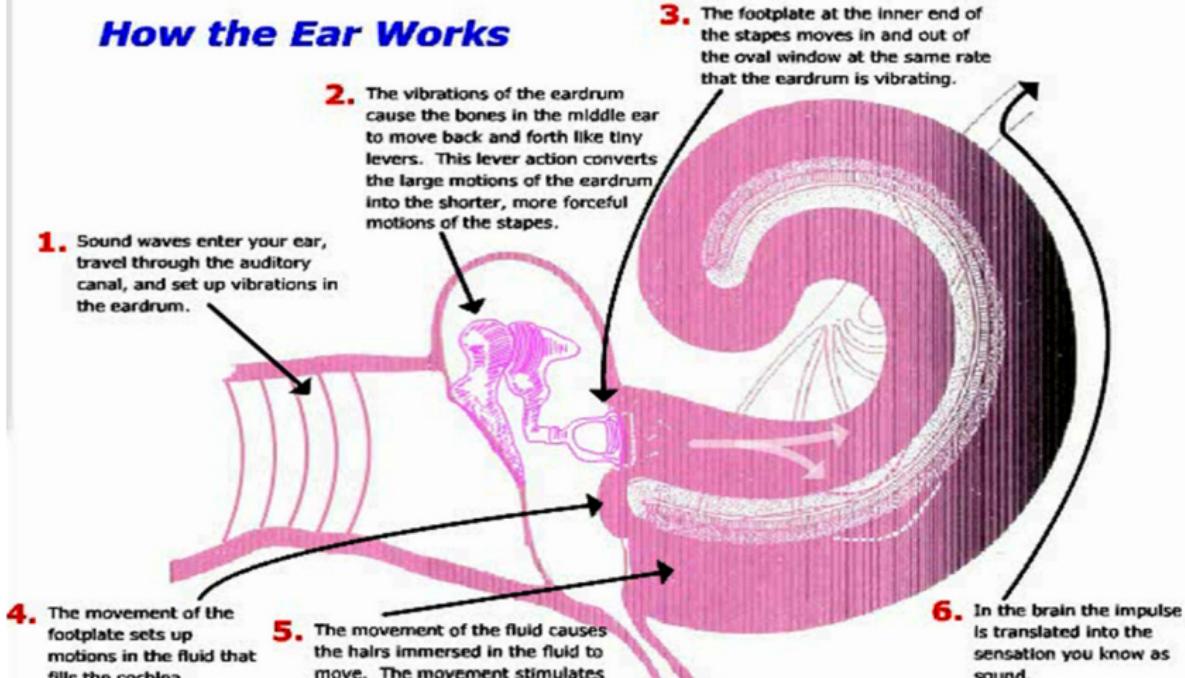
8. Source Separation

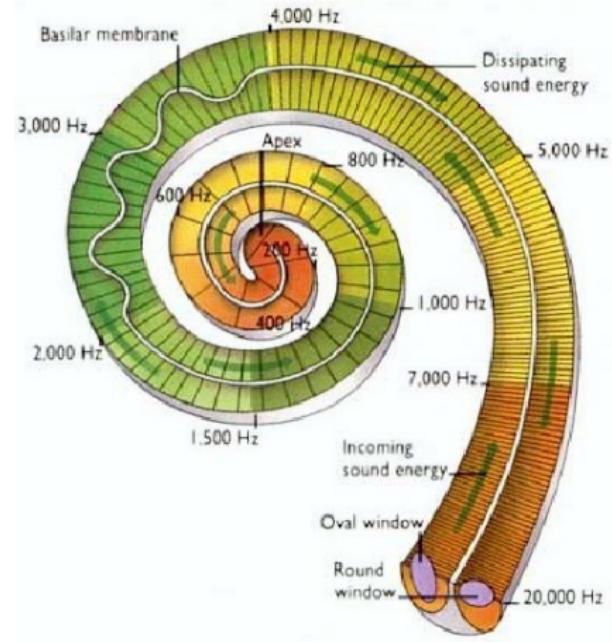


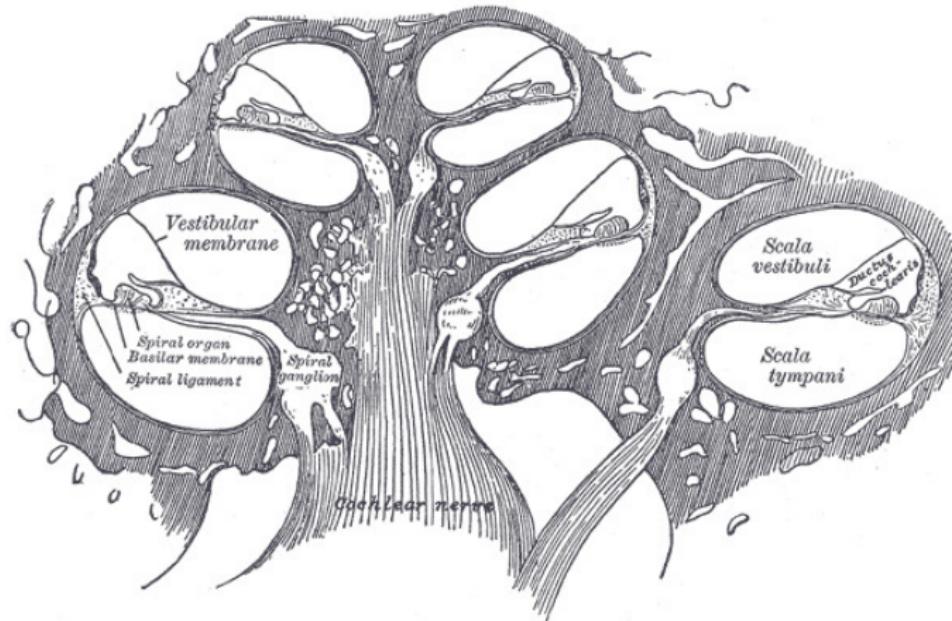
© 1990 Springer Verlag

Quelle: Zwicker, Fastl, 1999

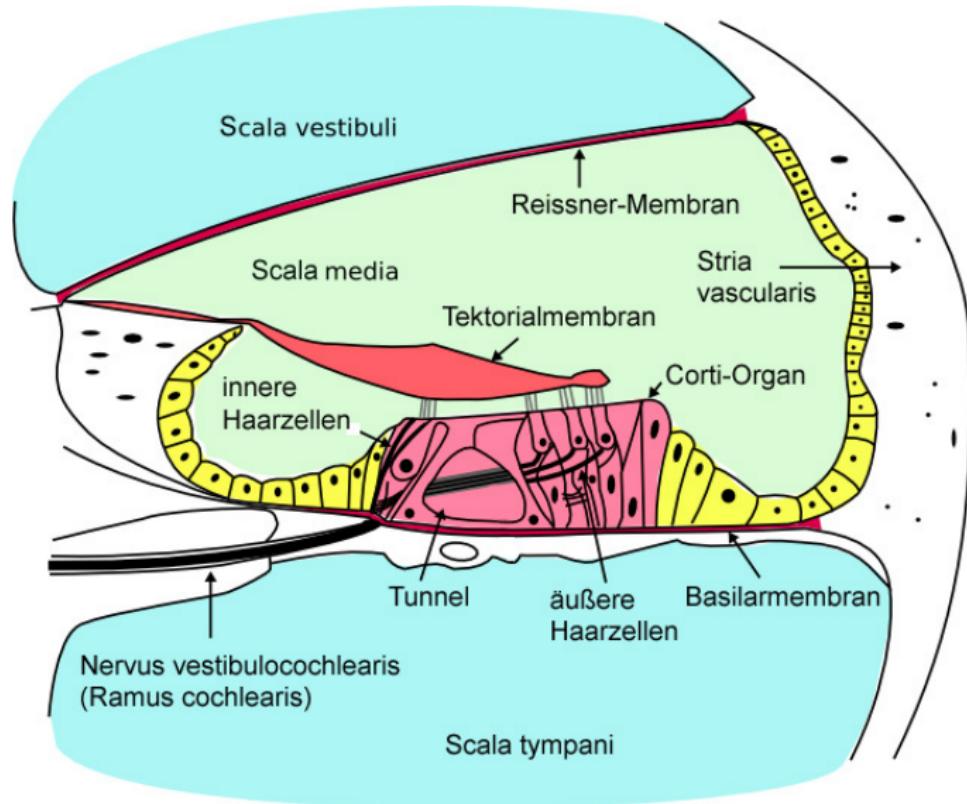
How the Ear Works



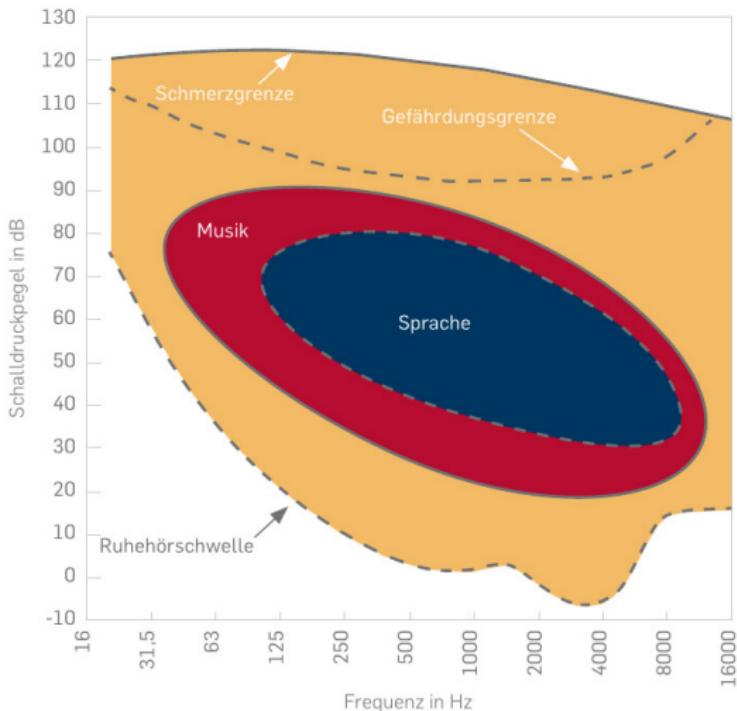




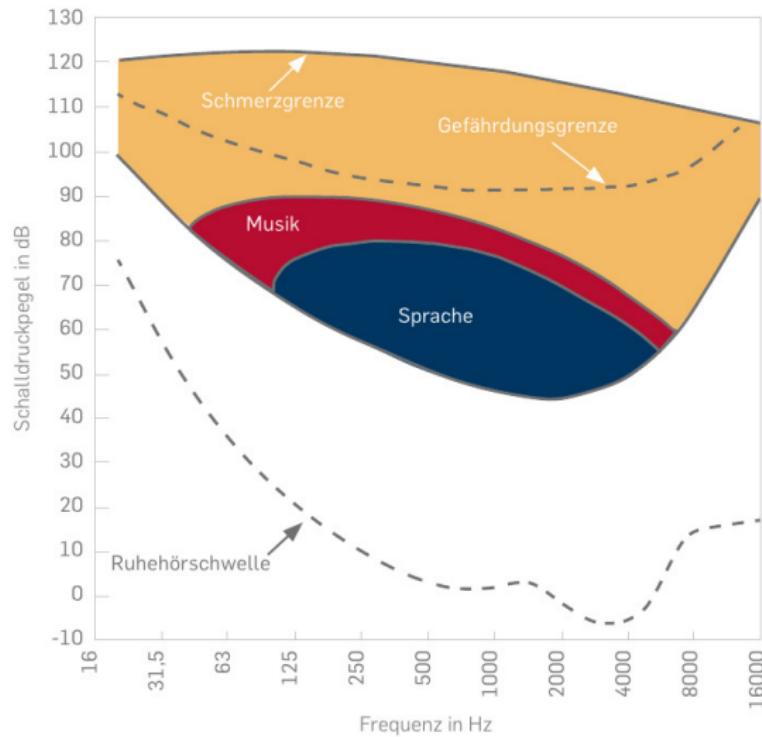
Quelle: Henry Gray, Gray's Anatomy, 2008



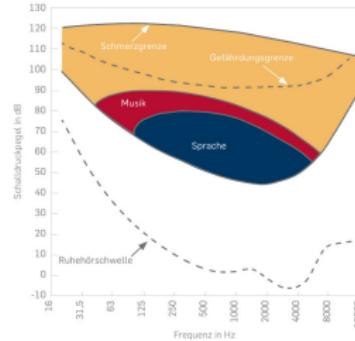
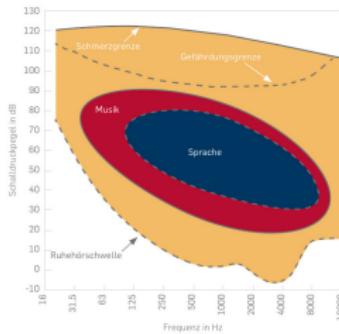
Short Movie



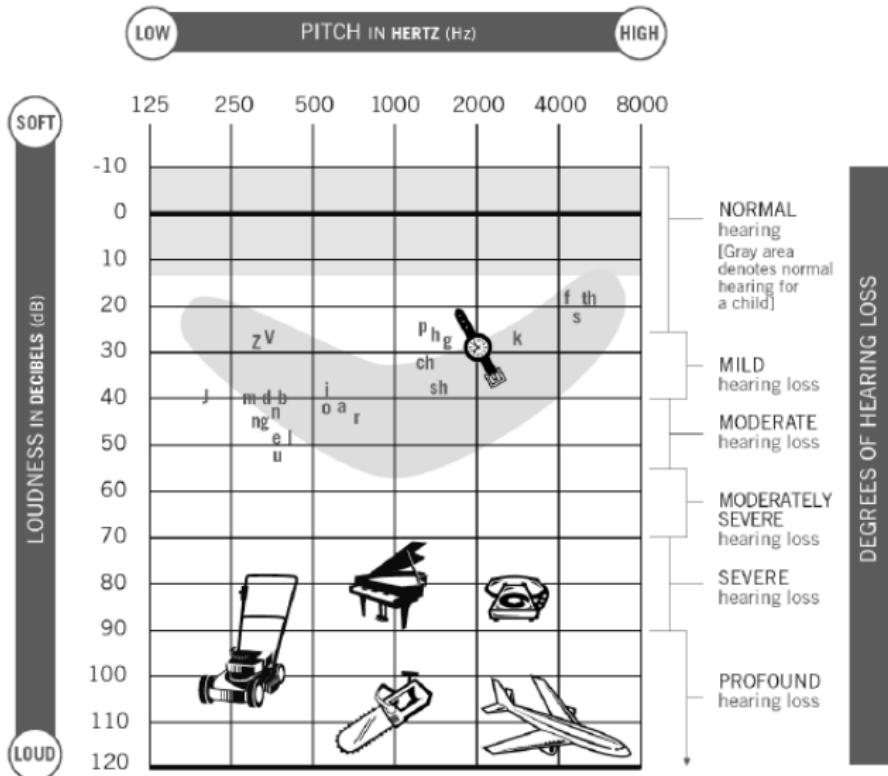
for the Hearing Impaired



For the Hearing Impaired

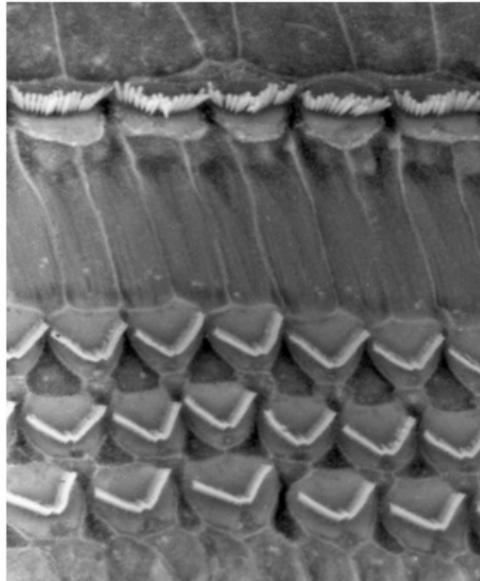


- For hearing impaired the sensation area is reduced.
- Hearing aids can not simply amplify sound.
- Instead, soft sounds are amplified more than louder sounds.
- This decreases the signal-to-noise-ratio (SNR)
- Noise reduction is required!

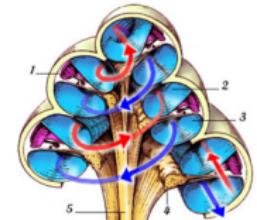
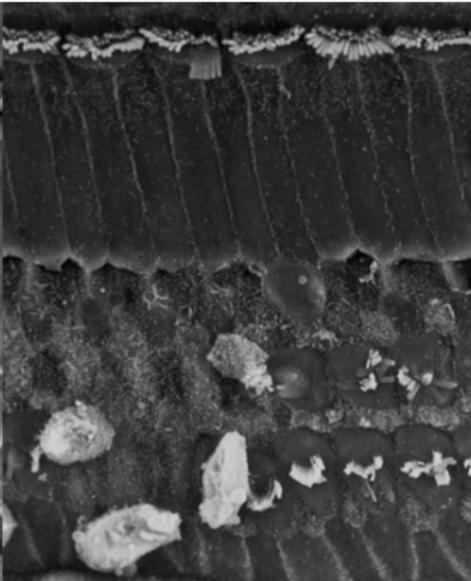


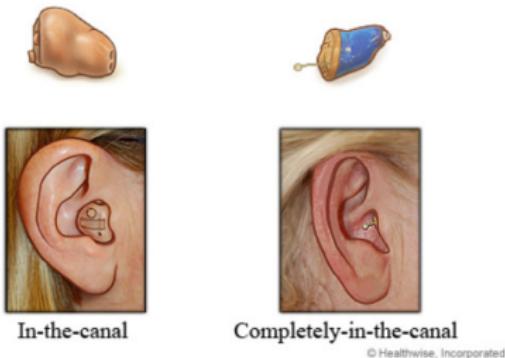
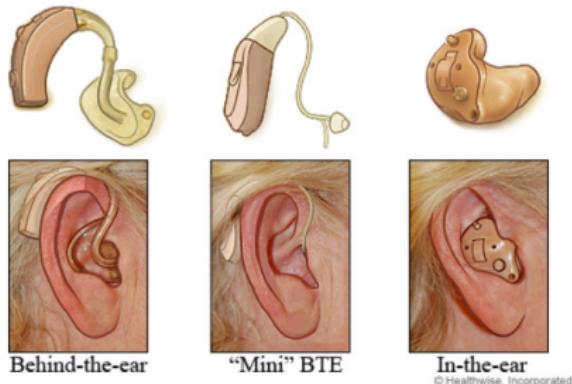
- Conductive hearing loss
 - Sound is not properly conducted through the outer and/or middle ear
 - Sound is perceived but attenuated.
 - Can often be healed and/or well treated with hearing aids
- Sensorineural hearing loss
 - defective inner ear, for instance dead or damaged hair cells
 - often co-occurrence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

healthy hair cells
(from above)

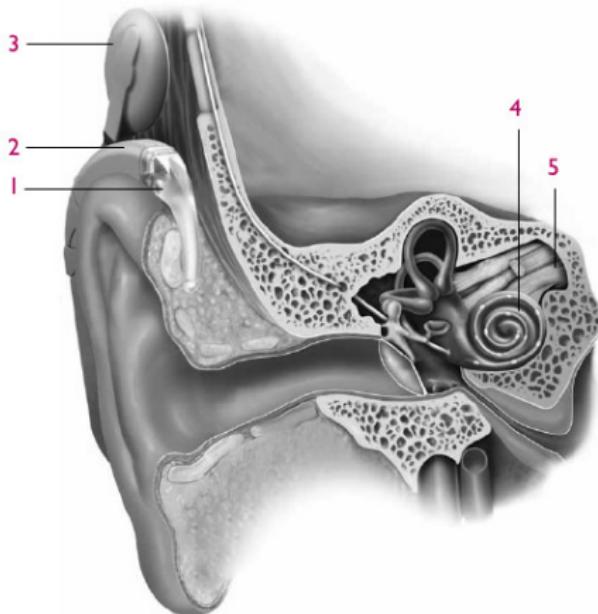


hair cells for sensorineural
deafness



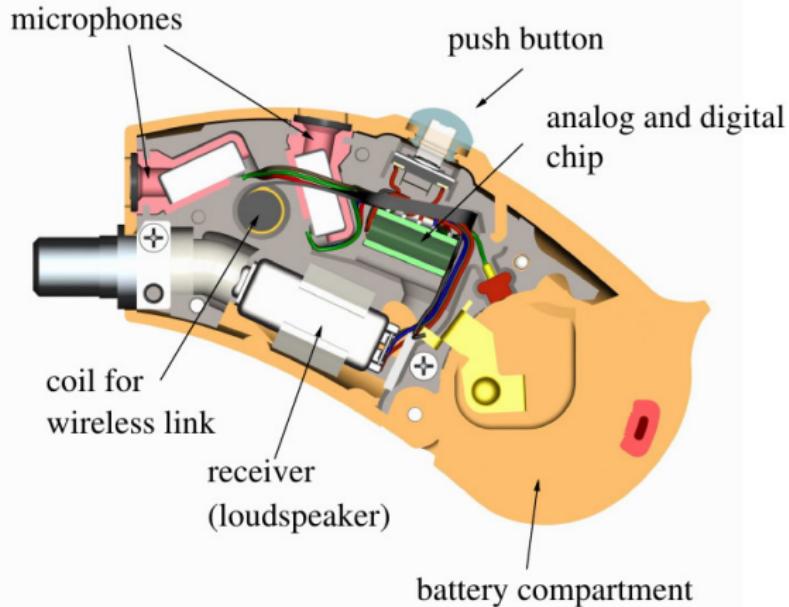


Cochlear Implants

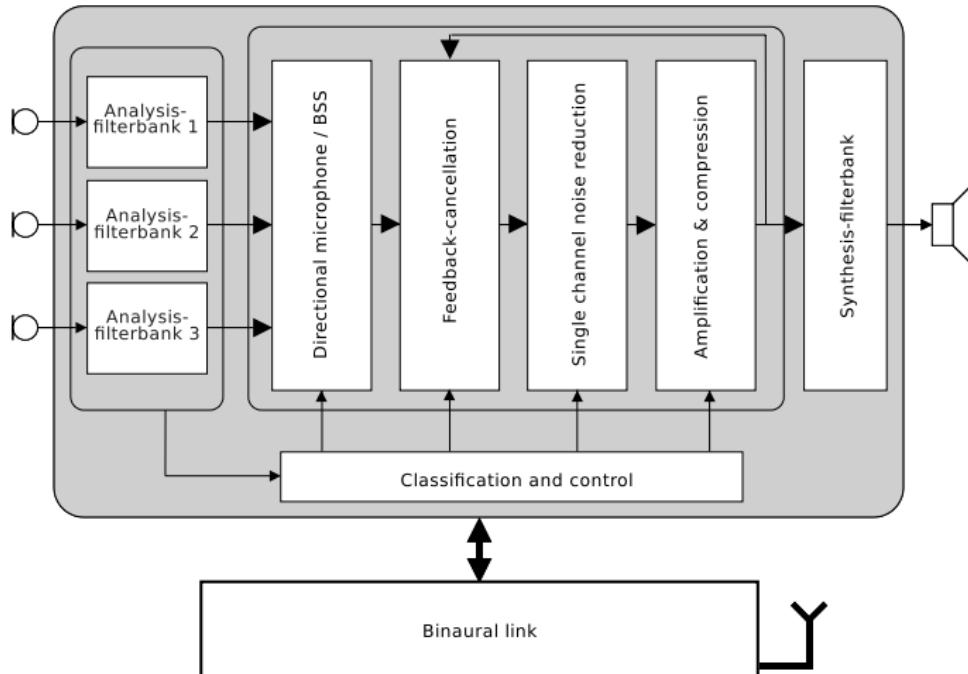


- 1 Sounds are picked up by the microphone.
- 2 The signal is then "coded" (turned into a special pattern of electrical pulses).
- 3 These pulses are sent to the coil and are then transmitted across the skin to the implant.
- 4 The implant sends a pattern of electrical pulses to the electrodes in the 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

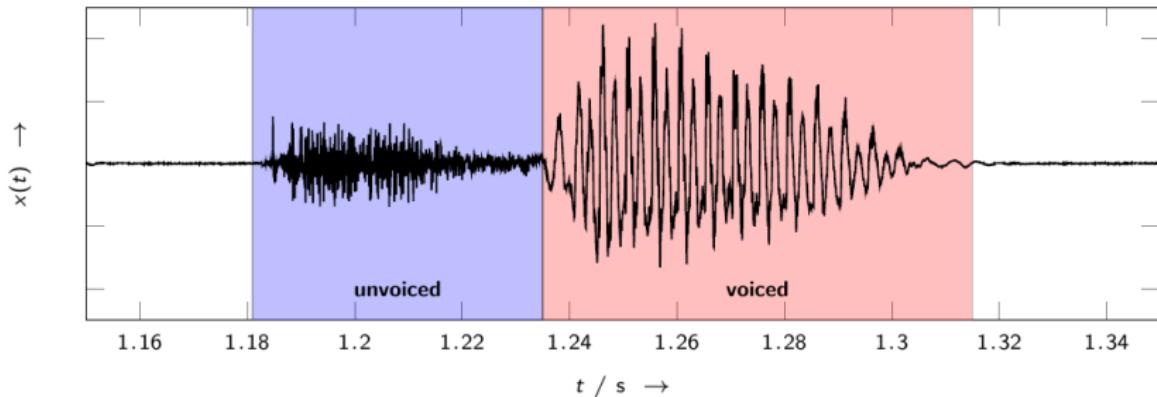
1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics
8. Source Separation
9. Automatic Speech Recognition



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



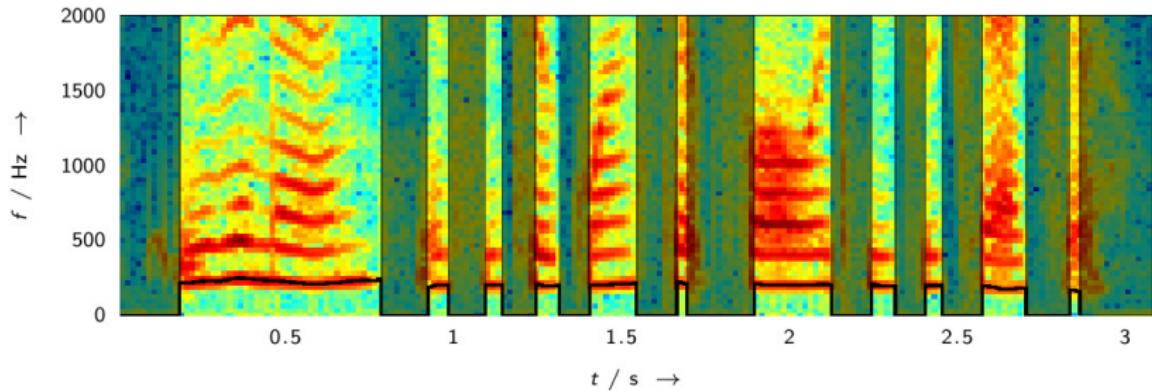
Fundamental Frequency Estimation

**■ Unvoiced speech:**

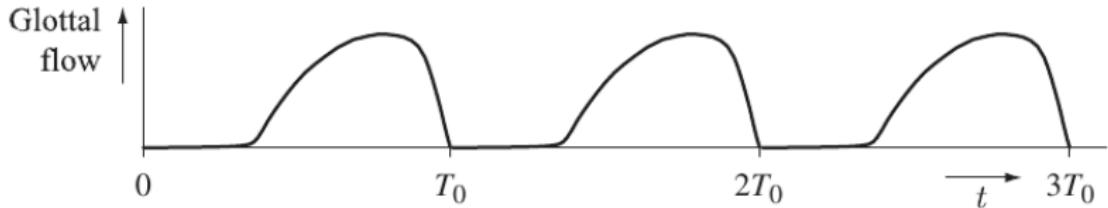
- noise excitation
- pitch not available

■ Voiced speech:

- periodic glottis excitation
- pitch available



- 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 synonymously used. Although, 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



© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

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

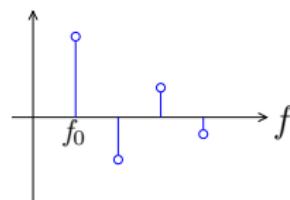
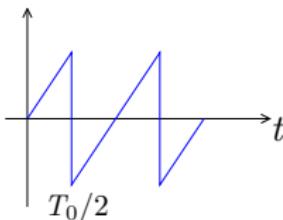
- Opening of glottis due to the increased pressure
- Air flows through the glottis, vocal cords are under tension
- Because of the opening of the glottis, the flow velocity increases while the pressure decreases (Bernoulli-effect)
- The vocal cords snap together, the air flow is interrupted
- The pressure 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 $f_0 = 1/T_0$:

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

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

Example:

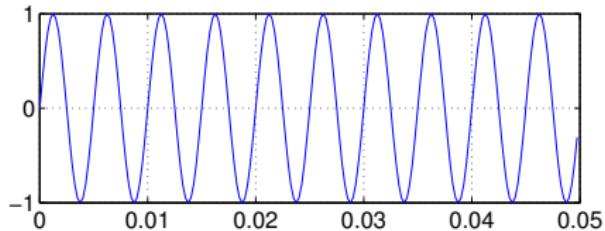


$$g(t) = \frac{1}{\pi f_0} \sum_{h=1}^{\infty} \frac{(-1)^{h-1}}{h} \times \sin(2\pi h f_0 t) \quad (2)$$

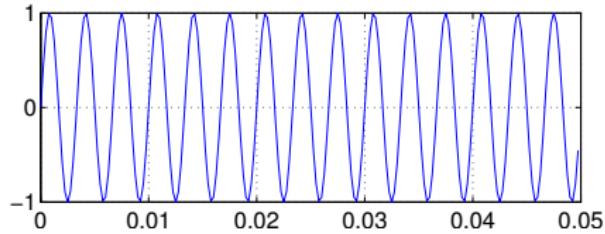
- Narrowband telephone speech is bandpass filtered between 300 Hz and 3400 Hz ("telephone voice")
- The lowest harmonic (fundamental frequency) is often not present in the signal. Still we can distinguish between male and female speakers.
- Example
 - ▶ 200 Hz Ton
 - ▶ 300 Hz Ton
 - ▶ ?

- Narrowband 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

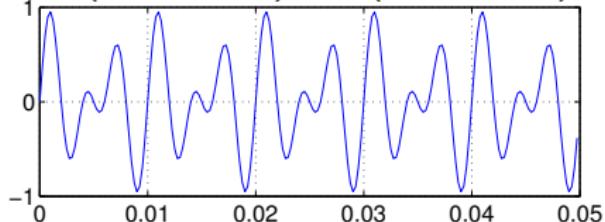
$$\sin(2\pi \cdot 200 \text{ Hz} \cdot t)$$



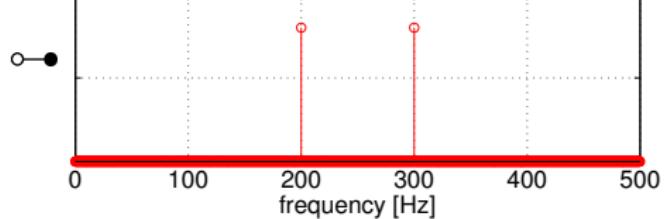
$$\sin(2\pi \cdot 300 \text{ Hz} \cdot t)$$



$$\sin(2\pi \cdot 200 \text{ Hz} \cdot t) + \sin(2\pi \cdot 300 \text{ Hz} \cdot t)$$

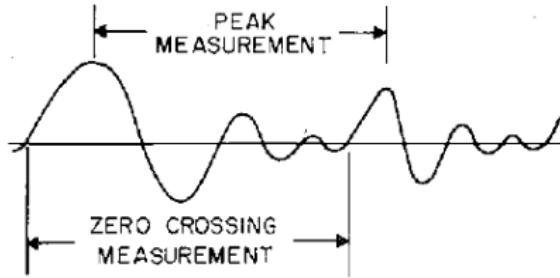


$$\text{DFT}\{\sin(2\pi \cdot 200 \text{ Hz} \cdot t) + \sin(2\pi \cdot 300 \text{ Hz} \cdot t)\}$$



- While typical average speech fundamental frequencies are between 50-300 Hz, frequencies below 300 Hz are not transmitted over the telephone channel.
- Narrowband 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

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

- Autocorrelation function

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

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

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

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

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

- Described by a Power Spectral Density constant over frequency (white)
- Power Spectral Density:

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

- Auto-Correlation function:

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

→ “White noise” implies that successive samples are *uncorrelated*

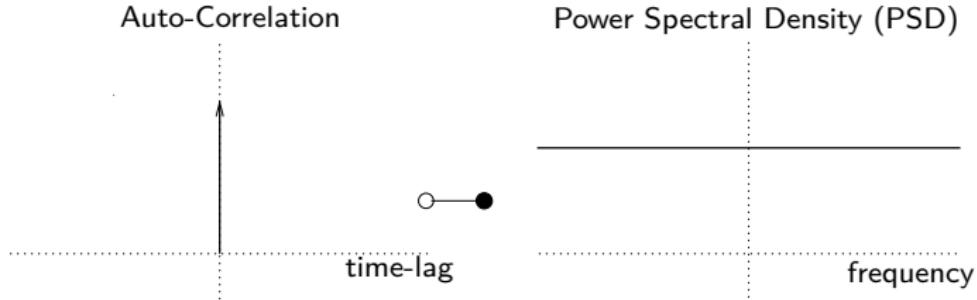
- Described by a Power Spectral Density constant over frequency (white)
- Power Spectral Density:

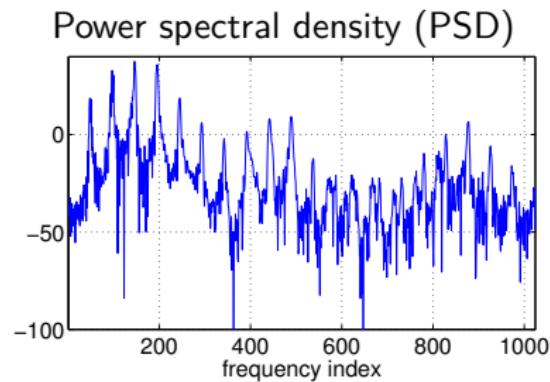
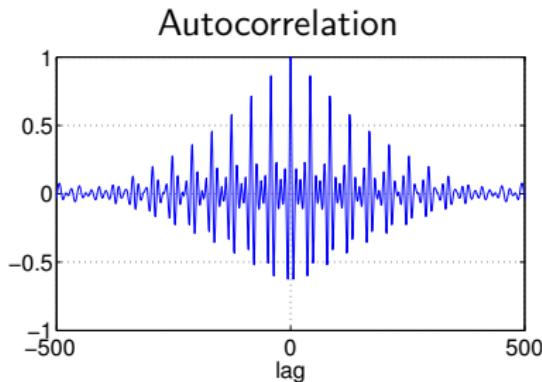
$$\Phi_X(\Omega) = \sigma_X^2$$

- Auto-Correlation function:

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

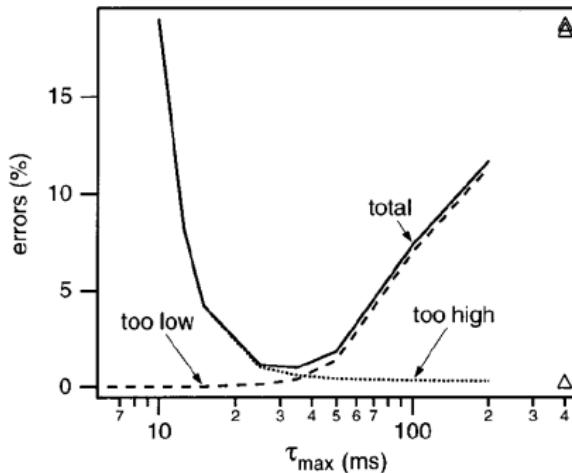
→ “White noise” implies that successive samples are *uncorrelated*





- $f_s = 8 \text{ kHz}$, Segment length: 128 ms, DFT length: 2048 (256 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 frequency $f_0 = 1/T_0$.

- 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)
- ≈ 30 ms is a good compromise (3 periods at $f_0 = 100$ Hz).



- Difference approach: for a signal that is periodic in N_0 , we have

$$x(n) - x(n + N_0) = 0, \quad \forall t$$

- The same holds for the square of the difference if $\lambda = N_0$

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

- Approach: find the λ that minimizes d_λ

→ This is the same as computing

$$d_\lambda = \hat{\varphi}_{x(n)}(0) + \hat{\varphi}_{x(n+N_0)}(0) - 2\hat{\varphi}_{x(n)}(N_0)$$

- If $\hat{\varphi}_{x(n+N_0)}(0) = \hat{\varphi}_{x(n)}(0)$ the ACF and the difference approach are equivalent

- 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 algorithm¹.

¹ A. d. Chevigné 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.

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
 - Fourier Theory and Complex Numbers
 - Linear Time-Invariant Systems
 - Discrete-time Fourier, DFT, and STFT
 - The z -Transform
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Spectral Analysis of Audio Signals

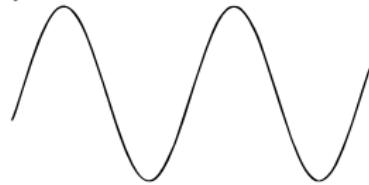
- **Purpose** of a spectral transformation is to transform a signal into a domain, where certain properties are more easily accessible.
- **Fourier transform** is a concept where a signal is decomposed into its frequency components.
 - Perfectly invertible
 - Mathematically elegant (orthogonal set of basis functions)
 - Comparably low computational complexity (FFT) $O(n \log n)$
 - Allows for a different view of the signal
- **Advantages for speech signals**
 - Formants and spectral harmonics are visible
 - Decorrelation of speech coefficients
 - Intuitive representation, as the human ear also performs a frequency analysis in the inner ear
 - Many filters (e.g. high-pass, low-pass) are easily described in frequency domain.

For speech signals, useful spectral representations are

- Short-time (discrete) Fourier transform (DFT, STFT)
- Short-time (discrete) Cosine transform (DCT)
- Eigenvalue/eigenvector decomposition (Karhunen-Loëve transform, KLT)
- (Generalized) singular value decomposition (SVD, GSVD)
- Filter bank coefficients
- Wavelet transform coefficients
- Parametric model coefficients: autoregressive parameters (LPC), reflection coefficients, log-area ratios, cepstral coefficients, Line Spectral Frequencies (LSF), mel-frequency cepstral coefficients (MFCC)

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
 - Fourier Theory and Complex Numbers
 - Linear Time-Invariant Systems
 - Discrete-time Fourier, DFT, and STFT
 - The z -Transform
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics

- pure tone, sinusoid



$$x(t) = A \sin(2\pi f_0 t + \phi)$$

- periodic signal \rightarrow sum of harmonics



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

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

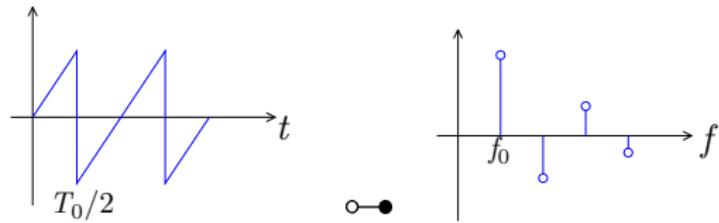
$$x(t) = \frac{a_0}{2} + \sum_{h=1}^{\infty} (a_h \cos(2\pi h f_0 t) + b_h \sin(2\pi h f_0 t))$$

- The coefficients a_h, b_h are obtained by computing the “similarity” between the signal and a sine/cosine as

$$a_h = \frac{2}{T_0} \int_0^{T_0} x(t) \cos(2\pi h f_0 t) dt$$

$$b_h = \frac{2}{T_0} \int_0^{T_0} x(t) \sin(2\pi h f_0 t) dt$$

Example:



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

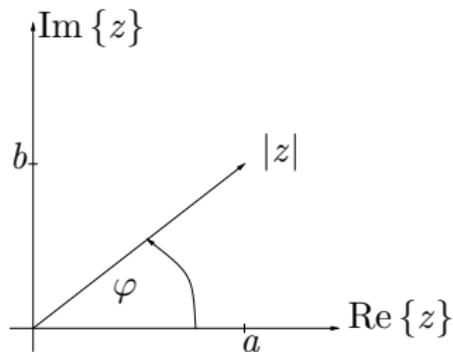
- We see that the Fourier series consists of the comparison with (1) sine and (2) cosine functions.
- Can be expressed much more elegantly by means of **complex numbers**
- For many spectral transformations, complex numbers are needed
- A complex number z is composed of a real a and an imaginary part b

$$z = \operatorname{Re}\{z\} + j\operatorname{Im}\{z\} = a + jb$$

- j is the **imaginary unit** and separates the real and imaginary parts

Visualization

- To understand the different representations of complex numbers, draw a diagram



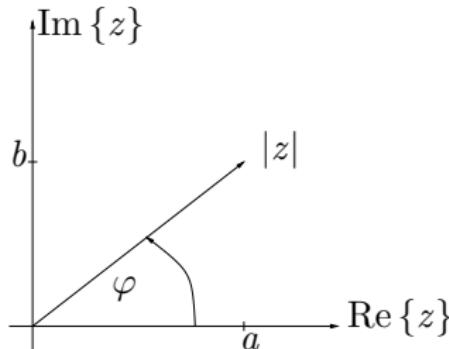
- Complex numbers can be represented by
 - real and imaginary parts (Cartesian coordinates)

$$z = a + jb$$

- Absolute value and phase (polar coordinates)

$$z = |z| e^{j\phi}$$

Different Representations



- Simple geometry reveals the relations
 - Pythagoras: $|z| = \sqrt{a^2 + b^2}$
 - Trigonometry:
 - $\varphi = \arctan\left(\frac{b}{a}\right)$
 - $a = |z| \cos \varphi$
 - $b = |z| \sin \varphi$

Euler relation (memorize!)

$$e^{j\varphi} = \cos \varphi + j \sin(\varphi)$$

Calculus

Addition is most easily done in Cartesian coordinates

$$z_1 + z_2 = (a_1 + a_2) + j(b_1 + b_2)$$

- Real and imaginary parts are added separately

Multiplication is most easily done in polar coordinates

$$z_1 z_2 = |z_1| |z_2| e^{j(\varphi_1 + \varphi_2)}$$

- Absolute values are multiplied, the phases add

Conjugate For a **complex conjugate** z^* of z , the sign of the imaginary part is flipped

$$z = a + jb = |z| e^{j\varphi}$$

$$z^* = a - jb = |z| e^{-j\varphi}$$

- mirror complex vector along Re-axis

Some conclusions

1. Transform $e^{j\pi}$ to Cartesian coordinates
2. Transform $e^{j\frac{\pi}{2}}$ to Cartesian coordinates
3. Express $\frac{1}{j}$ in terms of polar coordinates
4. Express $\frac{1}{j}$ in terms of Cartesian coordinates
5. Express $\sqrt{-1}$ in terms of polar coordinates
6. Express $\sqrt{-1}$ in terms of Cartesian coordinates

Lösung siehe Notizen

- The Fourier series representation seems a bit complicated

$$x(t) = \frac{a_0}{2} + \sum_{h=1}^{\infty} (a_h \cos(h\omega_0 t) + b_h \sin(h\omega_0 t))$$

- Introducing the complex notation yields a much more handy representation

$$c_h = \frac{1}{2}(a_h - jb_h); \quad c_h^* = \frac{1}{2}(a_h + jb_h)$$

- With Eulers relation we have

$$\cos(h\omega_0 t) = \frac{1}{2}(e^{jh\omega_0 t} + e^{-jh\omega_0 t}); \quad \sin(h\omega_0 t) = \frac{1}{j2}(e^{jh\omega_0 t} - e^{-jh\omega_0 t})$$

- For real-valued signals we have $c_{-h} = c^*$ and obtain the compact formulation

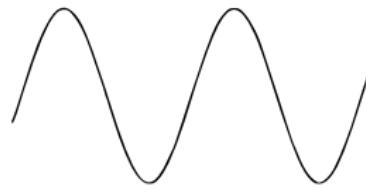
$$x(t) = \sum_{h=-\infty}^{\infty} c_h e^{jh\omega_0 t}$$

with

$$c_h = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-jh\omega_0 t} dt$$

- Pure tone, sinusoid

$$x(t) = A \sin(\omega t + \phi)$$



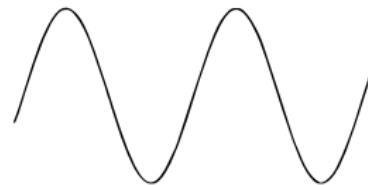
- Periodic signal: sum of harmonics

$$x(t) = \sum_{-\infty}^{\infty} c_h e^{j h \omega_0 t}$$



- Pure tone, sinusoid

$$x(t) = A \sin(\omega t + \phi)$$



- Periodic signal: sum of harmonics

$$x(t) = \sum_{-\infty}^{\infty} c_h e^{j h \omega_0 t}$$



- Arbitrary signal: integral over **all** frequencies

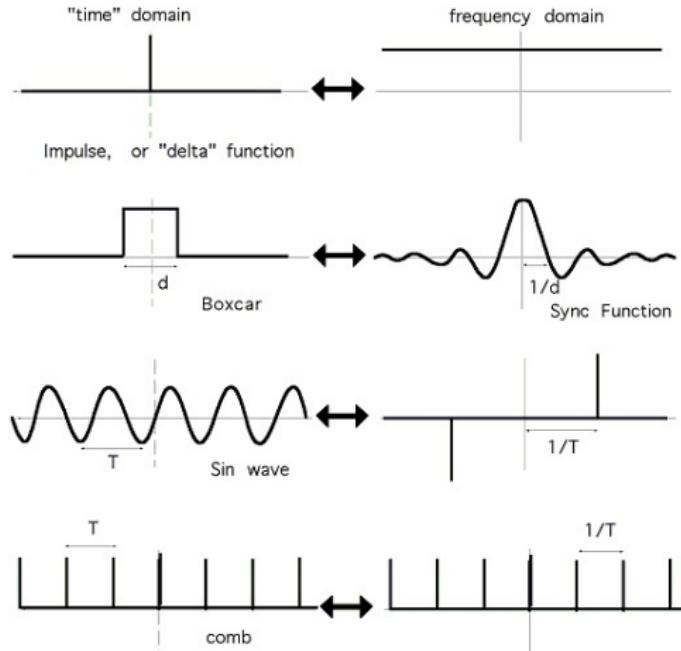
$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega) e^{j\omega t} d\omega$$



Continuous-time Fourier transform

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t} dt \quad x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)e^{j\omega t} d\omega$$

where $X(j\omega)$ is a continuous and non-periodic function of $\omega = 2\pi f$.



$$x(t) \circledleftarrow X(\omega)$$

$$h(t) \circledleftarrow H(\omega)$$

convolution $x(t) * h(t) \circledleftarrow X(\omega)H(\omega)$ multiplication
multiplication \circledleftarrow convolution

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
 - Fourier Theory and Complex Numbers
 - Linear Time-Invariant Systems
 - Discrete-time Fourier, DFT, and STFT
 - The z -Transform
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics



- $T\{\cdot\}$ defines the mapping from input $x(n)$ to output $y(n)$:

$$y(n) = T\{x(n)\}$$

- Linear and time-invariant system ($a, b \in \mathbb{C}$)

$$\begin{aligned} T\{ax_1(n) + bx_2(n)\} &= aT\{x_1(n)\} + bT\{x_2(n)\} \\ y(n - n_0) &= T\{x(n - n_0)\} \end{aligned}$$

$$\begin{aligned} y(n) = x(n) * h(n) &= \sum_{m=-\infty}^{\infty} x(m)h(n-m) \\ &= \sum_{m=-\infty}^{\infty} h(m)x(n-m) = h(n) * x(n) \end{aligned}$$



- In time domain, the relation between input and output is given by a **convolution**

$$y(n) = x(n) * h(n) = \sum_{m=-\infty}^{\infty} x(m)h(n-m)$$

- In frequency domain, this relation is simply given by a multiplication

$$Y(\Omega) = X(\Omega)H(\Omega)$$

- If the spectral transform is computed using e.g. the fast Fourier transform (FFT), computational complexity can be drastically reduced!

The Convolution Sum

Response of an LTI-System to Arbitrary Inputs

$$y(n) = T\{x(n)\} = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \quad (3)$$

- System response: $y(n) = T\{x(n)\}$
- Unit sequence response (impulse response): $h(n) \equiv y(n) = T\{\delta(n)\}$
- A relaxed LTI system is completely characterized by a single function $h(n)$, which is the response to the unit sample sequence $\delta(n)$
- The input-output relation Eq. (3) is called the **convolution sum**

The Convolution Sum

- The convolution sum $y(n_0) = \sum_{k=-\infty}^{\infty} x(k)h(n_0 - k)$ at time n_0 is obtained by
 1. *Folding.* Fold $h(k)$ about $k = 0$ to obtain $h(-k)$
 2. *Shifting.* Shift $h(-k)$ by n_0 (to the right if n_0 is positive) to obtain $h(n_0 - k)$
 3. *Multiplication.* Multiply $x(k)$ by $h(n_0 - k)$ to obtain $v_{n_0} \equiv x(k)h(n_0 - k)$
 4. *Summation.* Sum all the values of the product sequence v_{n_0} to obtain $y(n_0)$

Example

- The impulse response of an LTI system is

$$h(n) = \{1, 2, 1, -1\}$$

Determine the response of the system to the input signal

$$x(n) = \{1, 2, 3, 1\}$$

Example

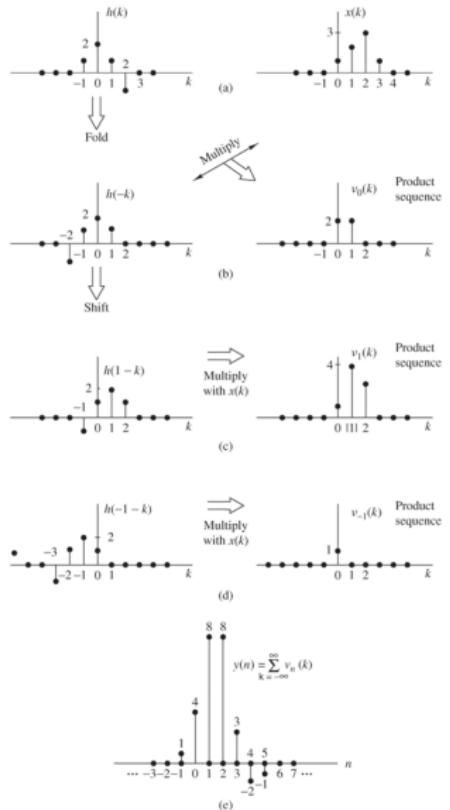


Figure 2.3.2 Graphical computation of convolution.

Notation The convolution is denoted by an asterisk “*”

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k)$$

Identity and shifting The unit sample sequence $\delta(n)$ is the identity element of convolution, i.e.

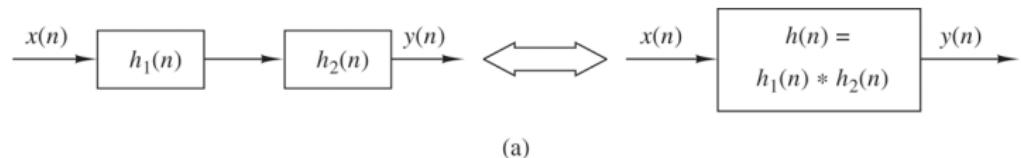
- $x(n) * \delta(n) = x(n),$
- $x(n) * \delta(n - k) = x(n - k)$

Commutative With a change of variable $k \leftarrow n - k$, we see that

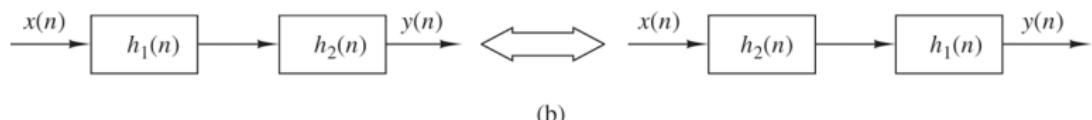
$$y(n) = x(n) * h(n) = h(n) * x(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k)$$

Associative

$$[x(n) * h_1(n)] * h_2(n) = x(n) * [h_1(n) * h_2(n)]$$



(a)



(b)

Figure 2.3.5 Implications of the associative (a) and the associative and commutative (b) properties of convolution.

Distributive

$$x(n) * [h_1(n) + h_2(n)] = x(n) * h_1(n) + x(n) * h_2(n)$$

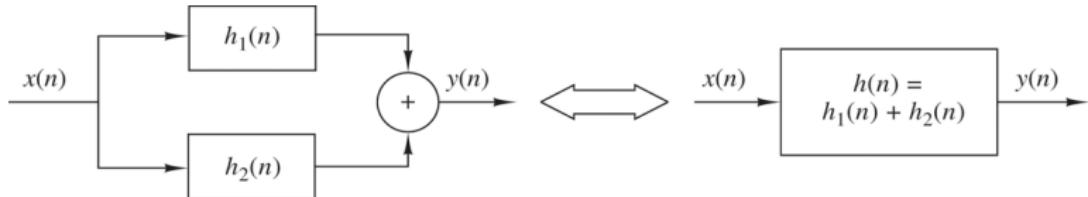
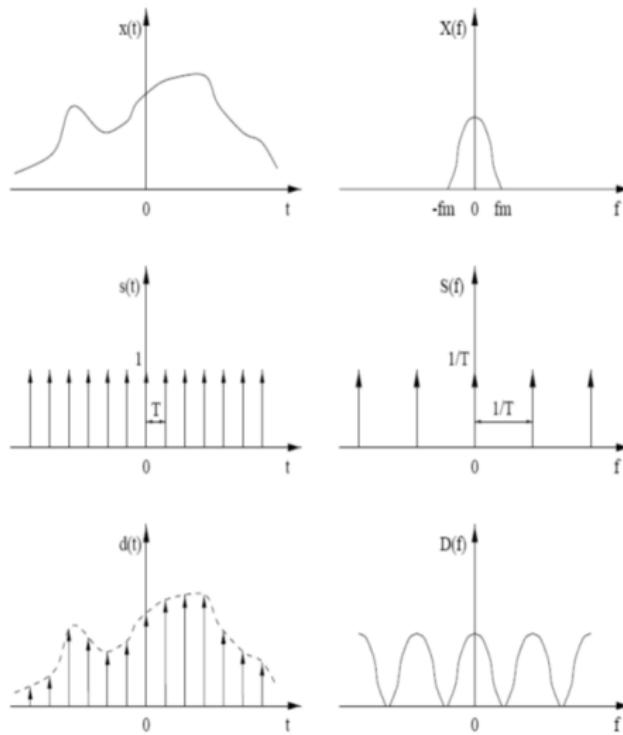


Figure 2.3.6 Interpretation of the distributive property of convolution: two LTI systems connected in parallel can be replaced by a single system with $h(n) = h_1(n) + h_2(n)$.

→Conversely, also means that: any LTI system can be decomposed into a parallel interconnection of subsystems

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
 - Fourier Theory and Complex Numbers
 - Linear Time-Invariant Systems
 - Discrete-time Fourier, DFT, and STFT
 - The z -Transform
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics



The Sampling Theorem

- From previous slide we see that after sampling the original spectrum (signal) can be reconstructed perfectly by eliminating the replicas
- If, however, the sampling period T_s is too large (the sampling frequency $f_s = 1/T_s$ is too low) the repeating spectra overlap \rightarrow perfect reconstruction not possible

Sampling Theorem

An analog signal with a bandwidth limited by f_m can be perfectly reconstructed from its samples if

$$f_s > 2f_m$$

- f_s : sampling rate
- f_m : maximum frequency in the signal

Practical Examples

- Humans can perceive sounds up to ≈ 20 kHz. What sampling rate would you recommend for audio?
- What is the sampling rate of an audio CD?
- Traditional telephone speech is sampled at 8 kHz. What is the highest audio frequency that we can perfectly represent?

Continuous-time Fourier transform

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t} dt \quad x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)e^{j\omega t} d\omega$$

where $X(j\omega)$ is a continuous and non-periodic function of $\omega = 2\pi f$.

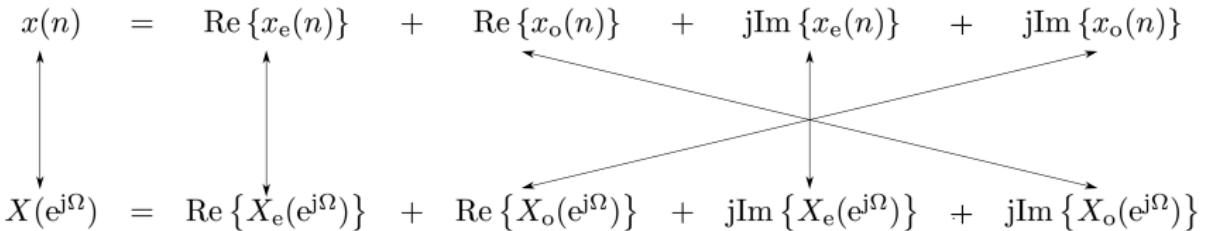
- Discrete-time signals $x(n)$:
 - $x(n) = x(nT)$, with the sampling period $T = 1/f_s$
 - $X_d(j\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega nT}$: continuous and periodic function of ω with period $2\pi/T$.
 - introduce $\Omega = \omega T, (0, 2\pi)$: normalized frequency

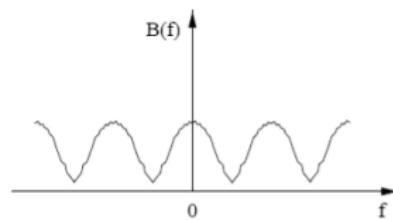
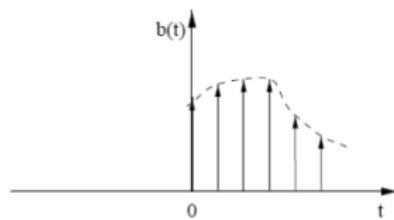
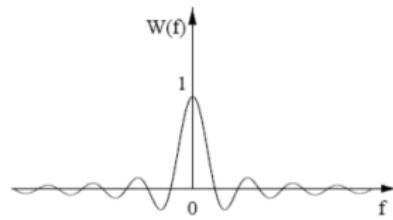
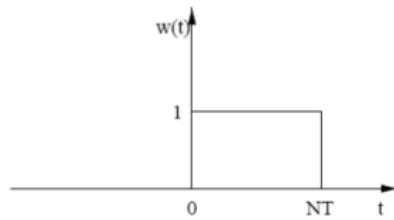
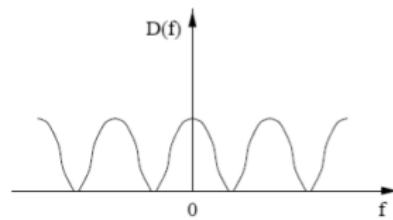
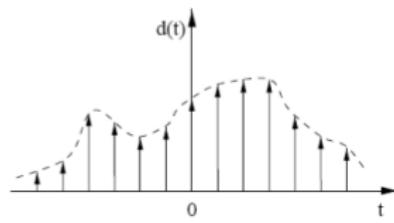
Discrete-time Fourier transform

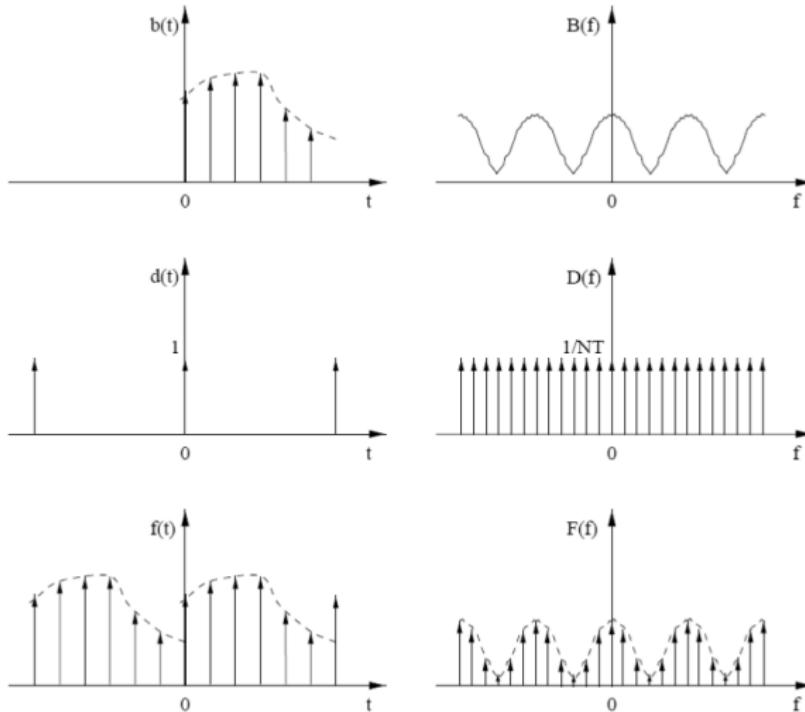
$$X(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j n \Omega} \quad x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega})e^{j n \Omega} d\Omega$$

Property	time domain	frequency domain
Definition	$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega}) e^{j\Omega n} d\Omega$	$X(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\Omega n}$
Linearity	$ax(n) + by(n)$	$aX(e^{j\Omega}) + bY(e^{j\Omega})$
Symmetry	real valued $x(n)$	$X(e^{j\Omega}) = X^*(e^{-j\Omega})$
even part	$x_e(n) = 0.5(x(n) + x(-n))$	$\text{Re}\{X(e^{j\Omega})\}$
odd part	$x_o(n) = 0.5(x(n) - x(-n))$	$j\text{Im}\{X(e^{j\Omega})\}$
Convolution	$x(n) * y(n)$	$X(e^{j\Omega}) Y(e^{j\Omega})$
Time Shift	$x(n - n_0)$	$e^{-j\Omega n_0} X(e^{j\Omega})$
Modulation	$x(n)e^{j\Omega_M n}$	$X(e^{j(\Omega - \Omega_M)})$
Parseval's theorem	$\sum_{n=-\infty}^{\infty} x(n)y^*(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\Omega}) Y^*(e^{j\Omega}) d\Omega$	

- Even part: $x_e(n) = 0.5(x(n) + x(-n))$
- Odd part: $x_o(n) = 0.5(x(n) - x(-n))$

$$\begin{aligned} x(n) &= \text{Re}\{x_e(n)\} + \text{Re}\{x_o(n)\} + j\text{Im}\{x_e(n)\} + j\text{Im}\{x_o(n)\} \\ X(e^{j\Omega}) &= \text{Re}\{X_e(e^{j\Omega})\} + \text{Re}\{X_o(e^{j\Omega})\} + j\text{Im}\{X_e(e^{j\Omega})\} + j\text{Im}\{X_o(e^{j\Omega})\} \end{aligned}$$






As we have seen

- Sampling in the time-domain \rightarrow periodic spectrum
- Sampling in the spectral domain \rightarrow periodic time-domain signal

We conclude

- Continuous time-domain signal $\circ\bullet$ continuous spectrum (Fourier Transform)
- Periodic time-domain signal $\circ\bullet$ discrete spectrum (Fourier series)
- Discrete time-domain signal $\circ\bullet$ periodic spectrum (Discrete-time Fourier transform (DTFT))
- Periodic and discrete time-domain signal $\circ\bullet$ periodic and discrete spectrum (Discrete Fourier transform (DFT))

Discrete Fourier Transform (DFT)

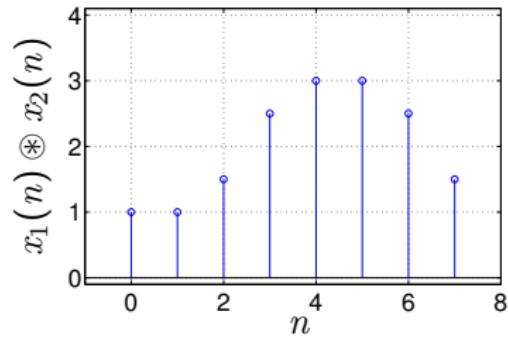
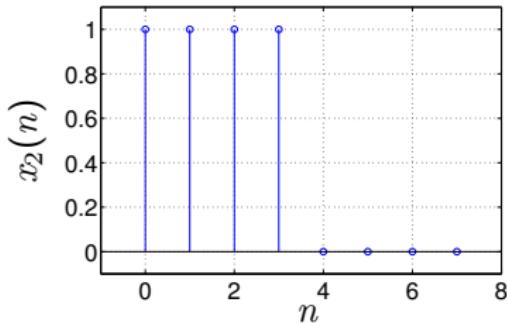
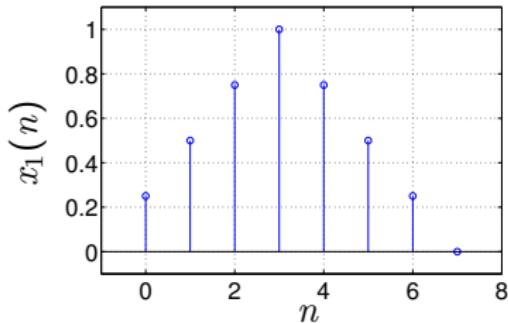
$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j \frac{2\pi}{N} kn} \quad x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j \frac{2\pi}{N} kn}$$

with $k = 0, \dots, N - 1$; $n = 0, \dots, N - 1$

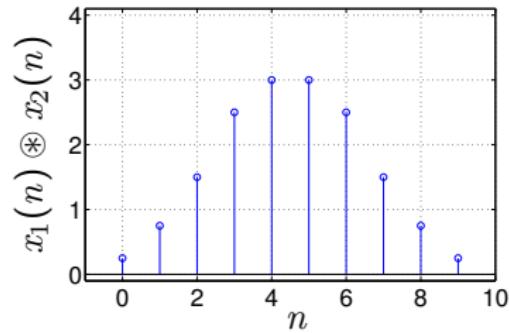
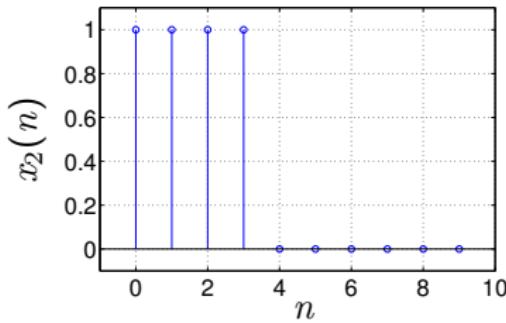
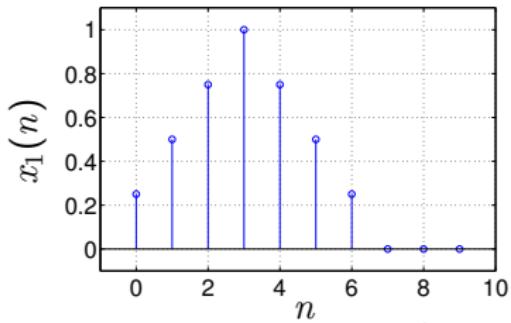
- The coefficients of the DFT are spaced by $\frac{2\pi}{N}$ on the normalized frequency axis.
- When the signal samples $x(n)$ are generated by sampling a continuous signal $x(t)$ with sampling rate f_s , the center frequencies of the DFT bins are $\Omega_k = \frac{2\pi k}{N}$, or $f_k = \frac{f_s}{N} k$.
- DFT coefficients $x(n)$ and $X(k)$ are periodic with period N

Property	time domain	frequency domain
Definition	$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j \frac{2\pi}{N} kn}$	$X_k = \sum_{n=0}^{N-1} x(n) e^{-j \frac{2\pi}{N} kn}$
Linearity	$ax(n) + by(n)$	$aX_k + bY_k$
Cyclic convolution	$x(n) \circledast y(n) = \sum_{\lambda=0}^{N-1} x(\lambda) y([n - \lambda]_{\text{mod } N})$	$X_k Y_k$
Multiplication	$x(n)y(n)$	$\frac{1}{N} \sum_{\lambda=0}^{N-1} X_\lambda Y_{[k-\lambda]_{\text{mod } N}}$
Parseval's theorem	$\sum_{n=0}^{N-1} x(n)y^*(n)$	$\frac{1}{N} \sum_{k=0}^{N-1} X_k Y_k^*$

- However, multiplying two DFT coefficient sets of length N results in a **cyclic convolution** with period N in the time domain.

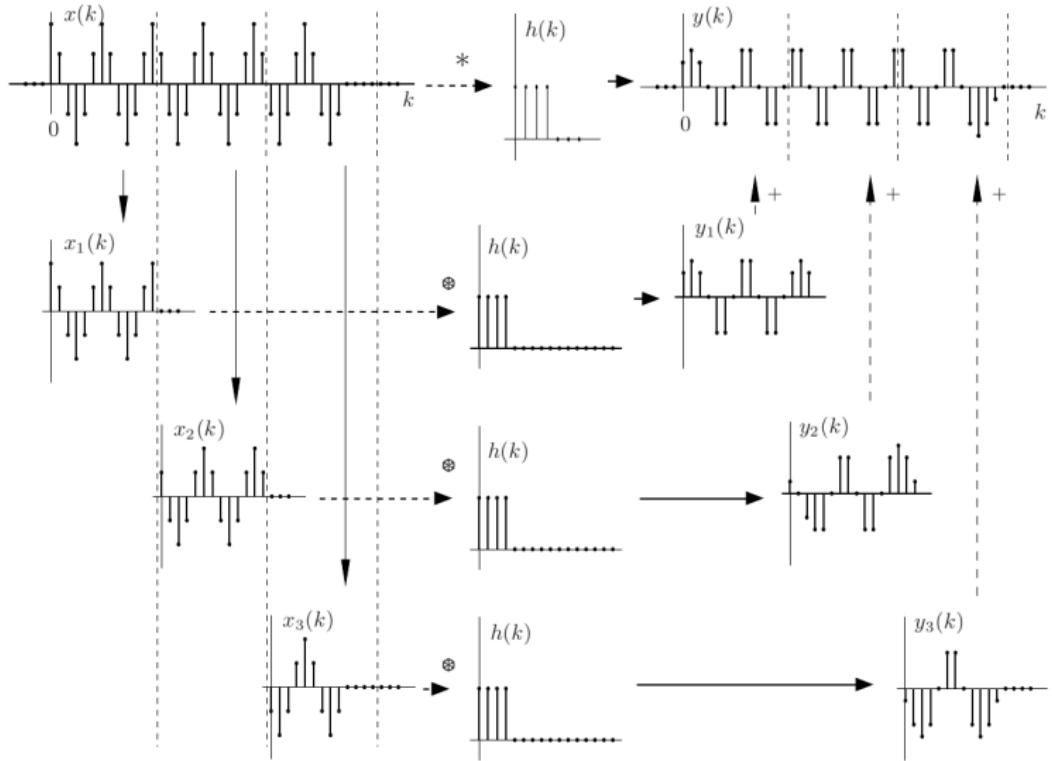


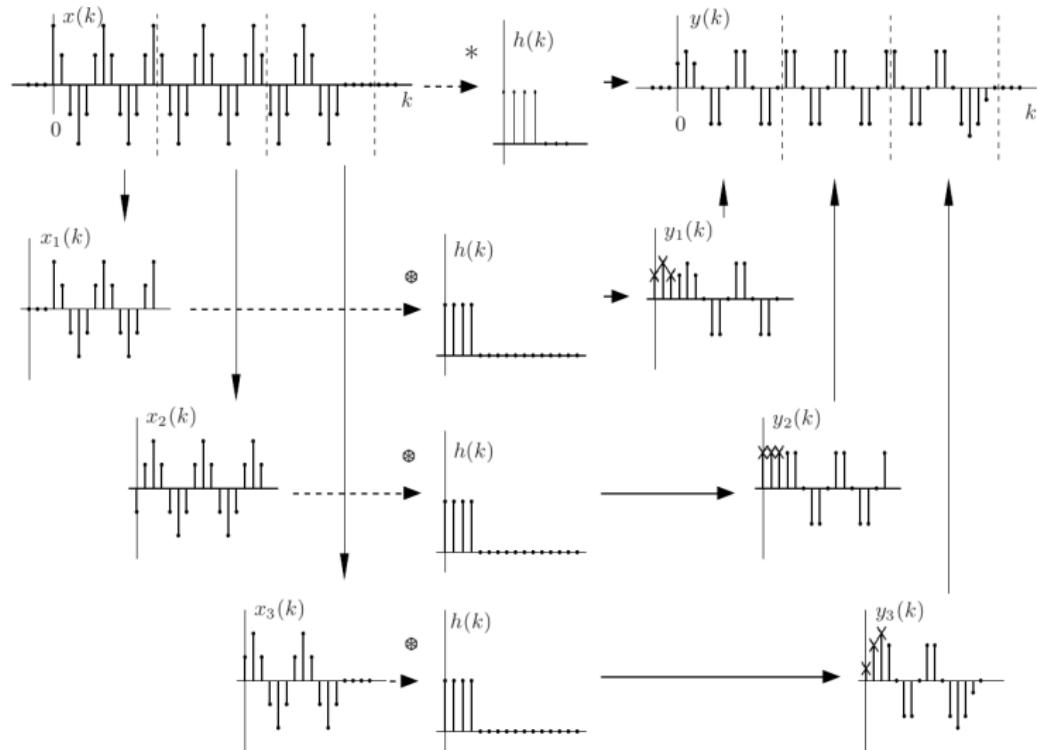
- Here N is chosen such that it equals the sum of nonzero entries minus one, i.e. $N = 7 + 4 - 1 = 10$.
- Cyclic convolution artifacts can be avoided by **zero-padding**!



Task: Compute linear filters in the DFT domain.

- Enforce linear convolution, thus avoid cyclic effects
- Partition signal into short segments
- Use zero-padding on input segments
- Use DFT and IDFT to compute linear convolution
- Use overlap-add or overlap-save procedure to construct output signal





Question

How does the frequency transform of a sinusoid look?

Question

How does the frequency transform of a sinusoid look?

Question

How does the frequency transform of a sinusoid look if a finite amount of data is available?

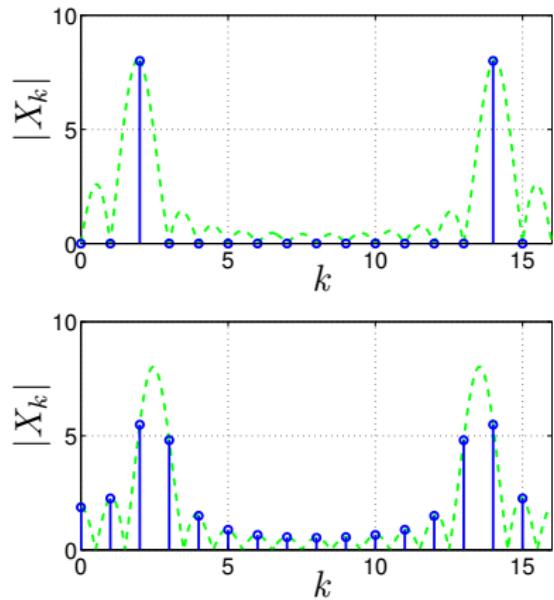
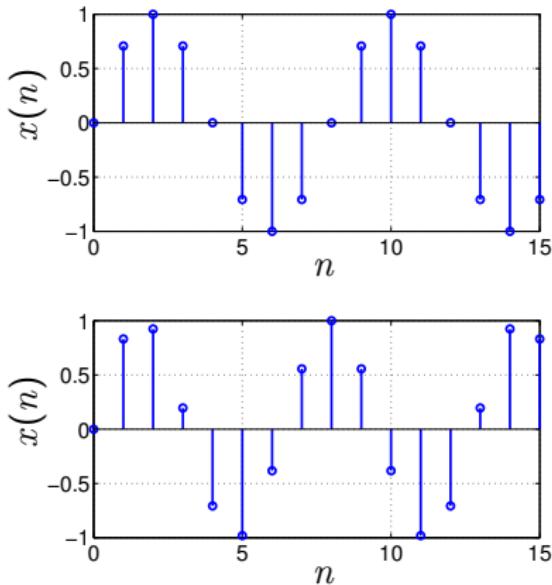
Question

How does the frequency transform of a sinusoid look?

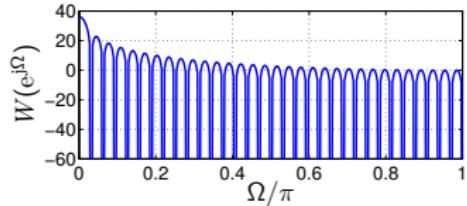
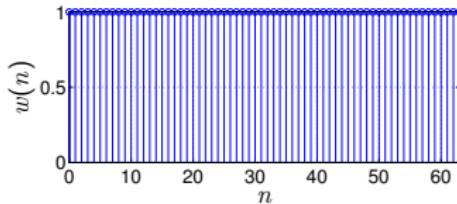
Question

How does the frequency transform of a sinusoid look if a finite amount of data is available?

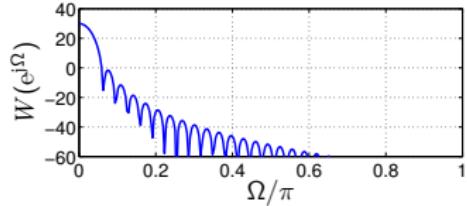
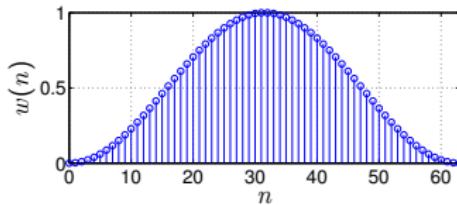
- Multiplication by windowing function turns into convolution of the corresponding spectra in the frequency domain.



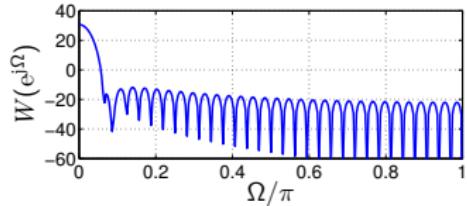
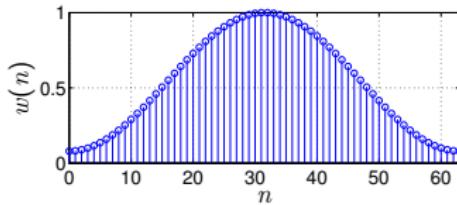
Rectangular window

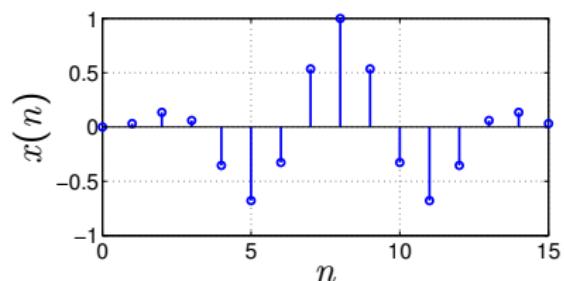
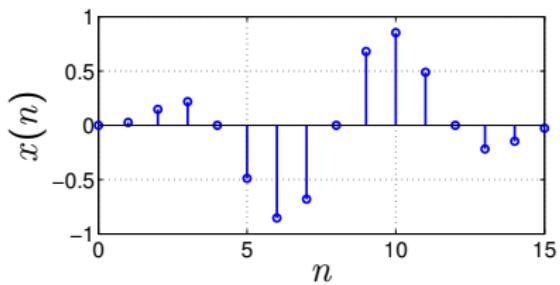


Hann window

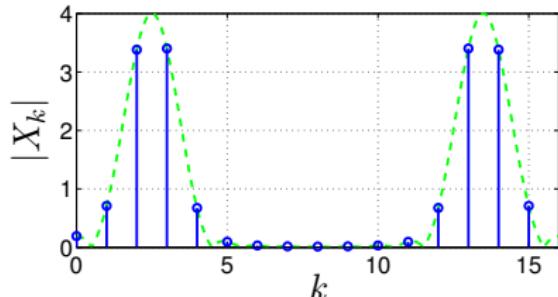
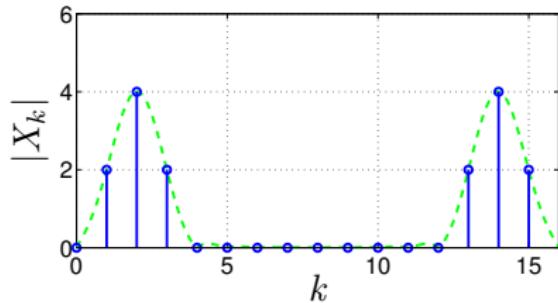


Hamming window



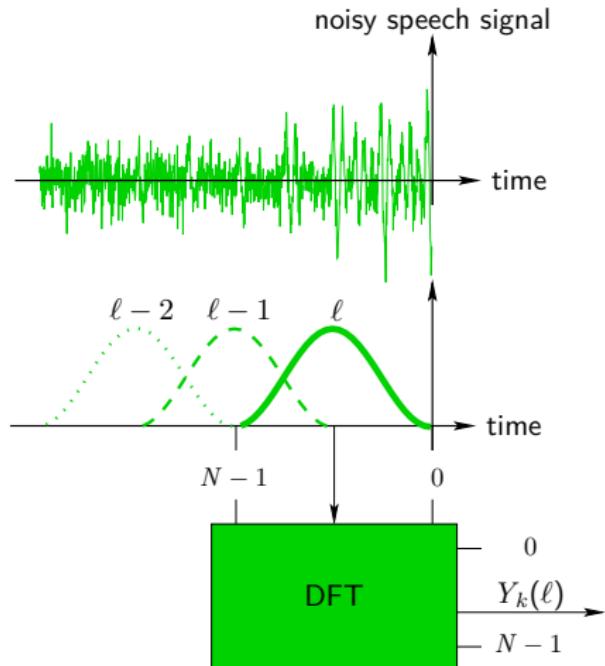


second one is modified using a hann-window



Conclusions

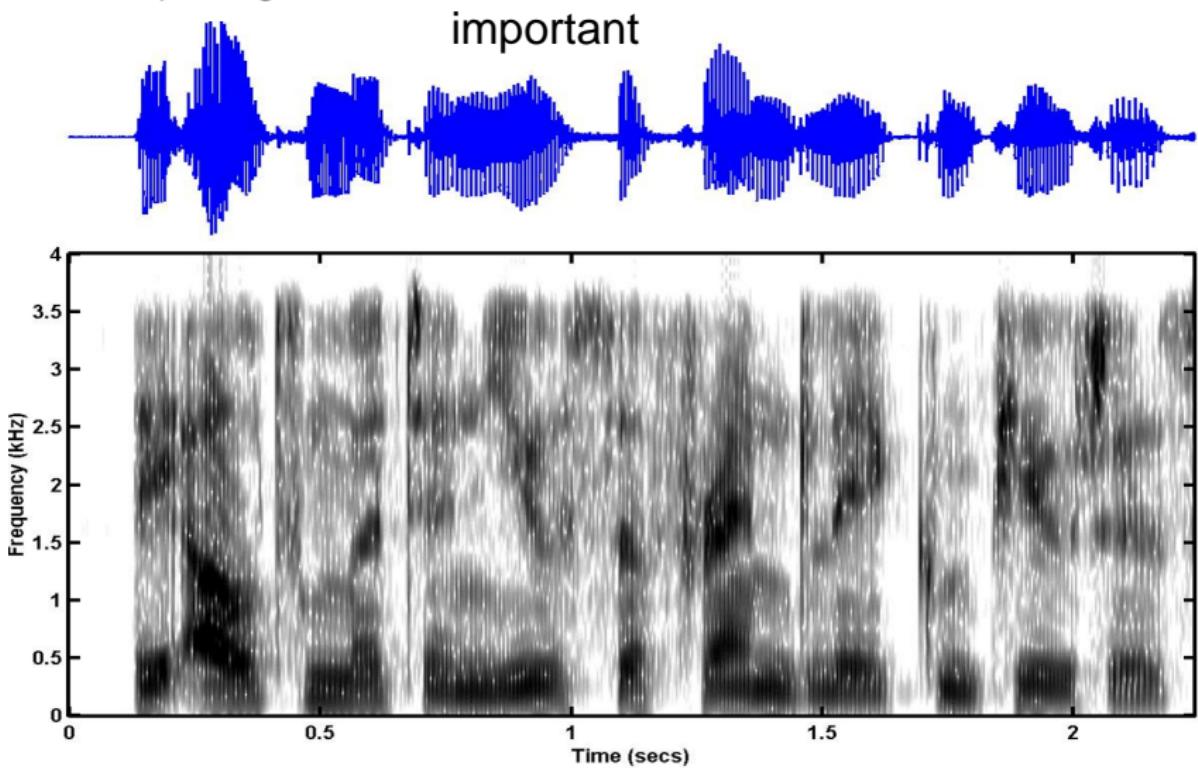
- The choice of the windowing function is a trade-off between spectral resolution and spectral leakage
- Rectangular (boxcar) window
 - Fourier transform is a Sinc-function
 - Narrow main-lobe → good frequency resolution
 - Large side-lobes → spectral leakage (bad)
- Tapered windows (Hann, Hamming, ...)
 - Main-lobe wider than for Rectangular window → slightly decreased frequency resolution
 - lower side-lobes → less spectral leakage (good)
 - Usually the preferred choice!



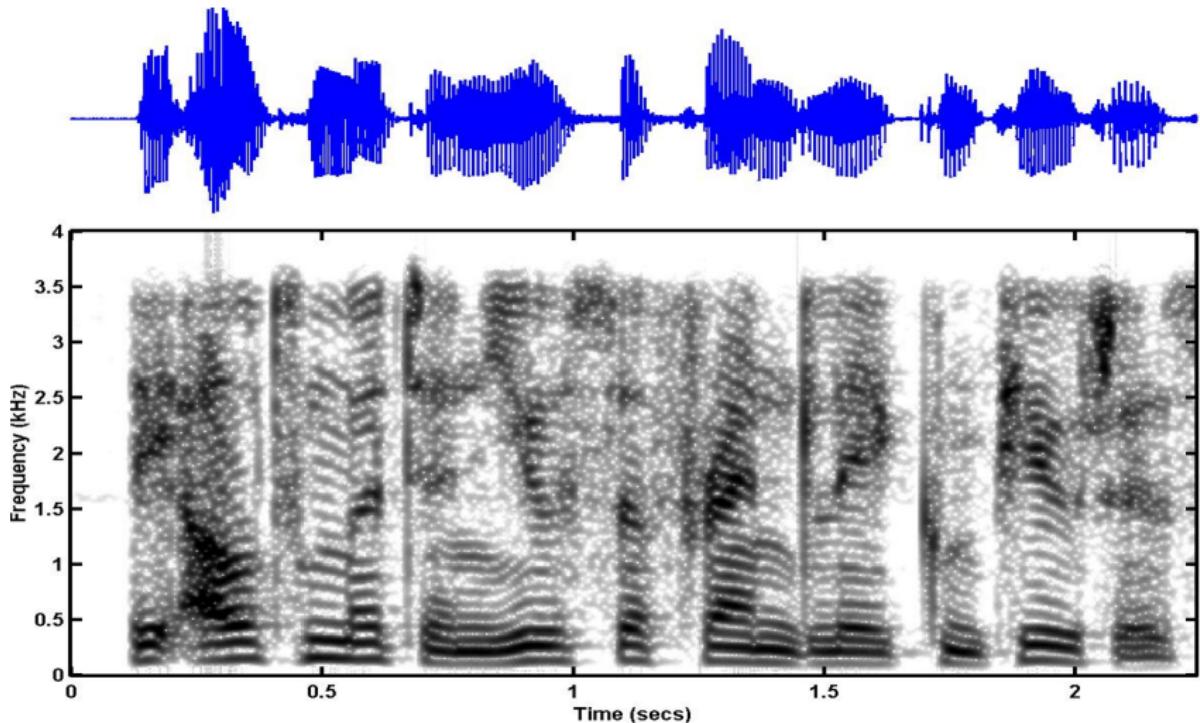
Time/Frequency Resolution

- STFT analysis always results in a trade-off between spectral resolution and temporal resolution
- Long spectral analysis windows (e.g. 32ms)
 - ✓ high spectral resolution
 - ✗ low temporal resolution
 - Narrowband spectrogram
- Short spectral analysis windows (e.g. 10ms)
 - ✗ low spectral resolution
 - ✓ high temporal resolution
 - Wideband spectrogram
- *Note: Adding zeros to the time-domain signal (aka zero-padding) merely interpolates the spectrum, but does not affect the resolution. In other words, zero-padding does not affect the ability to separate nearby sinusoids in the spectrum or nearby events in time-domain.*

Wideband Spectrogram



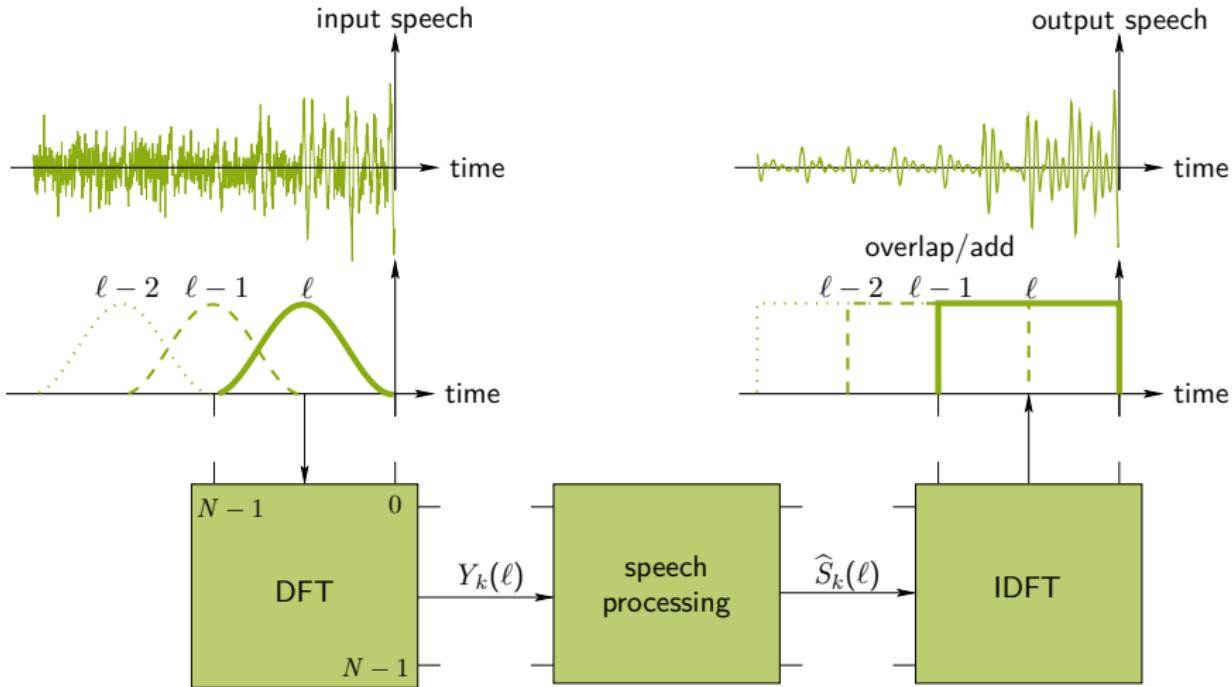
Narrowband Spectrogram



Wavesurfer demonstration

Quelle: <http://www.speech.kth.se/wavesurfer/>

- Different speech sounds in time and frequency domain,
- Different analysis lengths,
- Different windowing functions



1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
 - Fourier Theory and Complex Numbers
 - Linear Time-Invariant Systems
 - Discrete-time Fourier, DFT, and STFT
 - The z -Transform
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics

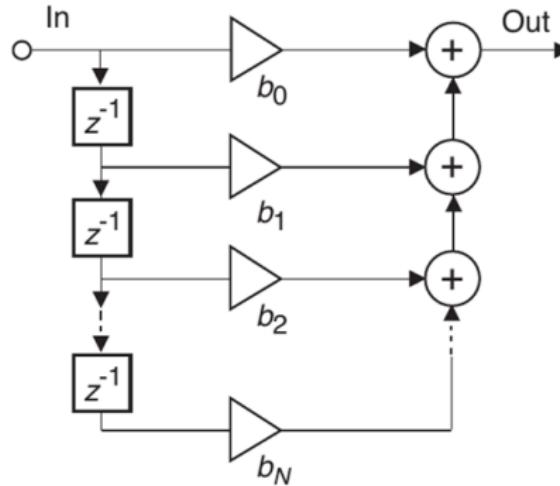
The z -transform of a discrete-time signal $x(n)$ is defined as the power series

$$X(z) \equiv \mathcal{Z}\{x(n)\} \equiv \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (4)$$

- with $z \in \mathbb{C}$ a complex-valued variable
- For $z = e^{j\Omega} \rightarrow$ discrete-time Fourier transform
- Well suited to analyze digital filters
 - Unit delay as a building block: $x(n - 1) \circledleftarrow z^{-1} X(z)$
 - Digital filtering can be expressed as a polynomial of z^{-1}

FIR Filter

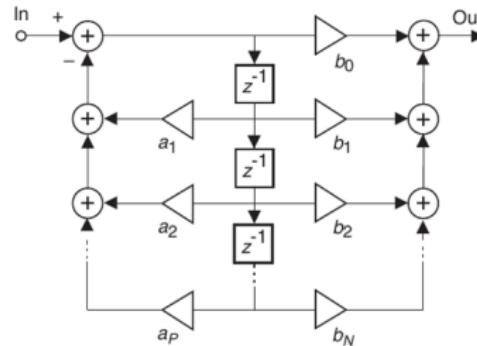
$$H_{\text{FIR}}(z) = \sum_{n=0}^{N-1} b_n z^{-n} = b_0 + b_1 z^{-1} + \cdots + b_{N-1} z^{-(N-1)}$$



IIR Filter

IIR = Infinite impulse response

$$H_{\text{IIR}}(z) = \frac{\sum_{n=0}^{N-1} b_n z^{-n}}{1 + \sum_{p=1}^{P-1} a_p z^{-p}} = \frac{b_0 + b_1 z^{-1} + \cdots + b_{N-1} z^{-(N-1)}}{1 + a_1 z^{-1} + \cdots + a_{P-1} z^{-(P-1)}}$$



1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
 - Tube model of the vocal tract
 - Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics
8. Source Separation

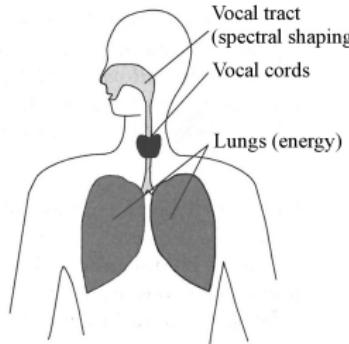


Universität Hamburg

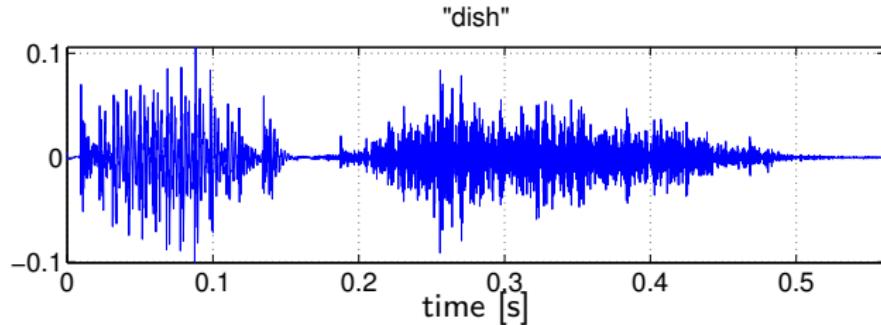
DER FORSCHUNG | DER LEHRE | DER BILDUNG

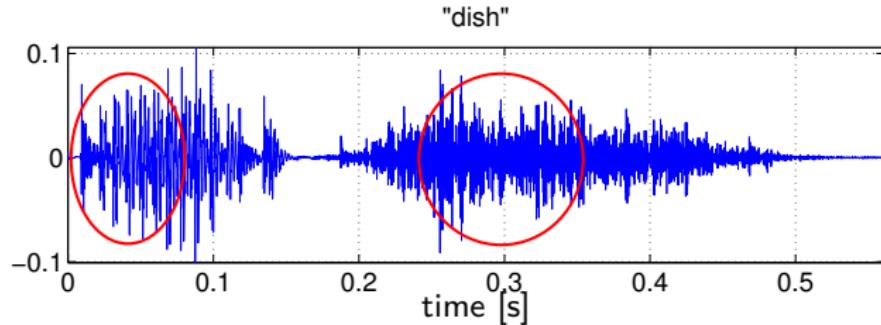


Vocal Tract Model and Linear Prediction

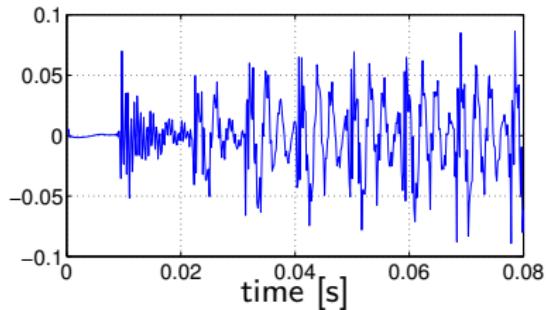


- Speech sounds:
 - voiced: periodic opening and closing of the vocal cords → fundamental period T_0 ,
 - unvoiced: open vocal cords, constriction somewhere in the vocal tract,
- Resonances of the vocal tract result in peaks in the spectral envelope of speech sounds (Formants).
- Different resonance frequencies result in different meanings of an utterance.

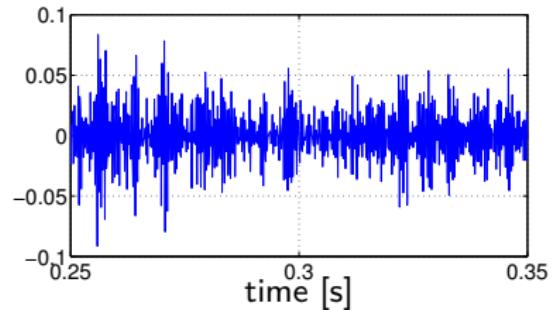




"di" of the word "dish"



"sh" of the word "dish"



Needed: parametric model for

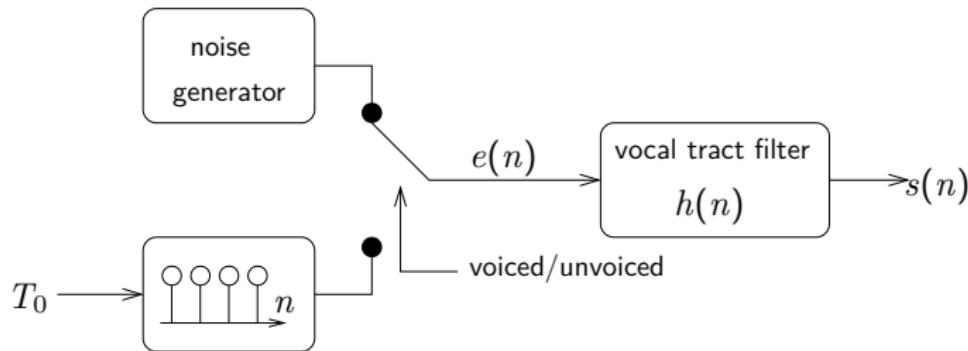
- the vocal tract filter function $h(n)$,
- the excitation signal $e(n)$.

Needed: parametric model for

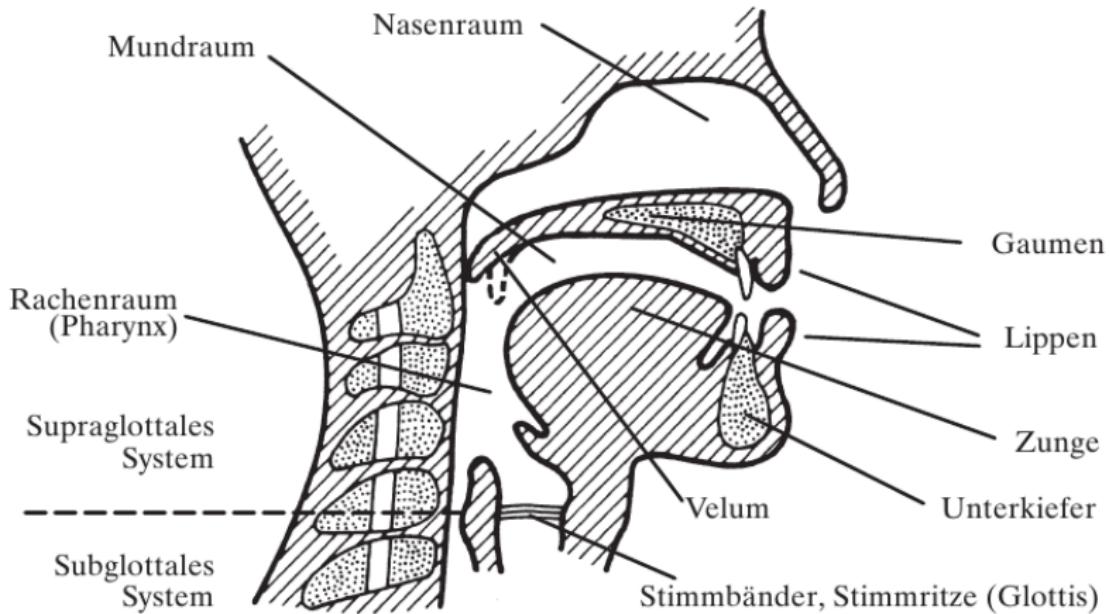
- the vocal tract filter function $h(n)$,
- the excitation signal $e(n)$.

Modeling the excitation $e(n)$

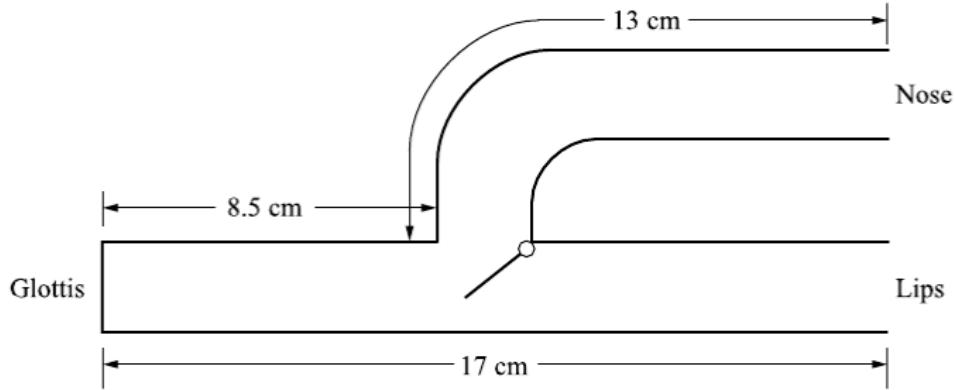
- Two state model
 - voiced: pulse train in time (periodic opening of the vocal cords).
(Parameter: fundamental period T_0 .)
 - unvoiced: excite with white Gaussian noise.



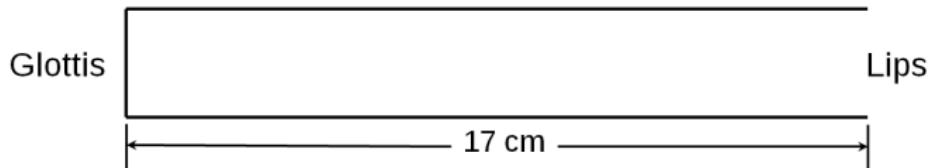
1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
 - Tube model of the vocal tract
 - Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics
8. Source Separation



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



© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

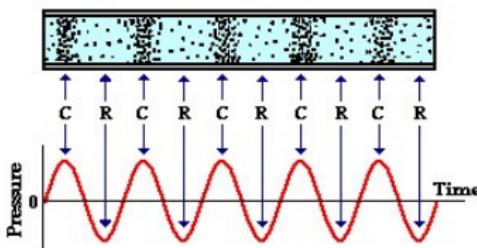


- For a closed velum, the model simplifies to a single tube
→ Nasals ("m", "n") are not well modeled.

- ↔ direction of oscillation
← direction of propagation



Sound is a Pressure Wave



NOTE: "C" stands for compression and "R" stands for rarefaction

- We define

- $p(x, t)$: acoustic pressure (*Schalldruck*)
- $u(x, t)$: sound particle velocity (*Schallschnelle*)
- $v(x, t) = u \cdot A$: volume velocity (*Schallfluss*)

- Acoustic impedance (compare Ohm's law)

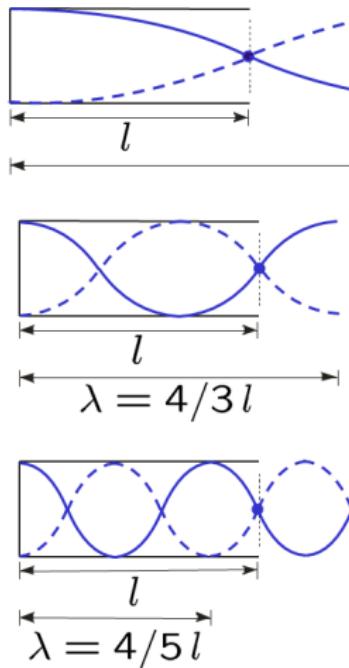
$$Z(x, t) = \frac{p(x, t)}{v(x, t)}$$

- Sudden changes in impedance result in reflections. Reflection factor

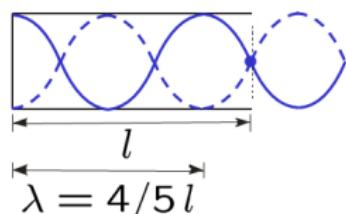
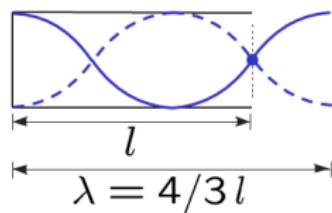
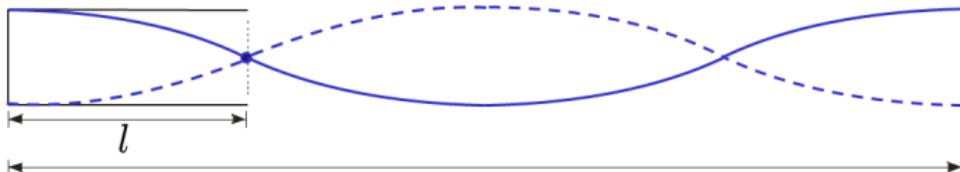
$$r_1 = \frac{Z_1 - Z_0}{Z_1 + Z_0} \quad \text{tube: } r_1 = \frac{A_0 - A_1}{A_0 + A_1}$$

- Special case: tube

- Closed ending of a tube (acoustically hard):
 $v(x, t) = 0$; $p(x, t)$ maximum; $Z_0(x, t) = \infty$; $r = -1$
- Open end of a tube (acoustically soft):
 $p(x, t) = 0$; $v(x, t)$ maximum; $Z_0(x, t) = 0$; $r = 1$



- The acoustic pressure is
 - Maximum at the closed end
 - zero at the open end.
- Standing waves occur at the specific resonance wavelengths $\lambda_n = \frac{4l}{2n+1}$



- $\lambda_n = \frac{4l}{2n+1}$
- Given a typical vocal tract length of 17 cm and the speed of sound $c = 340$ m/s we have

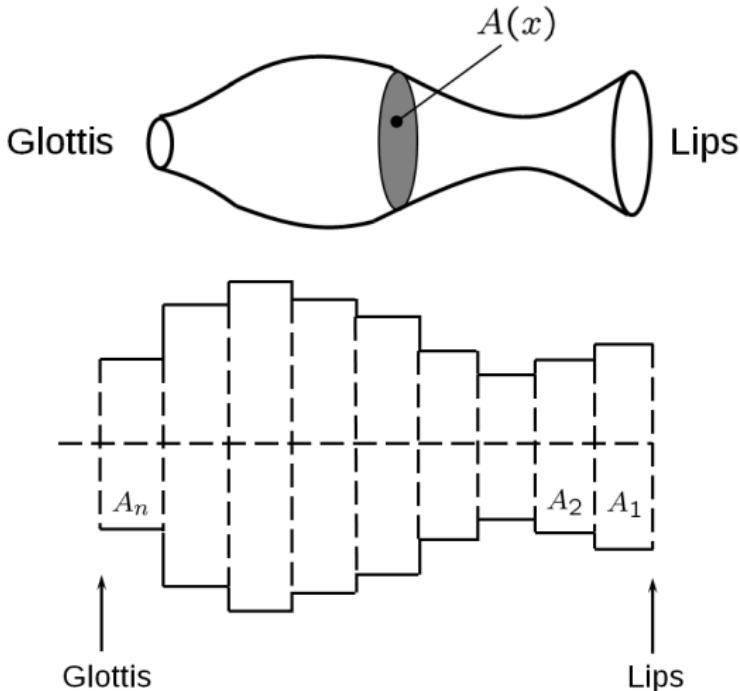
$$f_n = \frac{c}{\lambda} = (2n + 1)500 \text{ Hz} = \\ \{500 \text{ Hz}, 1500 \text{ Hz}, 2500 \text{ Hz}, 3500 \text{ Hz}\}$$

- Rule of thumb:
"One resonance per kHz".

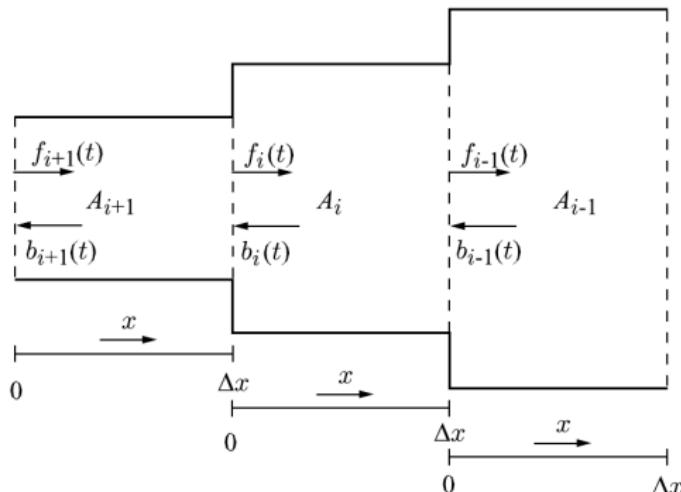
- From previous considerations we see that the first resonance occurs at one fourth of the wavelength
- The standard pitch of 440 Hz corresponds to a wavelength of

$$\lambda = \frac{340 \text{ m/s}}{440 \text{ Hz}} = 0.77 \text{ m}$$

- The organ pipe's length is $\lambda/4 \approx 20 \text{ cm}$.
- A 20 Hz-sinusoid has a wavelength of $\lambda = 17 \text{ m}$
- An Organ pipe with a resonance of 20 Hz requires a length of $\lambda/4 = 4.25 \text{ m}$.
- For an open pipe, the required length is twice as long, $\lambda/2 = 8.5 \text{ m}$.



Vary, Martin (2006), ©Wiley & Sons, Ltd

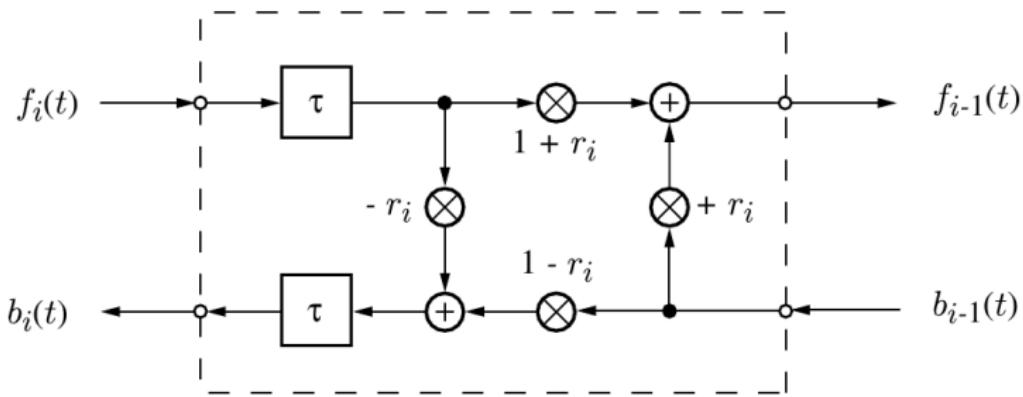


- Sudden changes in diameter/area result in reflections with reflection factor

$$-1 < r < 1$$

- Extreme cases
 - $r=1$: $A_{i-1} \rightarrow \infty$; open ending
 - $r=-1$: $A_{i-1} = 0$; closed ending
- forward $f(t)$ and backward $b(t)$ travelling waves occur.

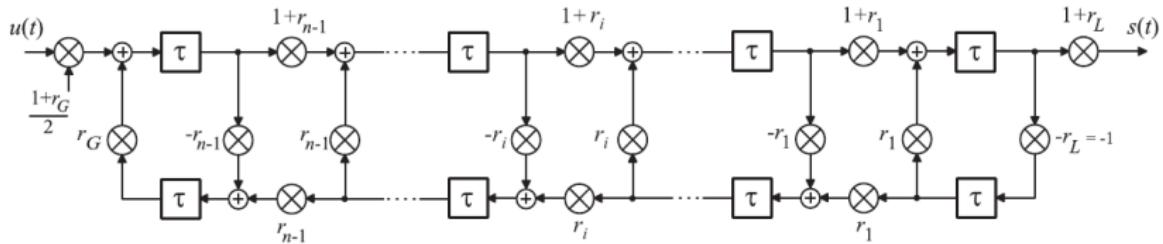
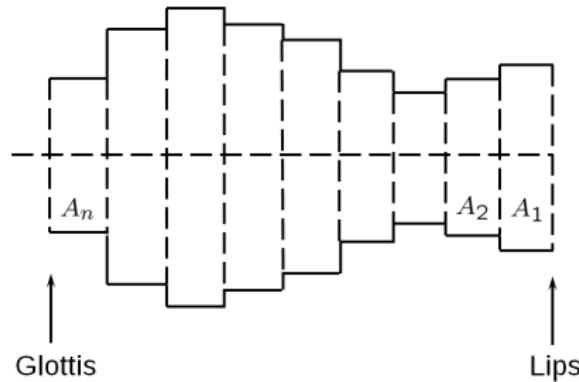
Quelle: Vary, Martin, Digital Speech Transmission, Wiley 2006

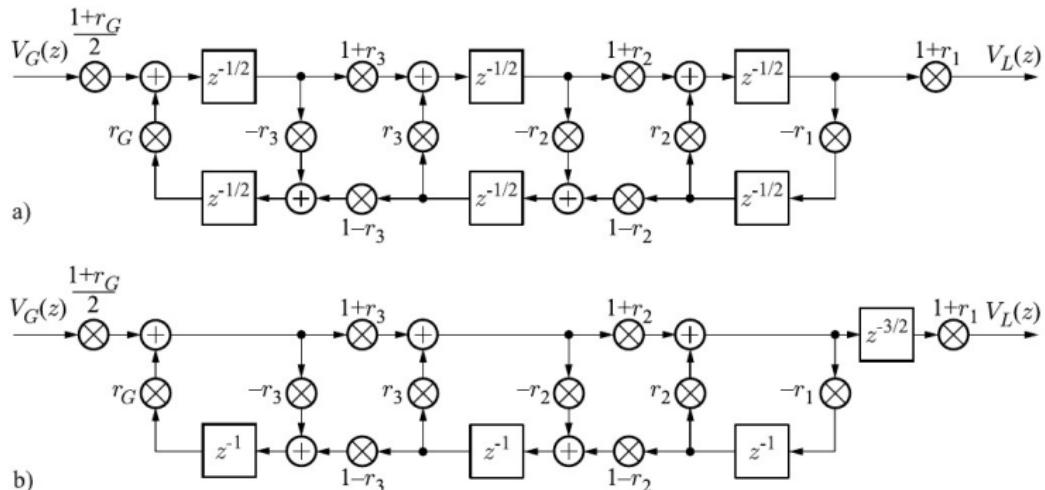


© 2006 John Wiley & Sons, Ltd
 Vary, Martin - Digital Speech Transmission

$$\begin{aligned}
 f_{i-1}(t) &= (1 + r_i) f_i(t - \tau) + r_i b_{i-1}(t) \\
 b_i(t) &= -r_i f_i(t - 2\tau) + (1 - r_i) b_{i-1}(t - \tau)
 \end{aligned}$$

n tube segments closed with glottis and lips

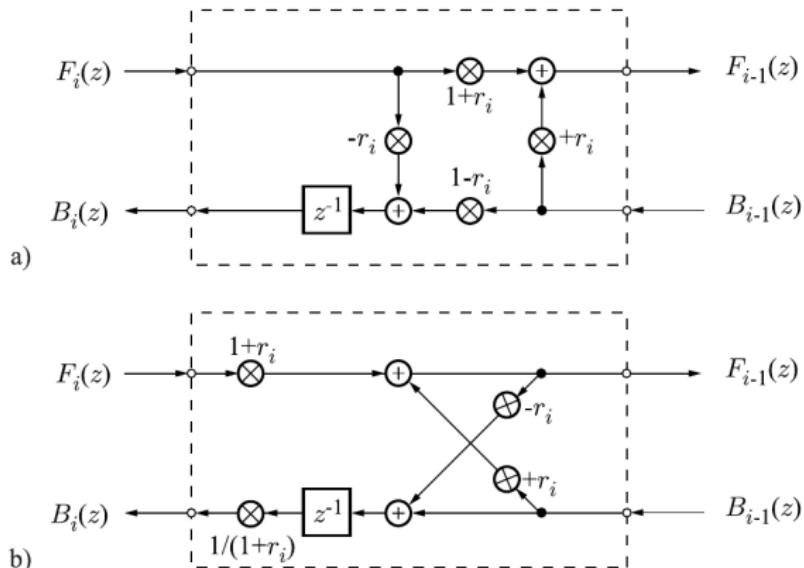




© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

Figure 2.12: Discrete-time models of the vocal tract (example $n = 3$)

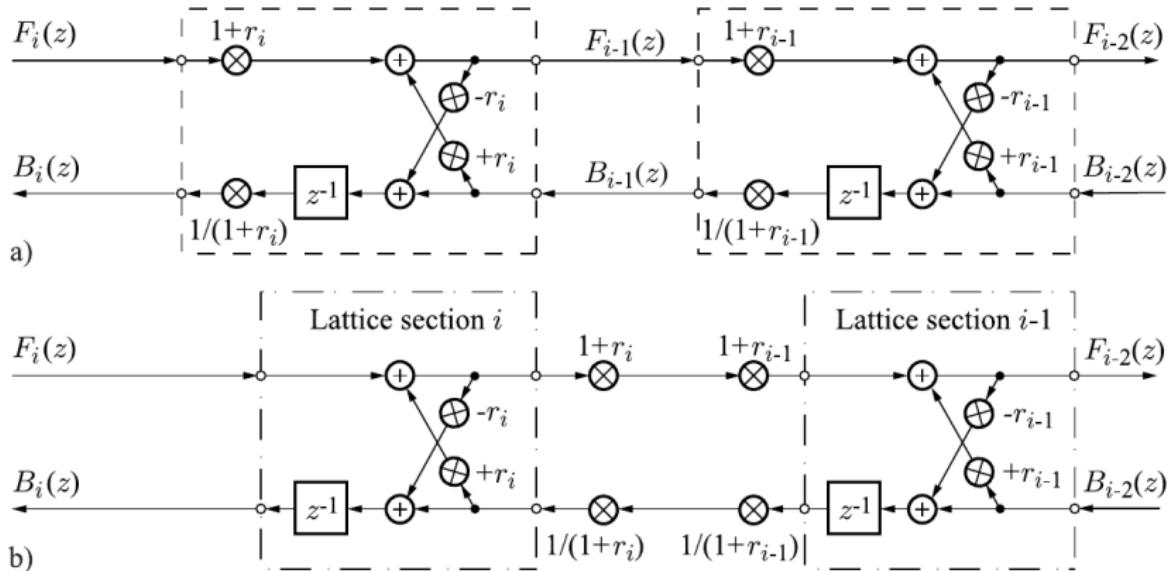
- a) Structure derived from the wave equations; $z^{-1/2} \equiv$ delay by τ
- b) Modified structure with delay elements $T = 2\tau \equiv z^{-1}$



© 2006 John Wiley & Sons, Ltd
 Vary, Martin - Digital Speech Transmission

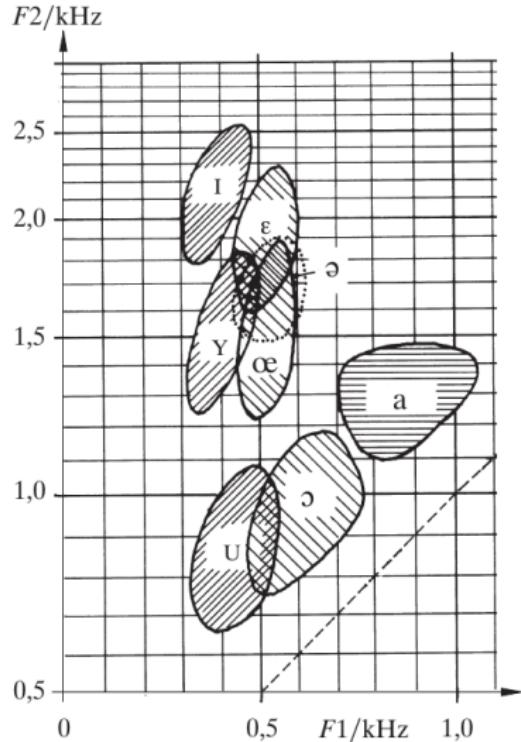
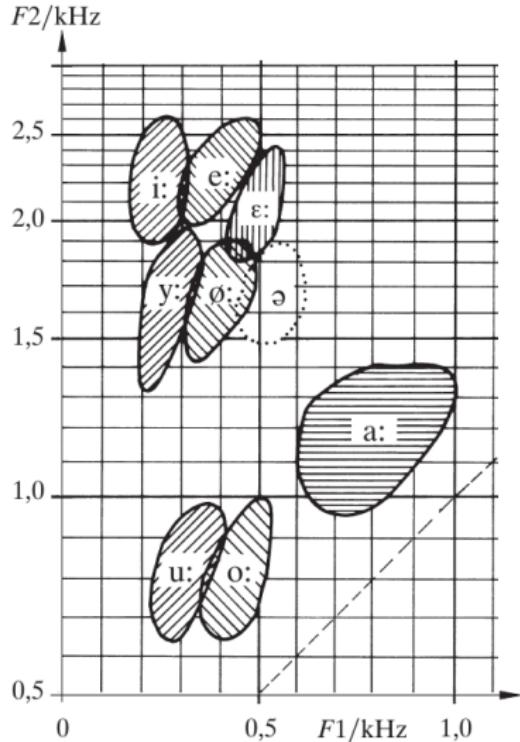
Figure 2.13: Equivalent filter sections

- a) Ladder structure
- b) Lattice structure



- In the lattice structure, two multiplications can be joined to one
→ higher efficiency

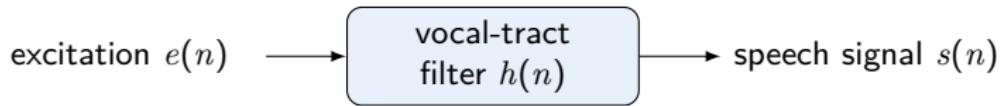
- In the source-filter model, excitation (source) and filter (vocal tract) are treated as being independent.
 - Formants: Peaks of the spectral envelope, resonances of the vocal tract
 - defines the meaning of a phone
 - Fundamental frequency: first peak of the spectral fine structure, and distance between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



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

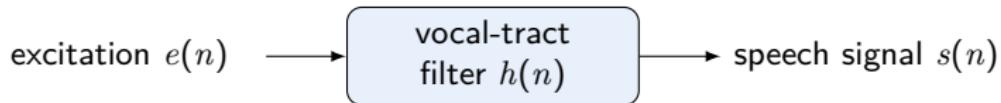
1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
 - Tube model of the vocal tract
 - Linear Prediction
5. Sampling, Quantization, and Speech Coding
6. Speech Enhancement
7. Machine Learning Basics
8. Source Separation

- The vocal tract is modeled using the impulse response $h(n)$.



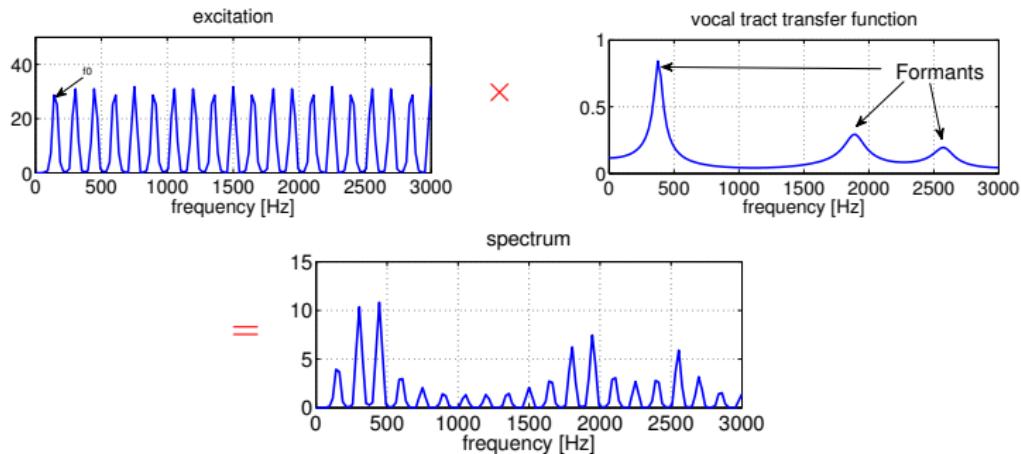
- $s(n) = e(n) * h(n)$ o—• $S(f) = E(f) \cdot H(f)$

- The vocal tract is modeled using the impulse response $h(n)$.



- $s(n) = e(n) * h(n)$ o—• $S(f) = E(f) \cdot H(f)$

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



- In discrete time-domain, we can describe the system by a convolution

$$s(n) = \sum_{m=0}^{\infty} h(m)e(n-m), \quad (5)$$

where the impulse response $h(k)$ can be infinitely long.

- We can describe our system by a *finite recursive* equation, as

$$s(n) = \sum_{m=0}^q b_m e(n-m) - \sum_{\nu=1}^p a_{\nu} s(n-\nu) \quad (6)$$



$$S(z) = E(z) \sum_{m=0}^q b_m z^{-m} - S(z) \sum_{\nu=1}^p a_{\nu} z^{-\nu} \quad (7)$$

- In discrete time-domain, we can describe the system by a convolution

$$s(n) = \sum_{m=0}^{\infty} h(m)e(n-m), \quad (5)$$

where the impulse response $h(k)$ can be infinitely long.

- We can describe our system by a *finite recursive* equation, as

$$s(n) = \underbrace{\sum_{m=0}^q b_m e(n-m)}_{\text{MA}} - \underbrace{\sum_{\nu=1}^p a_{\nu} s(n-\nu)}_{\text{AR}} \quad (6)$$



$$S(z) = E(z) \sum_{m=0}^q b_m z^{-m} - S(z) \sum_{\nu=1}^p a_{\nu} z^{-\nu} \quad (7)$$

- Referred to as *autoregressive moving-average (ARMA)* model.

- For the transfer function, we obtain

$$H(z) = \frac{S(z)}{E(z)} \stackrel{a_0=1}{=} \frac{\sum_{m=0}^q b_m z^{-m}}{\sum_{\nu=0}^p a_{\nu} z^{-\nu}} \quad (8)$$

- The transfer function thus consists of two polynomials.
- In general, every polynomial can be described by its roots. Thus, we obtain

$$H(z) = z^{p-q} b_0 \frac{\prod_{m=1}^q (z - z_{0m})}{\prod_{\nu=1}^p (z - z_{\infty\nu})} \quad (9)$$

- z_{0m} : roots of the numerator polynomial, zeros of $H(z)$,
- $z_{\infty\nu}$: roots of the denominator polynomial, poles of $H(z)$.
- ARMA-model \leftrightarrow pole-zero filter

- Exact modeling by pole-zero (or ARMA) model:

$$H(z) = b_0 z^{p-q} \frac{\prod_{m=1}^q (z - z_{0m})}{\prod_{\nu=1}^p (z - z_{\infty\nu})} = \frac{\sum_{m=0}^q b_m z^{-m}}{\sum_{\nu=0}^p a_\nu z^{-\nu}} \quad (10)$$

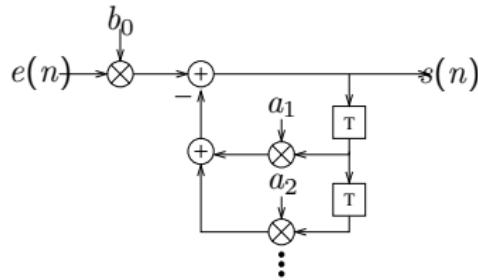
- We approximate the vocal tract filter as an autoregressive (AR) filter, also referred to as all-pole filter:

$$H(z) \approx b_0 z^p \frac{1}{\prod_{\nu=1}^p (z - z_{\infty\nu})} = \frac{b_0}{\sum_{\nu=0}^p a_\nu z^{-\nu}}. \quad (11)$$

- Accurate representation of $|H(z)|$, but not of the phase $\angle H(z)$. However, the ear is rather insensitive to phase changes.
- Poles correspond to resonances of the vocal tract (formants)

the ARMA signal model is given by

$$s(n) = \sum_{m=0}^q b_m e(n-m) - \sum_{\nu=1}^p a_\nu s(n-\nu)$$



The AR speech production model is given as

$$s(n) = b_0 \underbrace{e(n)}_{\text{innovation}} - \underbrace{\sum_{\nu=1}^p a_\nu s(n-\nu)}_{\text{past}}$$

■

$$s(n) = b_0 \underbrace{e(n)}_{\text{innovation}} - \underbrace{\sum_{\nu=1}^p a_\nu s(n-\nu)}_{\text{past}} \quad (12)$$

- According to this autoregressive speech model, successive speech samples are correlated.
- *Predict* the current speech sample from the previous samples

$$\hat{s}(n) = - \sum_{\nu=1}^p \hat{a}_\nu s(n-\nu) \quad (13)$$

- Thus, for $\hat{a}_\nu = a_\nu$ we can predict the speech signal up to the scaled excitation

$$\begin{aligned} d(n) &= s(n) - \hat{s}(n) \\ &= b_0 e(n). \end{aligned} \quad (14)$$

- Minimize power of prediction error $E(d^2(n))$ with respect to \hat{a}_ν , where $E(\cdot)$ is expectation (average over all possible d).
- Set first derivative to zero

$$\begin{aligned}
 0 &\stackrel{!}{=} \frac{\partial E(d^2)}{\partial \hat{a}_\nu} = E\left(2d(n) \frac{\partial}{\partial \hat{a}_\nu} \left(s(n) + \sum_{k=1}^p \hat{a}_k s(n-k)\right)\right) \\
 &= E(2d(n)s(n-\nu)) \\
 &= 2E\left\{\underbrace{\left(s(n) + \sum_{\mu=1}^p \hat{a}_\mu s(n-\mu)\right)}_{d(n)} s(n-\nu)\right\} \\
 \downarrow \quad &\varphi_s(\nu) = E(s(n)s(n-\nu)) \quad (\text{autocorrelation}) \\
 &= \varphi_s(\nu) + \sum_{\mu=1}^p \hat{a}_\mu \varphi_s(\nu - \mu).
 \end{aligned}$$

- Second derivative: $\frac{\partial^2 E(d^2)}{\partial \hat{a}_\nu^2} = E(2s^2(n - \nu)) \geq 0 \rightarrow \text{minimum.}$

Wiener-Hopf equations

$$\underbrace{\begin{pmatrix} \varphi_s(1) \\ \varphi_s(2) \\ \vdots \\ \varphi_s(p) \end{pmatrix}}_{\boldsymbol{\varphi}_s} = - \underbrace{\begin{pmatrix} \varphi_s(0) & \varphi_s(-1) & \dots & \varphi_s(1-p) \\ \varphi_s(1) & \varphi_s(0) & \dots & \varphi_s(2-p) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_s(p-1) & \varphi_s(p-2) & \dots & \varphi_s(0) \end{pmatrix}}_{\mathbf{R}_s} \underbrace{\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix}}_{\hat{\mathbf{a}}}$$

- $\hat{\mathbf{a}}_{\text{opt}} = -\mathbf{R}_s^{-1} \boldsymbol{\varphi}_s$
- In practice $\varphi_s(\nu)$ is estimated from short speech segments, and $E(\cdot)$ is replaced by a sum

$$\hat{\varphi}_s(\nu) = \sum_{n=n_1}^{n_2} \tilde{s}(n)\tilde{s}(n+\nu) \quad (15)$$

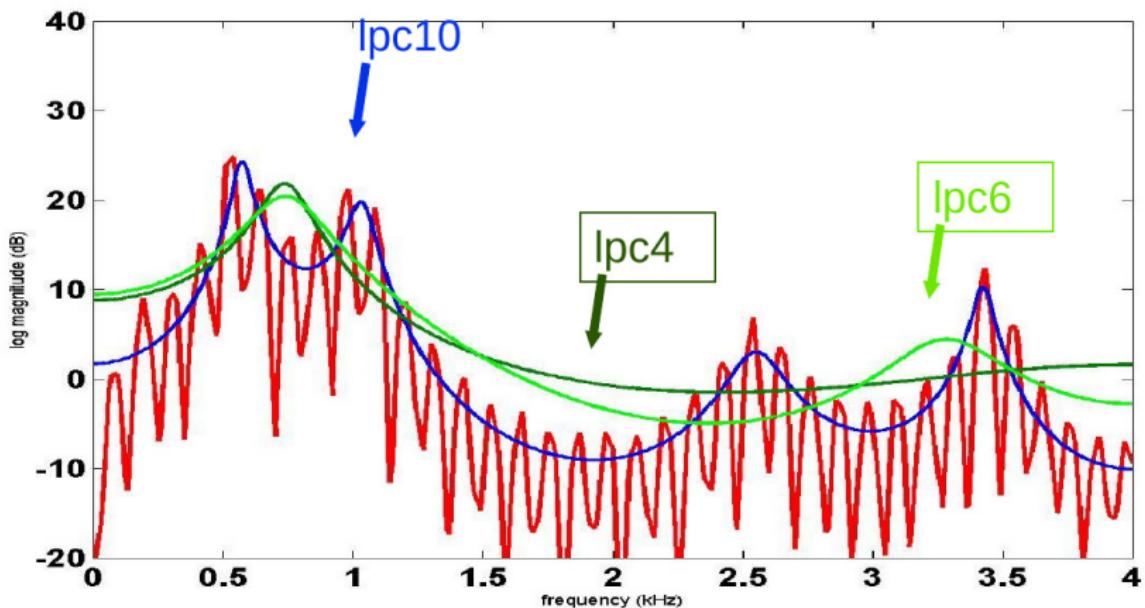
- $\hat{\varphi}_s(\nu) = \hat{\varphi}_s(-\nu)$,
- the correlation matrix \mathbf{R}_s is symmetric and Toeplitz,
- fast solutions exist (Levinson-Durbin recursion).
 - Levinson-Durbin yields both AR-coefficients and reflection coefficients at the same time.

- AR coefficients a_v , and reflection coefficients of the tube-model r_i both capture the same information,
- both are computed in the Levinson-Durbin recursion
- Reflection parameters of the tube-model can be computed from the signal itself.
- The relation between tube-segment *areas* is another representation

$$r_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}$$

Practical Considerations

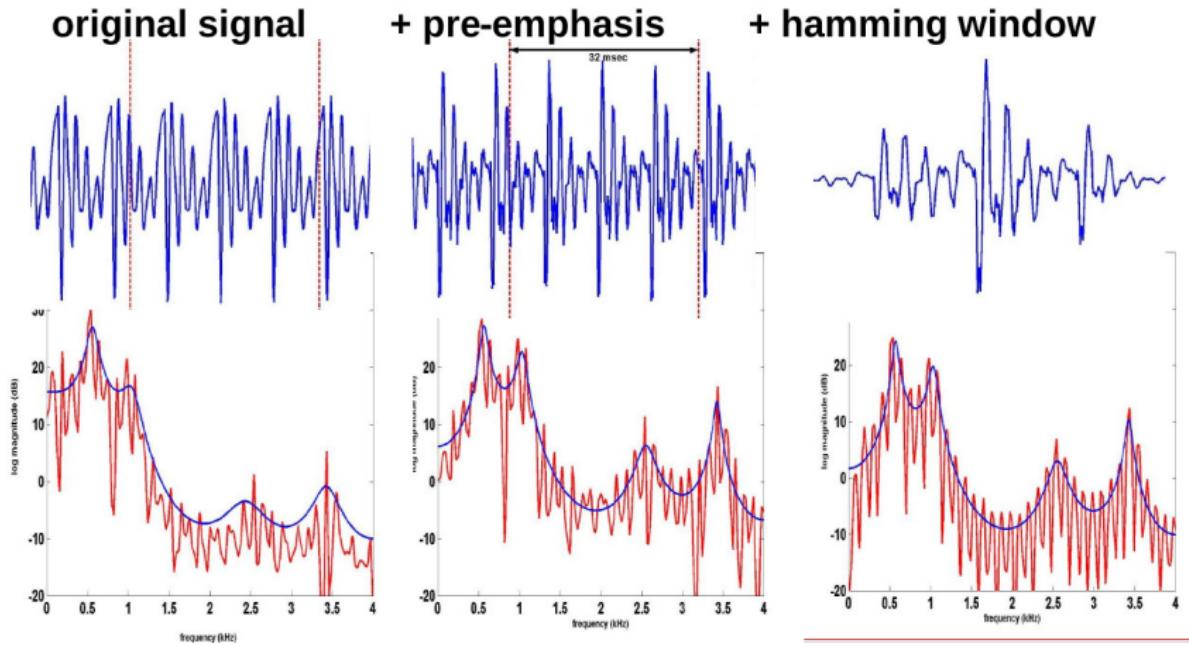
- | | |
|---------------------|---|
| Filter Order | <ul style="list-style-type: none">■ We need to model roughly 1 resonance per kHz and need 2 filter coefficients per resonance■ rule of thumb: $p = \text{filter order} = \text{sampling frequency} + 2$ |
| Segments | <ul style="list-style-type: none">■ Segment length over which we assume stationarity: 8-32 msec■ segment shift: 4-16 msec (e.g. 3/4 or 1/2 overlap) |
| Other | <ul style="list-style-type: none">■ Preemphasis■ Windowing: e.g. Hamming window, Hann window |



- The speech spectrum has a natural slope of roughly -6dB/octave
 - typical for many acoustic signals
 - note, that a flat spectrum (white signal) gives a 'high frequency' impression
- For LPC-analysis thus a first order pre-emphasis is applied to account for that slope

$$\widetilde{s(n)} = s(n) - \alpha s(n-1)$$

- typical parameter: $\alpha = 0.9\dots 1$
- After processing the pre-emphasis can be inverted.



- **Speech Analysis**

- Estimation of the spectral envelope
- Pitch tracking
- Formant tracking

- **Speech Coding**

- Basis for many speech codecs (LPC, CELP, ...)

- **Speech Synthesis**

- Independent control over source and filter parameters

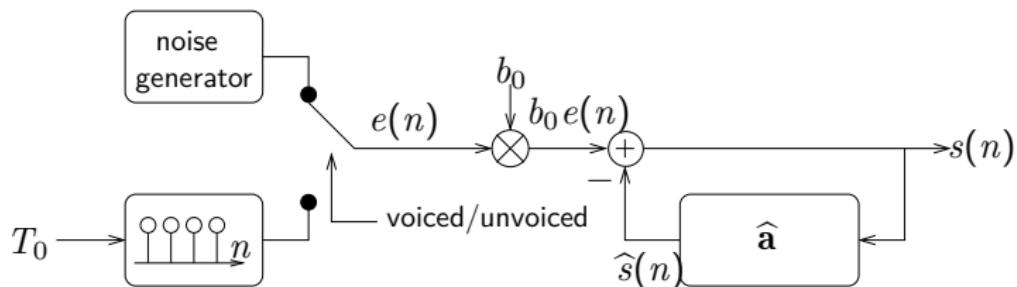
- **Speech Recognition**

- Estimation of the spectral envelope

- **Artifical Bandwidth Extensions**

- Separate treatment of source and filter parameters

- With the resulting model, intelligible speech can be transmitted with bit rates as low as 2.4 kbit/s.
- per speech segment we need to store/transmit
 - information if voiced or unvoiced
 - fundamental period T_0 (if voiced)
 - scale parameter b_0
 - 10 AR coefficients to model the vocal tract filter



Sound example

- only excitation $b_0e(n)$:
- only unvoiced:
- | V/UV:
- Coded speech (6.5 kbit/s):

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

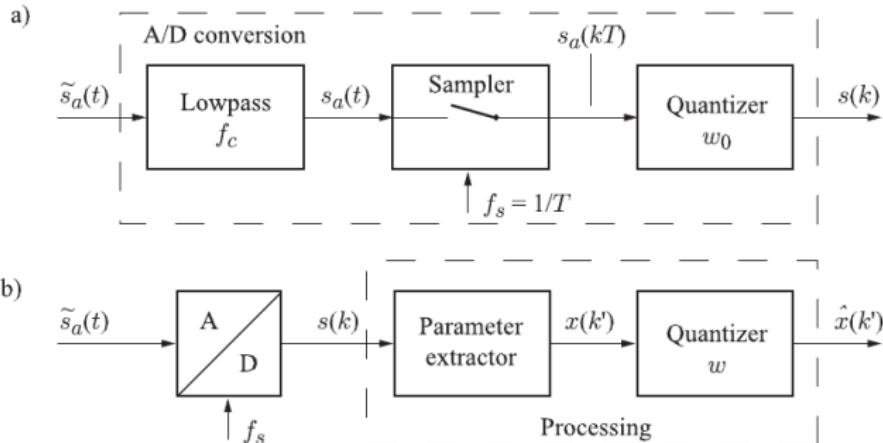


Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

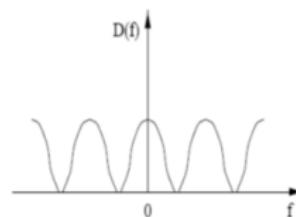
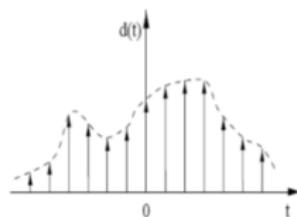
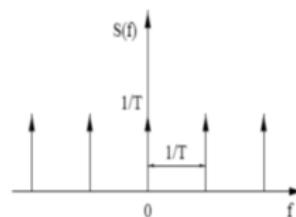
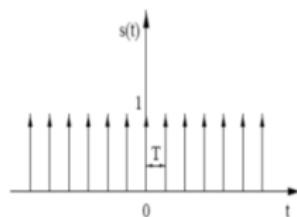
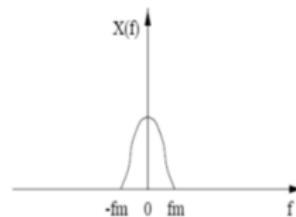
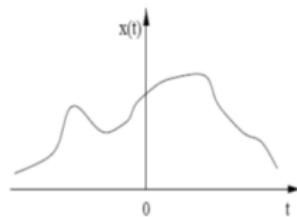


Sampling, Quantization, and Speech Coding



© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

- a) Quantization of signal samples
- b) Quantization of parameters



- From previous slide we see that if the sampling period T is chosen too large, the spectra (which are spaced by $1/T$) will overlap
- If the spectra overlap, a simple reconstruction using a lowpass-filter is not possible

Sampling Theorem

A signal can be perfectly reconstructed from its samples, if the sampling rate $f_s = 1/T$ is larger than two times the largest frequency f_m in the signal

$$f_s > 2f_m$$

- For audio signals we usually have following sampling rates

- From previous slide we see that if the sampling period T is chosen too large, the spectra (which are spaced by $1/T$) will overlap
- If the spectra overlap, a simple reconstruction using a lowpass-filter is not possible

Sampling Theorem

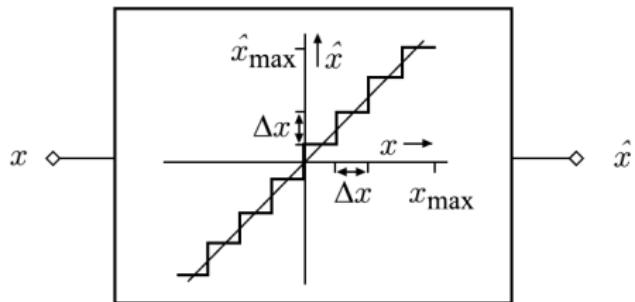
A signal can be perfectly reconstructed from its samples, if the sampling rate $f_s = 1/T$ is larger than two times the largest frequency f_m in the signal

$$f_s > 2f_m$$

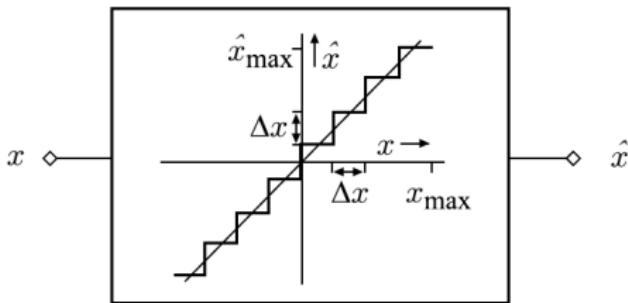
- For audio signals we usually have following sampling rates
 - 44.1 kHz or 48 kHz for Music
 - 8 kHz (ISDN / GSM) / 16 kHz (HD voice) / 32 kHz (HD Voice+) for telephony

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

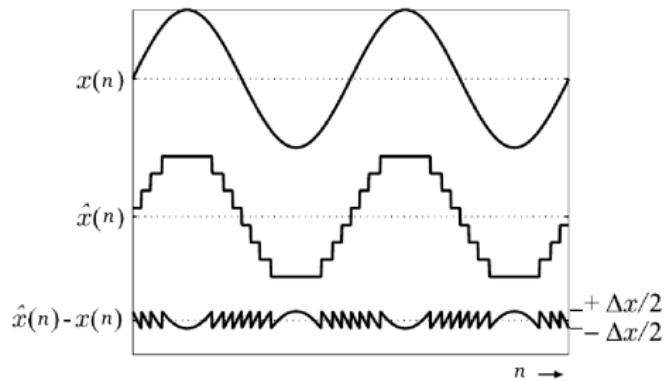
Midrise-Characteristic



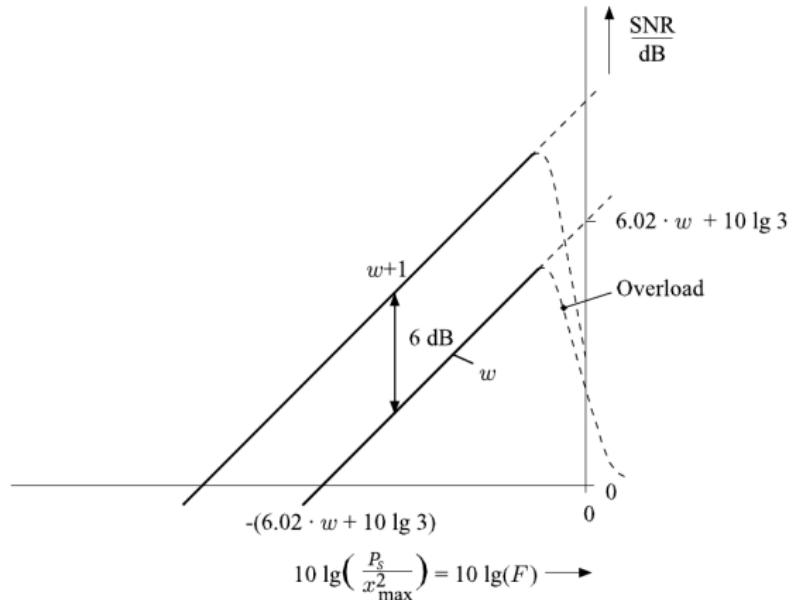
Midrise-Characteristic



Example: Sinusoid



©2006 Wiley & Sons, Vary, Martin

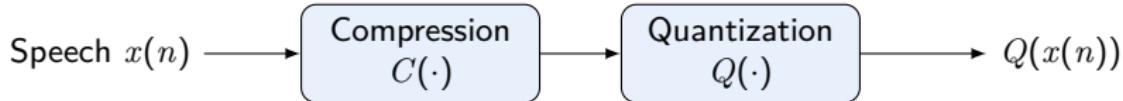


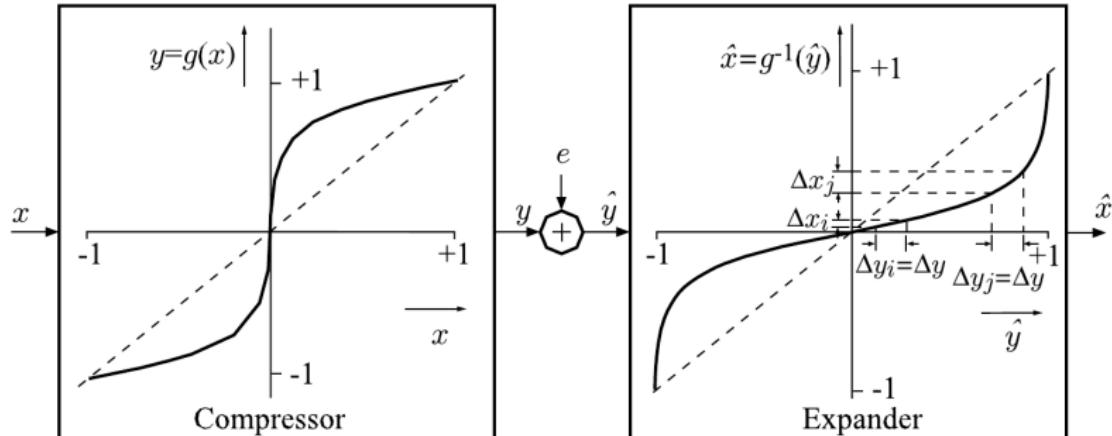
©1998 B.G. Teubner, Vary, Heute, Hess

- Rule of thumb: SNR increases by 6 dB for each spent bit w
- SNR is also dependent on the scale and distribution of the signal

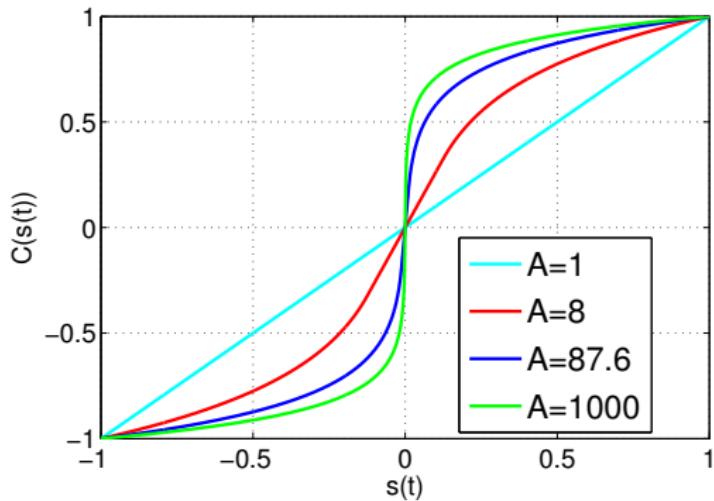
1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

- Motivation
 - Speech sample values have a highly 'zero-centered' distribution
 - Distortion to large values less audible than distortion to small values
 - A finer quantization of lower amplitude values would increase SNR.
- Approach
 - Compress speech signal with pseudo-logarithmic law
 - Quantize the compressed signal uniformly
- Logarithmic compression law
 - 8 bit logarithmic equivalent to 12 bit uniform quantization
 - 2 implementations are common
 - A-Law: Europe
 - μ -Law : North America and Japan





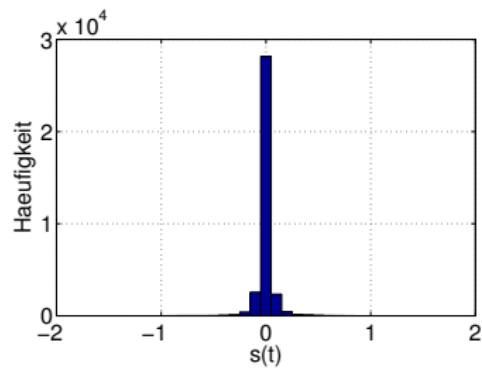
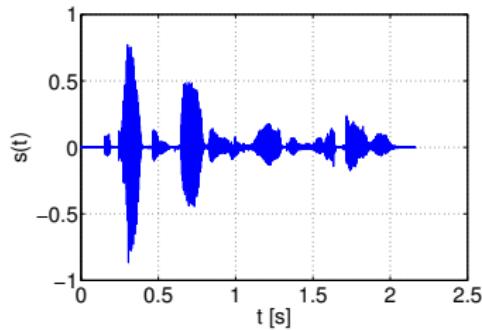
© 1998 B.G. Teubner
Vary, Heute, Hess - Digitale Sprachsignalverarbeitung



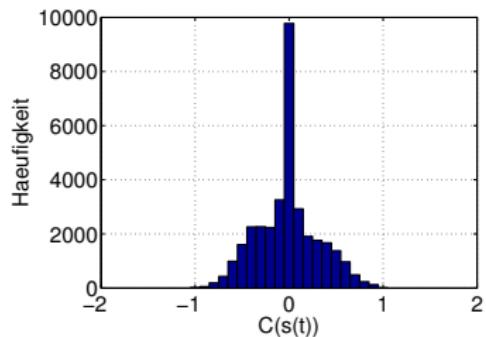
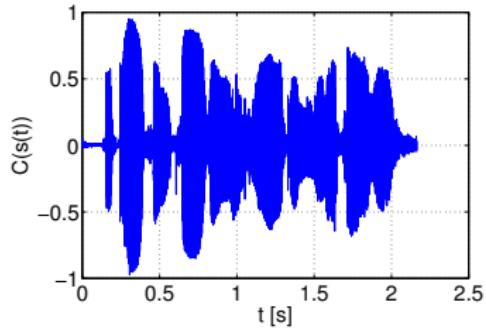
$$C(x) = \begin{cases} \text{sign}(x) \frac{1+\log(A|x|)}{1+\log(A)} & \frac{1}{A} \leq |x| \leq 1 \\ \text{sign}(x) \frac{A|x|}{1+\log(A)} & |x| < \frac{1}{A} \end{cases}$$

- 87.6 is a typical value for the A-law

Speech in time-domain



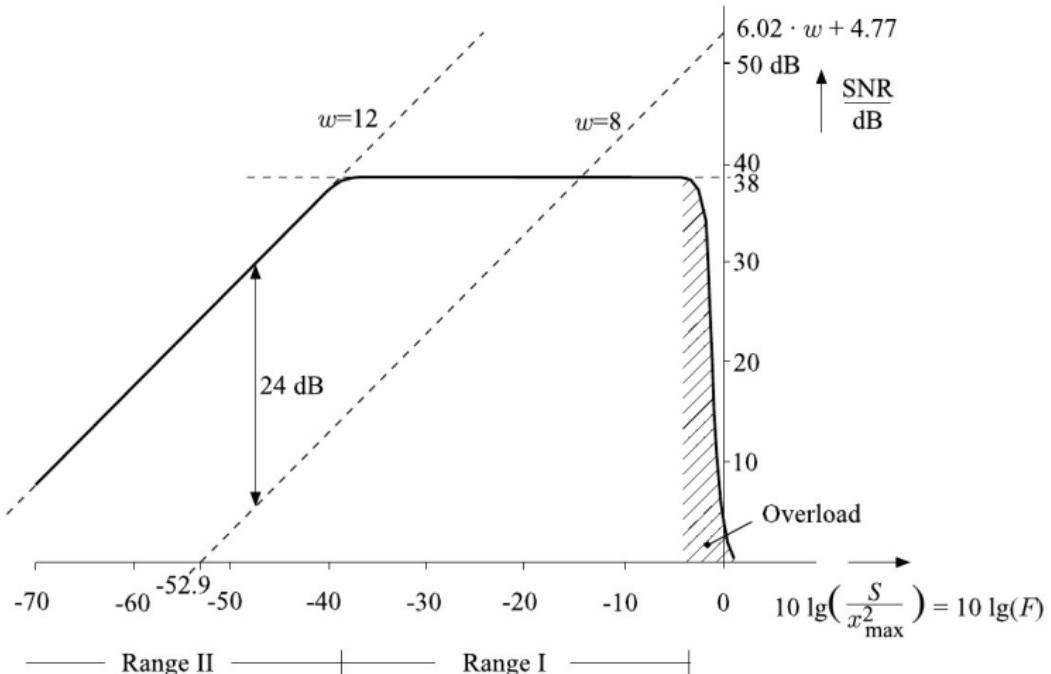
after compression



- low maximum gain ($-0.2 < s(t) < 0.2$)
 - Speech 
 - uniform Quantization (8bit) 
 - A-law, uniform Quantization (8bit) 

- low maximum gain ($-0.2 < s(t) < 0.2$)
 - Speech ➤
 - uniform Quantization (8bit) ➤
 - A-law, uniform Quantization (8bit) ➤

- high maximum gain ($-0.9 < s(t) < 0.9$)
 - Speech ➤
 - uniform Quantization (8bit) ➤
 - A-law, uniform Quantization (8bit) ➤



© 1998 B.G. Teubner
Vary, Heute, Hess - Digitale Sprachsignalverarbeitung

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

- Adaptive adjustment of the step size Δx with respect to the standard deviation

$$\Delta x(n) = c\hat{\sigma}_x(n)$$

- Adaptive quantization forward (AQF)

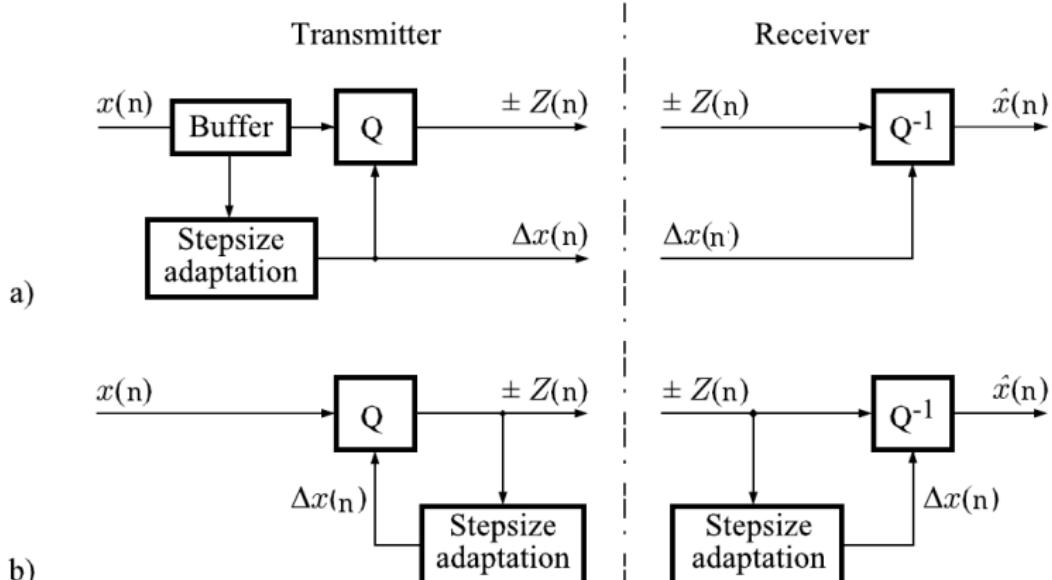
$$\hat{\sigma}_x^2(n) = \frac{1}{N} \sum_{m=0}^{N-1} x^2(n+m)$$

- Adaptive quantization backward (AQB)

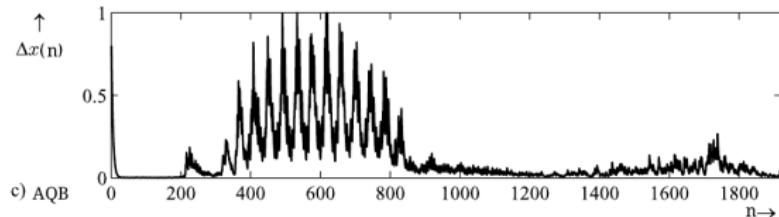
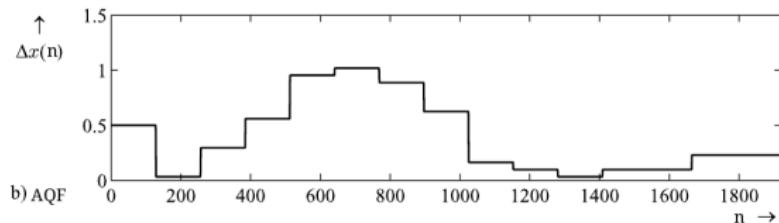
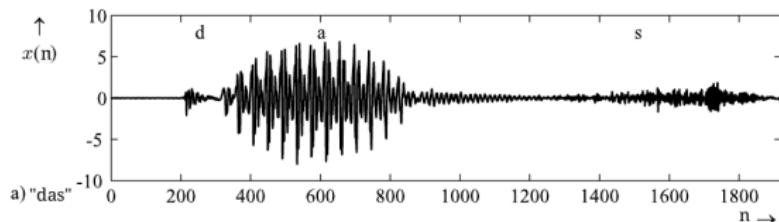
$$\hat{\sigma}_{\hat{x}}^2(n) = \alpha\hat{\sigma}_x^2(n-1) + (1-\alpha)\hat{x}^2(n-1), \quad 0 < \alpha < 1$$

- **Benefit:**

- Lower quantization noise level in low-level sounds
- Note: quantization noise is masked for high-level sounds
- Overall a low perceived quantization noise level



a) AQF b) AQB

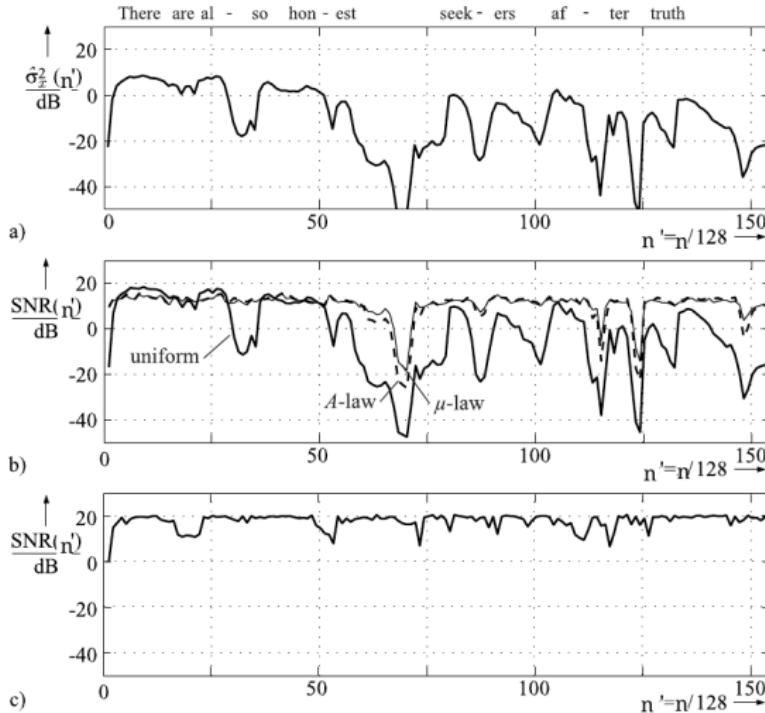


Example for the word "das":

→ For AQB the step size can be adapted more rapidly.

- Speech signal "das"
- AQF
- AQB

→ For an adaptive quantization, SNR is less dependent on signal level.



- a) σ_x^2
- b) uniform / A-law / μ -law
- c) AQF

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

- Simultaneous quantization of multiple random variables. Applications also in
 - Classification of vectors
 - Unsupervised learning
- Comprising L variables in one vector

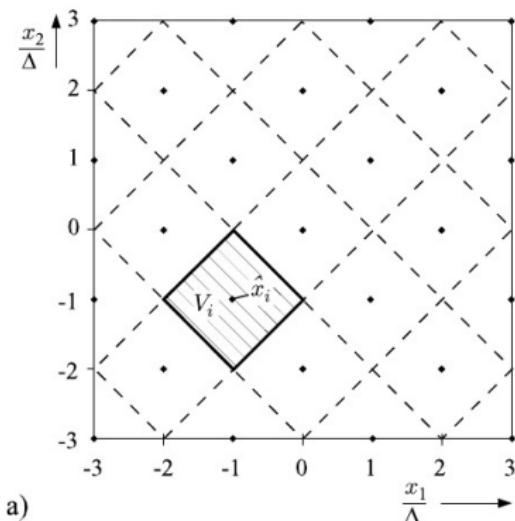
$$\mathbf{x} = (x_1, x_2, \dots, x_L)^T$$

- \mathbf{x} may represent e.g. successive time-samples or LPC-coefficients
- Task: Determine class representatives, so-called centroids

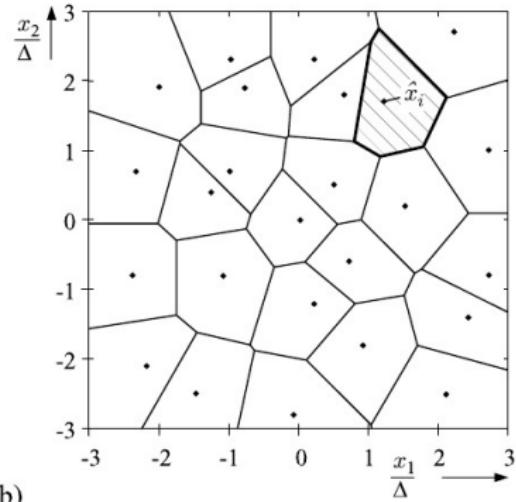
$$\hat{\mathbf{x}}_i = (\hat{x}_{i,1}, \hat{x}_{i,2}, \dots, \hat{x}_{i,L})^T$$

- Improved quantization (as compared to single variable quantization) by
 - Exploiting statistical dependencies between the elements of the vector.

- Residual noise is perceived as being less disturbing than for single variable quantization.
- Applications in signal transmission
 - Centroids are stored in a codebook and indexed
 - Sender and receiver have the same codebook available
 - → only codebook index needs to be transmitted!
- Comparison with scalar quantization
 - quantization levels ↔ centroids
 - represent the signal values in a certain range
 - quantization interval ↔ Voronoi-cells
 - each interval/cell is addressed by one index



a)



b)

© 2006 John Wiley & Sons, Ltd
 Vary, Martin - Digital Speech Transmission

- a) uniform resolution
- b) non-uniform resolution

- For a given vector \mathbf{x} , the selection of one centroid is obtained by minimizing a distance measure

$$d(\mathbf{x}, \hat{\mathbf{x}}_{\text{opt}}) = \min_i d(\mathbf{x}, \hat{\mathbf{x}}_i)$$

- The assignment of quantization index i and the centroid are stored in a

	i		$(\hat{x}_{i,1}, \hat{x}_{i,2})$
codebook	1	0001	(0.1, 0.9)
	2	0010	:
	:	:	:

- Coding of K centroids (codebook vectors) with equal probability of occurrence requires

$$w = \log_2(K) \text{ bit}$$

or

$$\tilde{w} = \frac{1}{L} \log_2(K) \frac{\text{bit}}{\text{vector element}}$$

- Frequently used measure of distance: weighted mean square error

$$d(\mathbf{x}, \hat{\mathbf{x}}_i) = (\mathbf{x} - \hat{\mathbf{x}}_i)^T \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}_i)$$

- In the simplest case, $\mathbf{W} = \mathbf{I}$ the identity matrix: *Euclidean distance*
- if \mathbf{W} is the inverse covariance matrix: *Mahalanobis distance*

- Vectors \mathbf{x} and K centroids (codebook entries) $\hat{\mathbf{x}}_i$ of dimension L
- Computation of the Euclidean (quadratic) distance (*nearest neighbor approach*)

$$(\mathbf{x} - \hat{\mathbf{x}}_i)^T (\mathbf{x} - \hat{\mathbf{x}}_i) = \|\mathbf{x} - \hat{\mathbf{x}}_i\|_2^2 \quad (16)$$

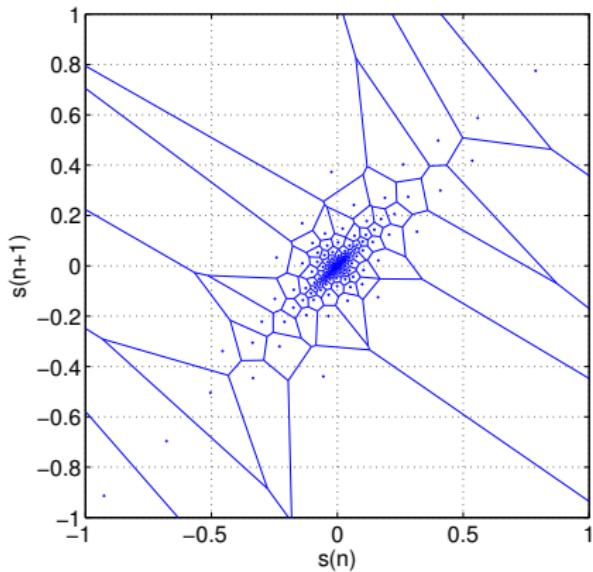
$$= \sum_{\mu=1}^L (x_\mu - \hat{x}_{i,\mu})^2 \quad (17)$$

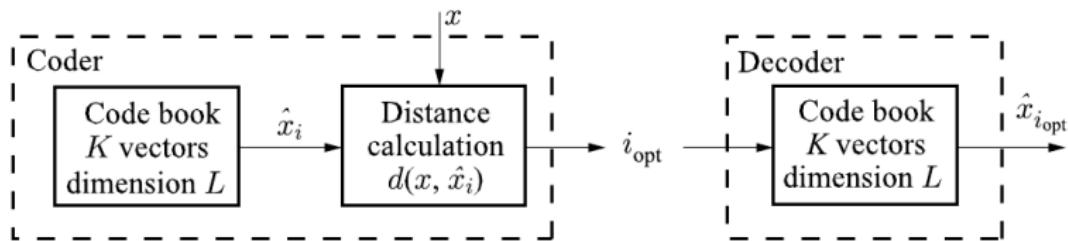
- L differences, L squares, $(L - 1)$ additions,
 $\rightarrow 3L - 1$ operations per vector $\hat{\mathbf{x}}_i$.
- For K codebook entries $\rightarrow K(3L - 1)$ operations.
- Example: Codebook with $K = 1024$ entries (centroids), $L = 40$ vector components, 8 kHz sampling rate
 $8000/40 \cdot 1024 \cdot 119 \approx 24.6$ MOPS (Mega operations per second)
 - $8000/40$, as one vector represents 40 samples
- \Rightarrow *full search, exhaustive search, brute-force* often too complex.
 Better: tree-structure

- Iterative Linde-Buzo-Gray (LBG) algorithm [Linde, Buzo, Gray, 1980]
(similar to K-means algorithm in data clustering)
 1. Set number of centroids $\rightarrow K$
 2. Take N vectors for training
 3. Initialize by choosing K centroids (e.g. randomly, uniformly)
 4. Use distance measure to determine nearest neighbor in codebook
 5. Compute center of gravity of data assigned to i th centroid \rightarrow new centroid
 6. go to 4 until stopping critereon is fulfilled (convergence or maximal number of iterations reached)

For speech samples, we have

- High correlation, \rightarrow Voronoi example exhibits high density at diagonal
- Small amplitudes are frequent \rightarrow Density of centroids in the origin is high

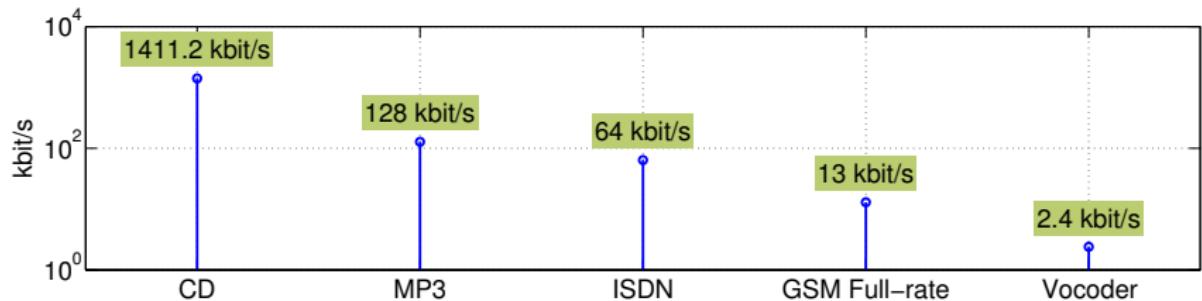


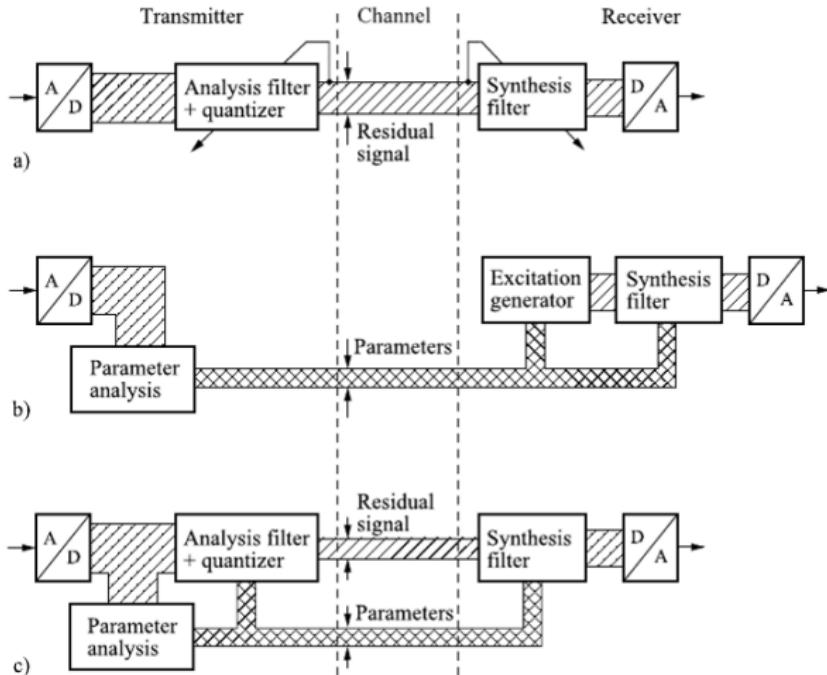


© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Waveform Coding
 - Parametric Coding
 - Hybrid Coding
 - Perceptual Coders

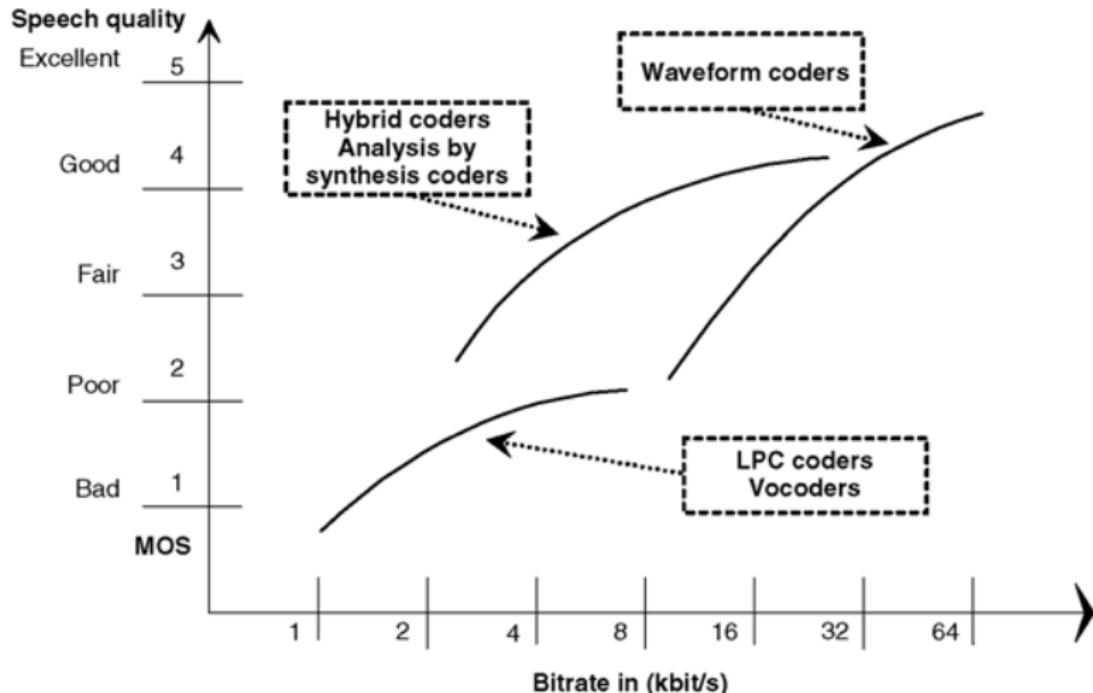
- Transmission of speech signals
- Reduction of redundant and irrelevant information
 - No exact representation of phase information (small changes in phase are not perceived)
 - If naturalness of synthesized speech is not important, even more information reduction is possible.
- Quantization of samples or parameters (e. g. LPC-coefficients) necessary.



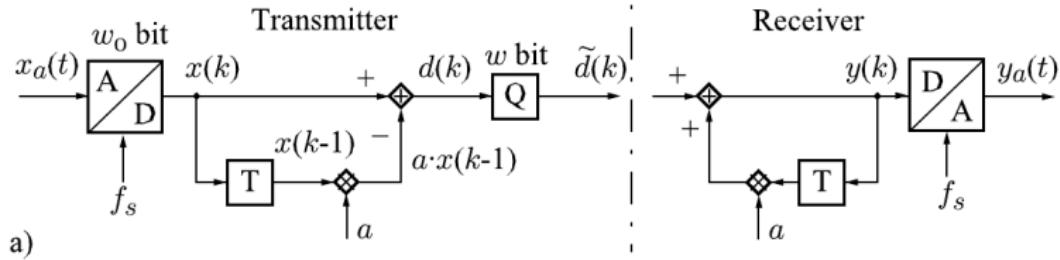


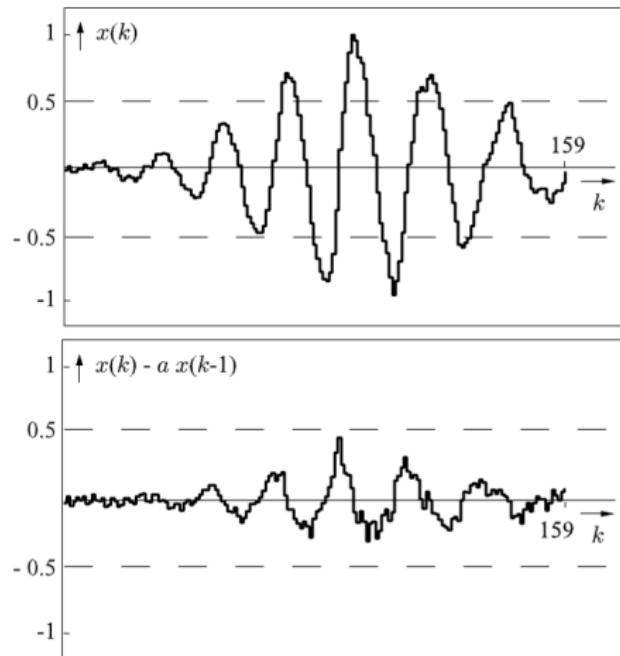
© 1998 B.G. Teubner
Vary, Heute, Hess - Digitale Sprachsignalverarbeitung

a) Waveform coding | b) Parametric coding | c) Hybrid coding



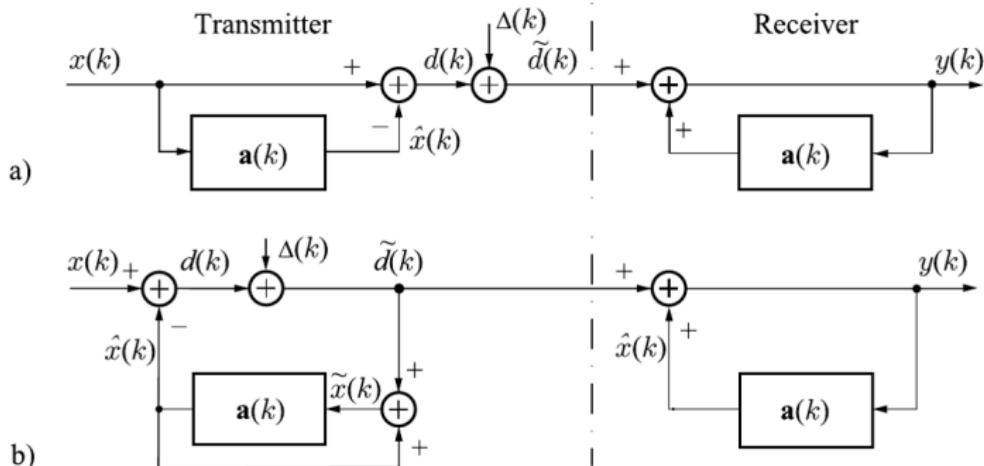
1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Waveform Coding
 - Parametric Coding
 - Hybrid Coding
 - Perceptual Coders





© 1998 B.G. Teubner
Vary, Heute, Hess - Digitale Sprachsignalverarbeitung

- For closed-loop (b), the parameters a do not have to be transmitted



© 1998 B.G. Teubner
 Vary, Heute, Hess - Digitale Sprachsignalverarbeitung

(ADPCM)

ADPCM closed-loop DPCM with adaptive backward quantization (AQB)

ITU G.726 (includes G.721):

- G.721: ADPCM @ 32 kbit/s (4 bits/sample)
- G.726: bit rates from 16-40 kbit/s
- Lower bitrate but somewhat audible distortions

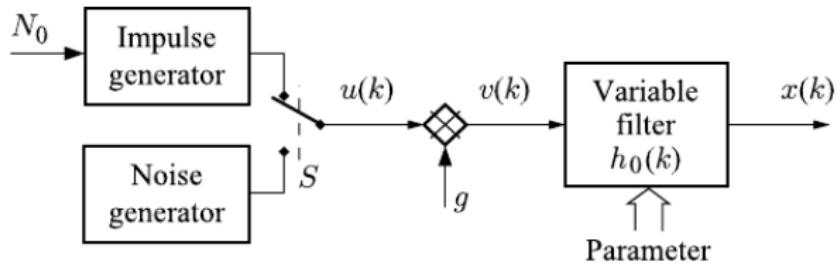
ITU G.727 (G.722)

- Wide-band ADPCM
- 64 kbit/s
- 7 kHz bandwidth

■ Waveform coding demonstration

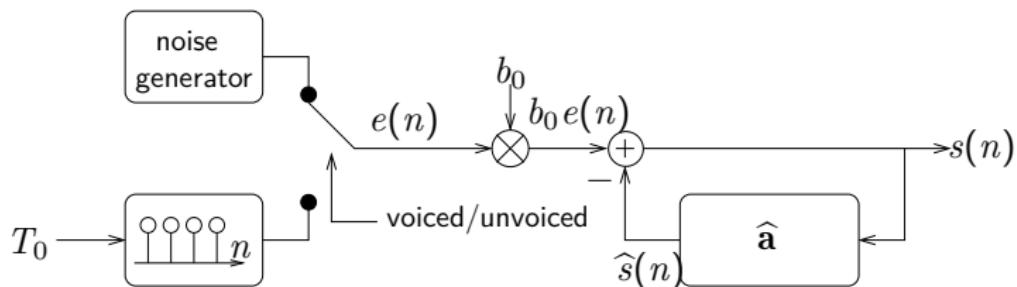
- ▶ PCM (128 kbit/s)
- ▶ G.711 (e.g. ISDN) (64 kbit/s)
- ▶ G.726 (e.g. DECT) (32 kbit/s)

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Waveform Coding
 - Parametric Coding
 - Hybrid Coding
 - Perceptual Coders



© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

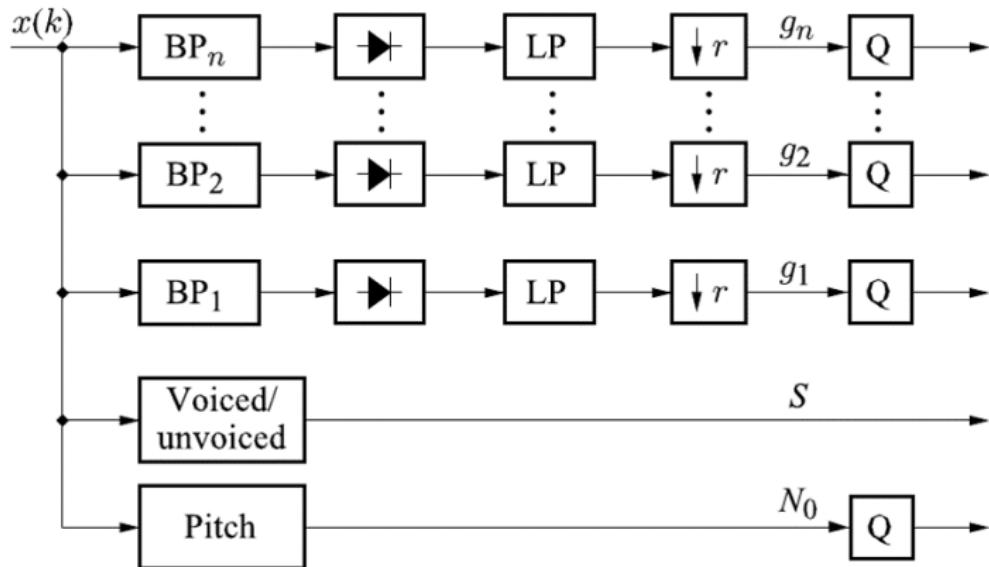
- With the resulting model, intelligible speech can be transmitted with bit rates as low as 2.4 kbit/s.
- per speech segment we need to store/transmit
 - information if voiced or unvoiced
 - fundamental period T_0 (if voiced)
 - scale parameter g
 - 10 AR coefficients to model the vocal tract filter



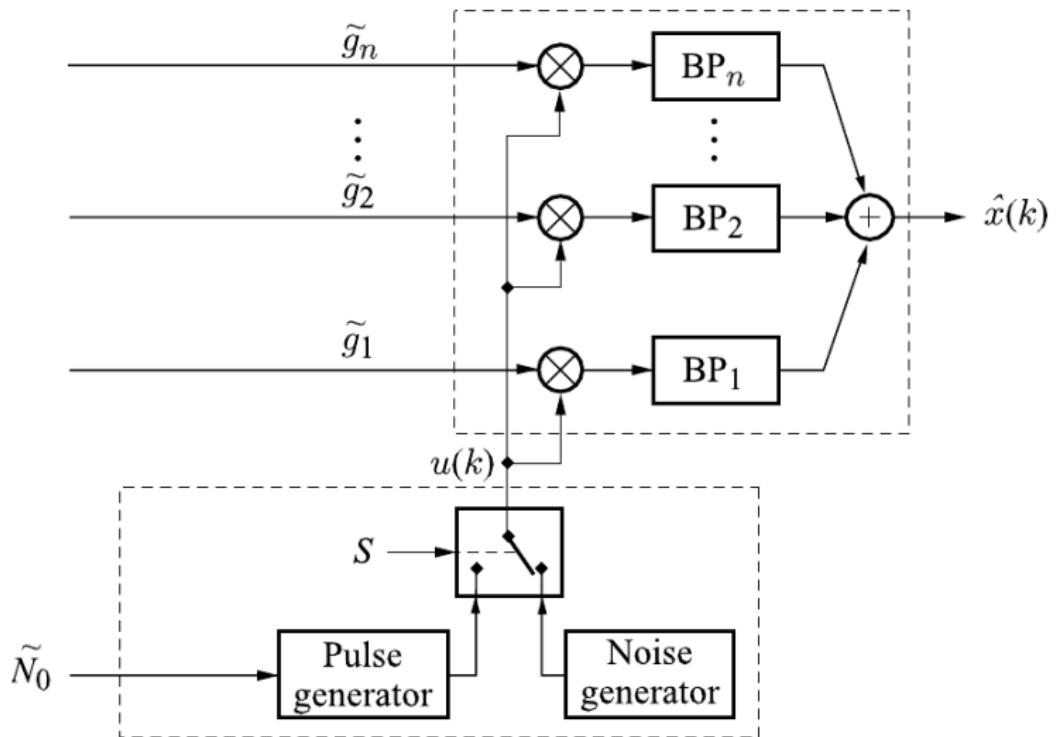
Sound example

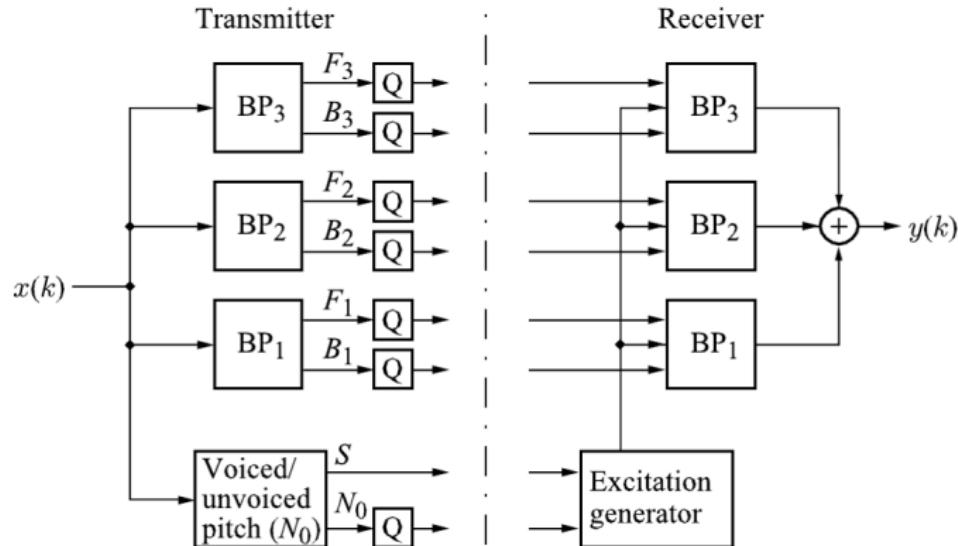
- only excitation $b_0 e(n)$:
- | only unvoiced:
- | V/UV:
- Coded speech (6.5 kbit/s):

Transmitter



Receiver





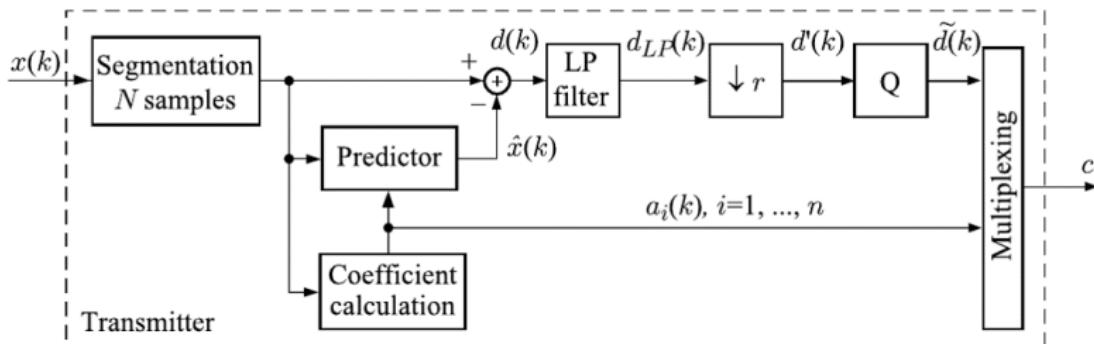
© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

- Very sensitive with respect to additive noise
- For music and general audio signals, the signal model is not well fulfilled, which may result in poor signal quality

1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Waveform Coding
 - Parametric Coding
 - Hybrid Coding
 - Perceptual Coders

- 0.5 - 1.5 bit/sample
- Residual signal is quantized rather roughly with respect to amplitude and/or time resolution
- Synthesis filter transmitted as side information
- Analysis-by-Synthesis

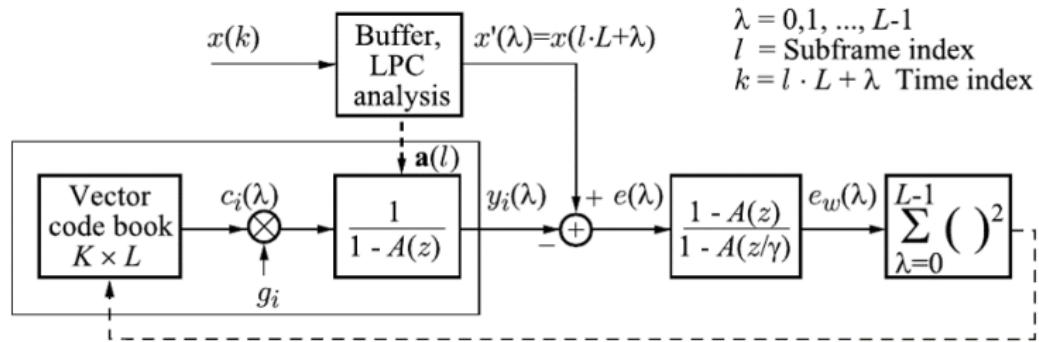
	Filter	Residue
LPC	Reflection Coefficients	Fundamental Frequency + Energy
RELP	Vector-quantized reflection coefficients	low-pass filter + downsampling
CELP	Vector-quantized Spectral Pairs	Line Long Term Predictor + vector quantizer

**Components**

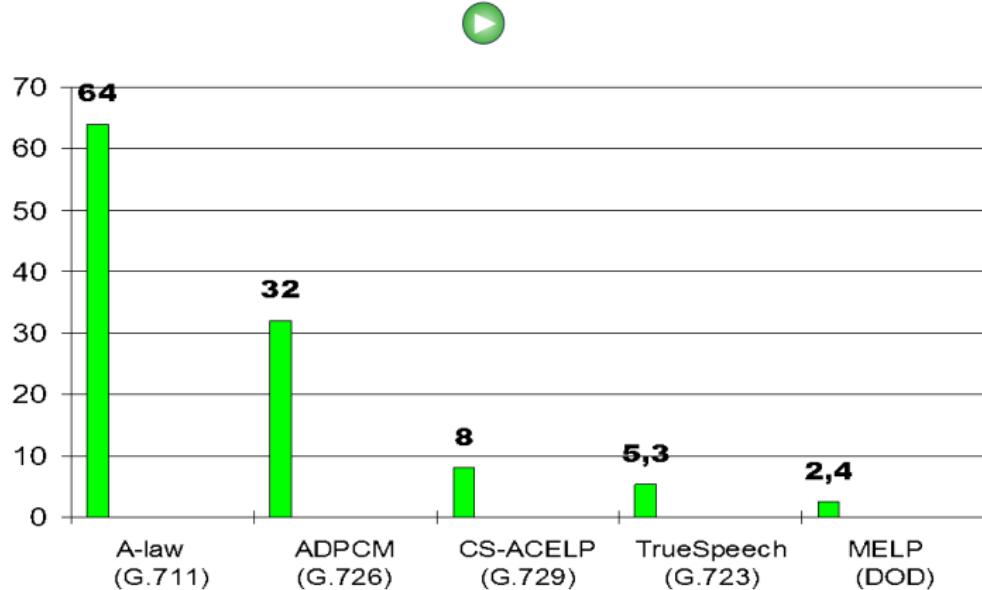
- LPC vector quantizer
- Low-pass filter for residual (0-500 Hz)
- Missing frequencies reconstructed based on replicas of 0-500 Hz range
- Residual quantizer: low bit rate waveform quantizer

Assessment

- Moderate to good quality, depending on bit rate
- Replaced by more modern CELP style coders (next)



- Vector quantization of
 - LPC coefficients
 - Excitation
- Choice of best codebook vector using **analysis-by-synthesis**

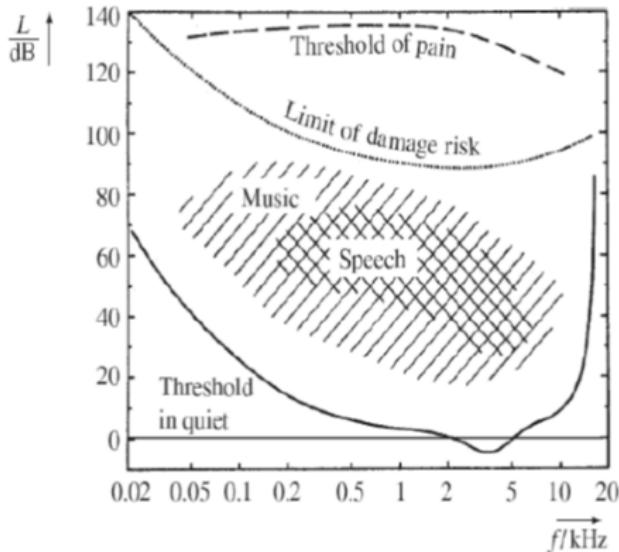


1. Introduction
2. Fundamental Frequency Estimation
3. Spectral Analysis of Audio Signals
4. Vocal Tract Model and Linear Prediction
5. Sampling, Quantization, and Speech Coding
 - Uniform Quantization
 - Non-Uniform Quantization
 - Adaptive Quantization
 - Vector Quantization
 - Speech Coding
 - Perceptual Coders
6. Speech Enhancement

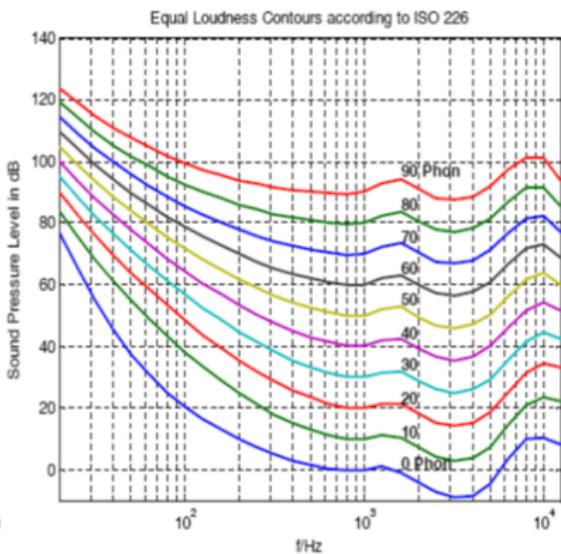
Basic Properties of mp3 and Successors

- Hearing sensitivity levels
- Perceptual coders exploit the properties of the human ear:
 - subband filters:** similar to the peripheral auditory processing in the inner ear
 - frequency masking:** a strong tone masks a soft tone nearby
 - forward masking:** a strong burst has a masking tail (temporal masking)
 - adaptive quantization:** use the bits where the information is (and where it can be perceived)
- Universal audio coding approach. Examples: mp3, AAC, ...
- For speech, same quality can be achieved at lower bitrate using source-optimized speech coders

The Hearing Area / Equal Loudness Curves



The hearing area [Zwicker, Fastl 1999]



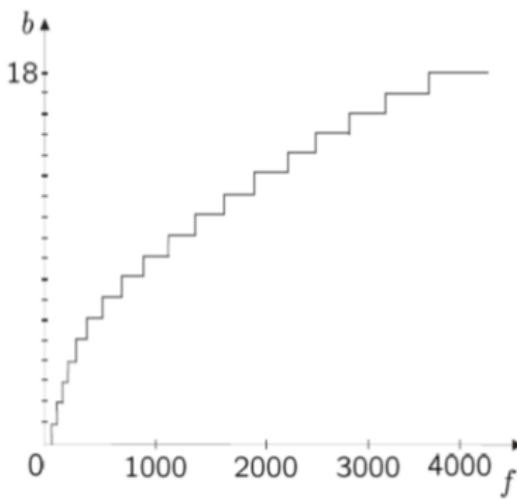
Equal Loudness Curves [ISO226]

Spectral Resolution of the Human Auditory System

Bark	f_{low} (in Hz)	f_{high} (in Hz)
0	0	100
1	100	200
2	200	300
3	300	400
4	400	510
5	510	630
6	630	770
7	770	920
8	920	1080
9	1080	1265
10	1265	1480
11	1480	1715
12	1715	1990
13	1990	2310
14	2310	2690
15	2690	3125
16	3125	3675
17	3675	4350

Frequency-to-Bark-Transformation:

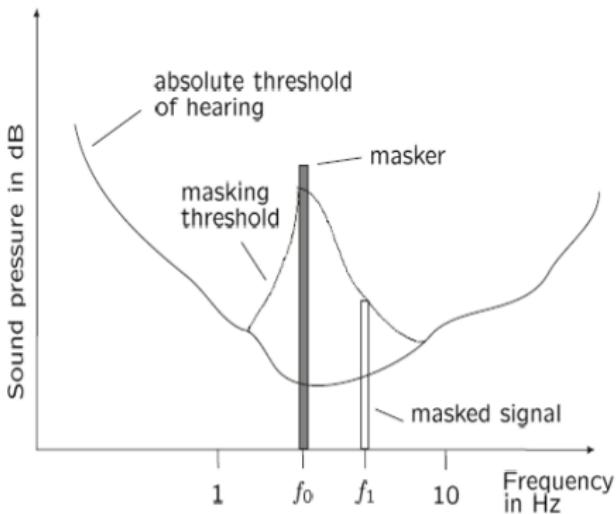
$$b = 13 \cdot \arctan\left(0.76 \frac{f}{1000}\right) + 3.5 \cdot \arctan\left(\frac{f}{7500}\right)^2$$



Frequency Masking

Masking effect:

If two sounds occur simultaneously,
the weaker signal can be made
inaudible by the stronger signal,
i.e., the masker, if they are close
enough in frequency and time.



1. Introduction
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
 - Single Channel Speech Enhancement
 - Multichannel Speech Enhancement
7. Machine Learning Basics
8. Source Separation



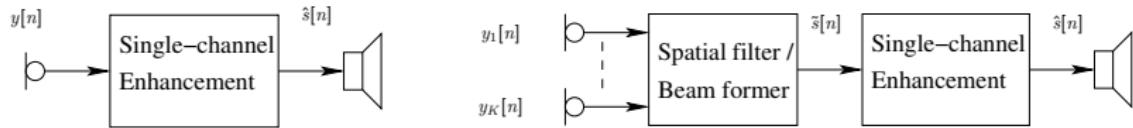
Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



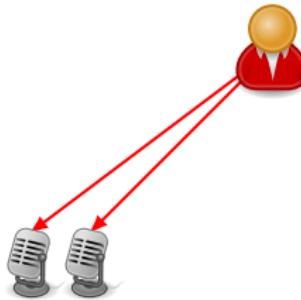
Speech Enhancement

- Speech acquisition in a noisy environment from a single observation
 - size and cost limitations may allow for only one microphone,
 - multichannel algorithms can be decomposed in to spatial processing and single channel postfilter.



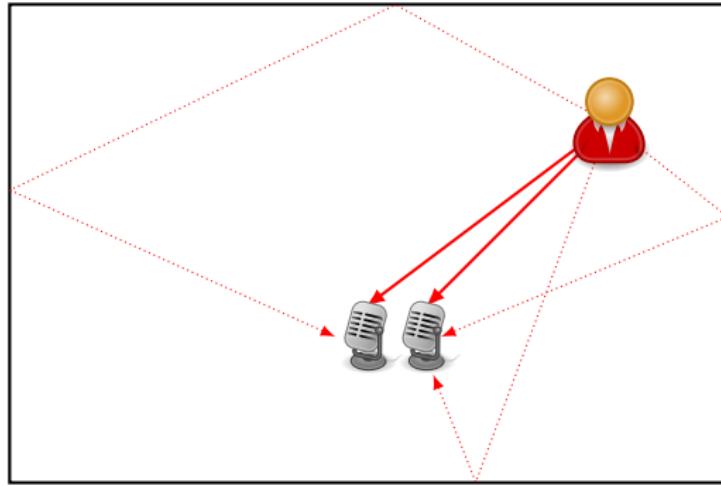
- Applications: hearing instruments, cell phones, speech recognition.





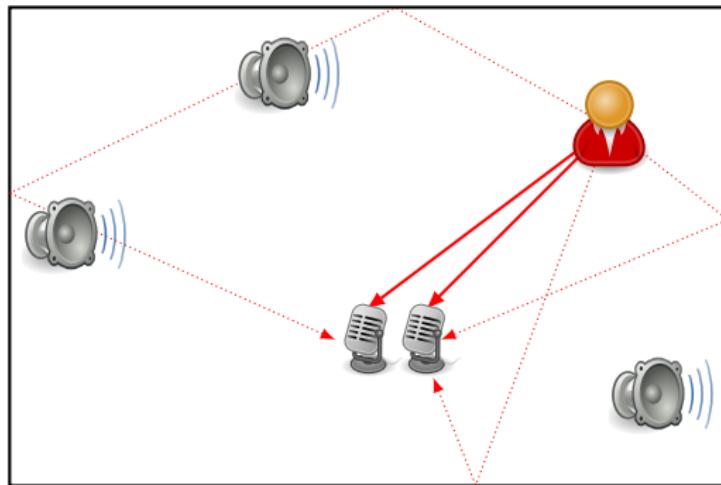
- Signal model: $y_m(t) = s_m(t)$

Goal: Extract desired source $s(t)$ from recorded mixture $y_m(t)$



- Signal model: $y_m(t) = s(t) * h_m(t)$
- Conversation disturbed by
 - reflections from the walls

Goal: Extract desired source $s(t)$ from recorded mixture $y_m(t)$

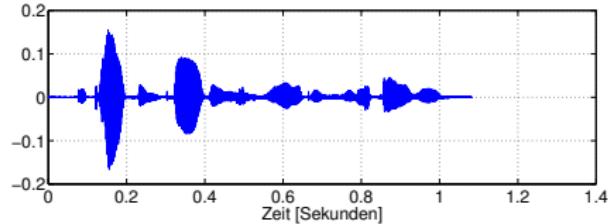
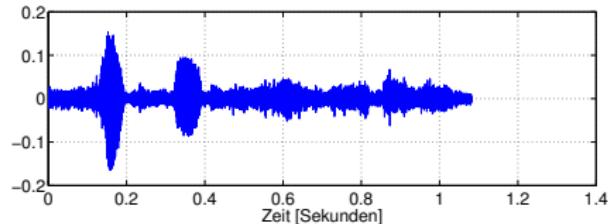


- Signal model: $y_m(t) = s(t) * h_m(t) + \sum_{i=1}^I n_{i,m}(t)$
- Conversation disturbed by
 - reflections from the walls
 - additive noise

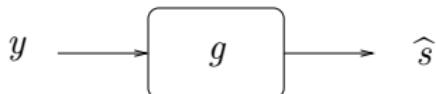
Goal: Extract desired source $s(t)$ from recorded mixture $y_m(t)$

- Speech signal may be disturbed by
 - additive noise sources
 - reverberation
- Noisy reverberant speech is captured by one microphone.
- How can we distinguish between speech and noise?
- Find transformation space in which the speech and noise signals are well separable.
- Exploit different statistical properties of speech and noise signals.

1. Introduction
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
 - Single Channel Speech Enhancement
 - Multichannel Speech Enhancement
7. Machine Learning Basics
8. Source Separation

Clean Speech $s(n)$ Noisy Speech $y(n) = s(n) + n(n)$ 

- Problem: estimate clean speech s given the noisy signal $y = s + n$.



- In this context *linear* means that g is not a function of the input y .
- The clean speech estimate is obtained by convolution $\hat{s} = y * g$.
- Minimize the mean of the squared error (MMSE) between the true speech signal s and the estimate \hat{s}

$$g = \arg \min_g E((s - \hat{s})^2)$$

Time-domain Wiener filter

■

$$\hat{s}(n) = g(n) * y(n) = \sum_{\nu=-\infty}^{\infty} g(\nu)y(n-\nu)$$

■

$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$

Time-domain Wiener filter



$$\hat{s}(n) = g(n) * y(n) = \sum_{\nu=-\infty}^{\infty} g(\nu)y(n-\nu)$$



$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$



$$\sum_{\nu=-\infty}^{\infty} g(\nu)\varphi_{yy}(n-\nu) = \varphi_{ys}(n)$$

Time-domain Wiener filter



$$\hat{s}(n) = g(n) * y(n) = \sum_{\nu=-\infty}^{\infty} g(\nu)y(n-\nu)$$



$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$



$$\sum_{\nu=-\infty}^{\infty} g(\nu)\varphi_{yy}(n-\nu) = \varphi_{ys}(n)$$



$$G(e^{j\Omega}) = \frac{\Phi_{YS}(e^{j\Omega})}{\Phi_{YY}(e^{j\Omega})}$$

Time-domain Wiener filter



$$\hat{s}(n) = g(n) * y(n) = \sum_{\nu=-\infty}^{\infty} g(\nu)y(n-\nu)$$



$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$

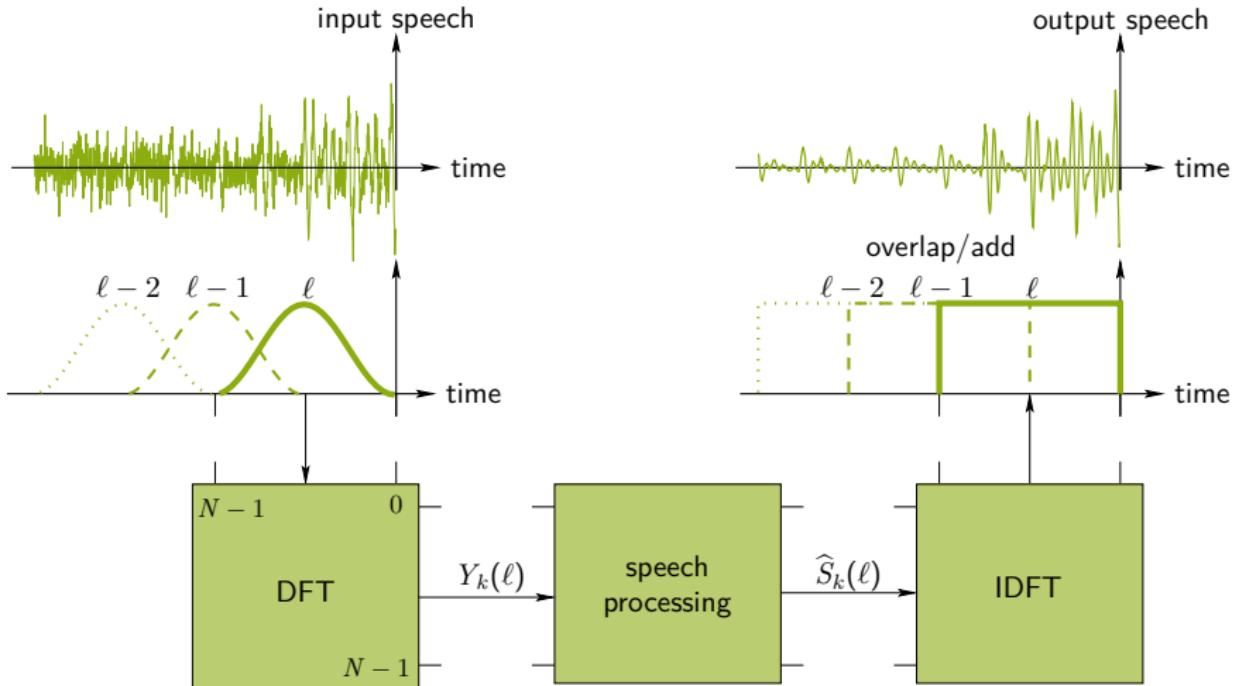


$$\sum_{\nu=-\infty}^{\infty} g(\nu)\varphi_{yy}(n-\nu) = \varphi_{ys}(n)$$



$$G(e^{j\Omega}) = \frac{\Phi_{YS}(e^{j\Omega})}{\Phi_{YY}(e^{j\Omega})} = \frac{\Phi_{SS}(e^{j\Omega})}{\Phi_{SS}(e^{j\Omega}) + \Phi_{NN}(e^{j\Omega})}$$

- $\Phi_{SS}(e^{j\Omega})$: speech power spectral density (Leistungsdichthespektrum)
- $\Phi_{NN}(e^{j\Omega})$: noise power spectral density



- Signal model: $Y_k(\ell) = S_k(\ell) + N_k(\ell)$
- Convolution in time-domain turns into multiplication in the frequency domain

$$\hat{s}(n) = y(n) * g(n) \quad \circ\bullet \quad \hat{S}_k(\ell) = G_k(\ell) Y_k(\ell)$$

- We assume that the DFT decorrelates the spectral coefficients
- We can treat every time-frequency point (k, ℓ) independently
-

$$\min(E(|S_k - G_k Y_k|^2))$$

- Signal model: $Y_k(\ell) = S_k(\ell) + N_k(\ell)$
- Convolution in time-domain turns into multiplication in the frequency domain

$$\hat{s}(n) = y(n) * g(n) \quad \circledast \quad \hat{S}_k(\ell) = G_k(\ell) Y_k(\ell)$$

- We assume that the DFT decorrelates the spectral coefficients
- We can treat every time-frequency point (k, ℓ) independently
-

$$\min(E(|S_k - G_k Y_k|^2))$$

$$G_k = \frac{E(S_k Y_k^*)}{E(|Y_k|^2)}$$

- Signal model: $Y_k(\ell) = S_k(\ell) + N_k(\ell)$
- Convolution in time-domain turns into multiplication in the frequency domain

$$\hat{s}(n) = y(n) * g(n) \quad \circledast \quad \hat{S}_k(\ell) = G_k(\ell) Y_k(\ell)$$

- We assume that the DFT decorrelates the spectral coefficients
- We can treat every time-frequency point (k, ℓ) independently
-

$$\min(E(|S_k - G_k Y_k|^2))$$

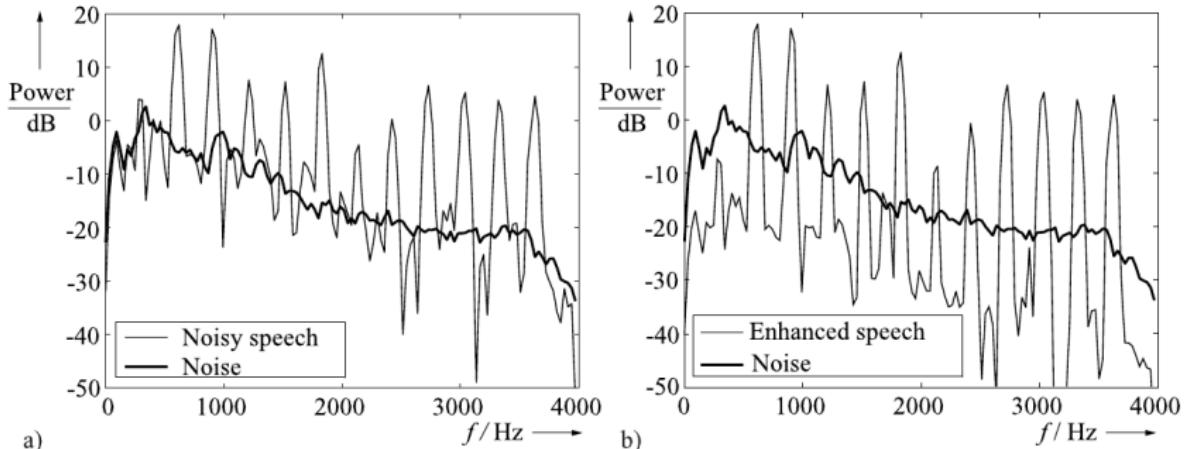
$$G_k = \frac{E(S_k Y_k^*)}{E(|Y_k|^2)} = \frac{E(|S_k|^2)}{E(|S_k^2|) + E(|N_k^2|)} = \frac{\sigma_{s,k}^2}{\sigma_{s,k}^2 + \sigma_{n,k}^2}$$

- General definition of the variance: $\text{var}(X) = \mathbb{E}(|X - \mathbb{E}(X)|^2)$
- For a zero-mean random variable $\mathbb{E}(|S_k|^2)$ is the variance, i.e.

$$\mathbb{E}(|S_k|^2) = \sigma_{s,k}^2$$

- $$G_k = \frac{\sigma_{s,k}^2}{\sigma_{s,k}^2 + \sigma_{n,k}^2}$$
- Comparing to the time-domain solution, we can also interpret the variance $\sigma_{s,k}^2$ as the power spectral density $\Phi_{SS}(e^{j\Omega})$.

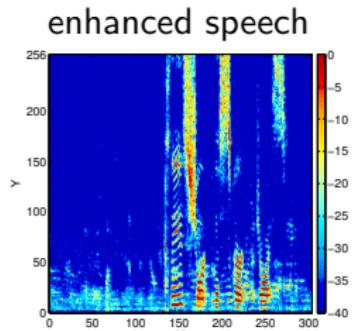
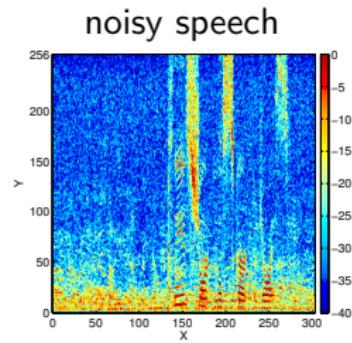
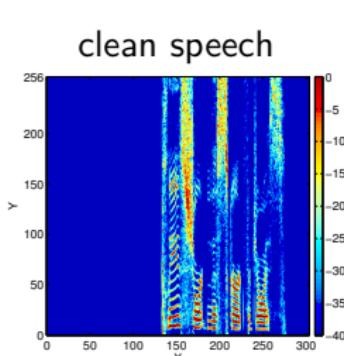
$$\hat{S}_k(\ell) = \frac{\sigma_{s,k}^2(\ell)}{\sigma_{s,k}^2(\ell) + \sigma_{n,k}^2(\ell)} Y_k(\ell)$$



© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

- a) Short time spectrum of the noisy signal and the estimated noise PSD.
- b) Short-time spectrum of the enhanced signal and the estimated noise PSD.

$$\hat{S}_k(\ell) = \frac{\sigma_{s,k}^2(\ell)}{\sigma_{s,k}^2(\ell) + \sigma_{n,k}^2(\ell)} Y_k(\ell)$$



- Noise dominated time-frequency points are attenuated.
- Good performance depends on a robust estimation of the speech and noise PSDs!

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2)$$

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2) = \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) + 2\text{Re}\{\mathbb{E}(SN^*)\} \quad (18)$$

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2) = \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) + 2\text{Re}\{\mathbb{E}(SN^*)\} \quad (18)$$

$$= \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) \quad (19)$$

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2) = \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) + 2\text{Re}\{\mathbb{E}(SN^*)\} \quad (18)$$

$$= \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) \quad (19)$$

■

$$\mathbb{E}(|S|^2) = \mathbb{E}(|Y|^2) - \mathbb{E}(|N|^2) \quad (20)$$

$$= \mathbb{E}(|Y|^2) \left(1 - \frac{\mathbb{E}(|N|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (21)$$

$$= \mathbb{E}(|Y|^2) \left(\frac{\mathbb{E}(|S|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (22)$$

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2) = \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) + 2\text{Re}\{\mathbb{E}(SN^*)\} \quad (18)$$

$$= \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) \quad (19)$$

-

$$\mathbb{E}(|S|^2) = \mathbb{E}(|Y|^2) - \mathbb{E}(|N|^2) \quad (20)$$

$$= \mathbb{E}(|Y|^2) \left(1 - \frac{\mathbb{E}(|N|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (21)$$

$$= \mathbb{E}(|Y|^2) \left(\frac{\mathbb{E}(|S|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (22)$$

- Note that the spectral subtraction gain function $\frac{\mathbb{E}(|S|^2)}{\mathbb{E}(|Y|^2)}$ resembles the Wiener filter, but is applied to PSDs! Note the $\mathbb{E}(\cdot)$!

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

2

- Idea: Subtract noise PSD from noisy PSD
- Note that

$$\mathbb{E}(|Y|^2) = \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) + 2\text{Re}\{\mathbb{E}(SN^*)\} \quad (18)$$

$$= \mathbb{E}(|S|^2) + \mathbb{E}(|N|^2) \quad (19)$$

■

$$\mathbb{E}(|S|^2) = \mathbb{E}(|Y|^2) - \mathbb{E}(|N|^2) \quad (20)$$

$$= \mathbb{E}(|Y|^2) \left(1 - \frac{\mathbb{E}(|N|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (21)$$

$$= \mathbb{E}(|Y|^2) \left(\frac{\mathbb{E}(|S|^2)}{\mathbb{E}(|Y|^2)}\right) \quad (22)$$

- Note that the spectral subtraction gain function $\frac{\mathbb{E}(|S|^2)}{\mathbb{E}(|Y|^2)}$ resembles the Wiener filter, but is applied to PSDs! Note the $\mathbb{E}(\cdot)$!
- Spectral subtraction is related to the square-root of the Wiener filter.

² S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

- In practice, several variants of spectral subtraction exist, which can be controlled by parameters ν and μ , as

$$\widehat{|S|^2} = |Y|^2 \left(1 - \left(\frac{\overline{|N|^2}}{\overline{|Y|^2}} \right)^\mu \right)^\nu$$

- here, the overbars indicate some temporal smoothing.
- Proposals with and without temporal smoothing exist.
 - Trade-off between musical noise and speech distortions.
- Spectral Subtraction and its variants are somewhat heuristically motivated.
- Note that, in general

$$\overline{|S|^2} \neq \overline{|Y|^2} - \overline{|N|^2}$$

- General definitions
 - Θ : estimated quantity; Y : observation
 - $p(\Theta | Y)$: *posterior*
 - $p(Y | \Theta)$: *likelihood*
 - $p(\Theta)$: *prior*
 - $p(Y)$: *evidence model*
- related by Bayes' theorem

$$P(A | B) = \frac{P(A, B)}{P(B)} = \frac{P(B | A)P(A)}{P(B)}$$

- Often, finding a model for likelihood and prior is easier than finding a model for the posterior directly \rightarrow Bayes' theorem.

- Maximum likelihood (ML) estimation

$$\hat{\Theta}^{\text{ML}} = \arg \max_{\theta} p(Y | \Theta)$$

- Maximum a posteriori (MAP) estimation

$$\hat{\Theta}^{\text{MAP}} = \arg \max_{\theta} p(\Theta | Y) = \arg \max_{\theta} p(Y|\Theta)p(\Theta)$$

- MMSE

$$\hat{\Theta}^{\text{MMSE}} = \arg \min_{\hat{\Theta}} \mathbb{E}\left(|\Theta - \hat{\Theta}|^2\right) = \mathbb{E}(\Theta | Y)$$

- In the derivation we have seen so far, the Wiener filter is constrained to be a linear function in Y .
- Q: Can we find a better filter by dropping the linear constraint?

- In the derivation we have seen so far, the Wiener filter is constrained to be a linear function in Y .
- Q: Can we find a better filter by dropping the linear constraint?
- Goal:

$$\widehat{S} = \arg \min_{\widehat{S}} E\left(\left(S - \widehat{S}\right)^2 | Y\right)$$

where \widehat{S} is some function of the noisy input Y , i.e. $\widehat{S} = f(Y)$.

- In the derivation we have seen so far, the Wiener filter is constrained to be a linear function in Y .
- Q: Can we find a better filter by dropping the linear constraint?
- Goal:

$$\widehat{S} = \arg \min_{\widehat{S}} E\left(\left(S - \widehat{S}\right)^2 \mid Y\right)$$

where \widehat{S} is some function of the noisy input Y , i.e. $\widehat{S} = f(Y)$.

- It can be shown that the non linearly constrained MMSE estimator results in the conditional mean estimator

$$\widehat{S} = E(S \mid Y) \tag{23}$$

$$= \int_{-\infty}^{\infty} Sp(S \mid Y) dS \tag{24}$$

- The form of the resulting generalized MMSE estimator depends on the chosen models for the prior and likelihood
- Example: Gaussian distribution for prior and likelihood

- prior
$$p(S) = \frac{1}{\pi\sigma_s^2} \exp\left(-\frac{|S|^2}{\sigma_s^2}\right)$$

- likelihood
$$p(Y | S) = \frac{1}{\pi\sigma_n^2} \exp\left(-\frac{|Y - S|^2}{\sigma_n^2}\right)$$

- evidence model
$$p(Y) = \frac{1}{\pi(\sigma_n^2 + \sigma_s^2)} \exp\left(-\frac{|Y|^2}{(\sigma_n^2 + \sigma_s^2)}\right)$$

- The form of the resulting generalized MMSE estimator depends on the chosen models for the prior and likelihood
- Example: Gaussian distribution for prior and likelihood

- prior $p(S) = \frac{1}{\pi\sigma_s^2} \exp\left(-\frac{|S|^2}{\sigma_s^2}\right)$
- likelihood $p(Y | S) = \frac{1}{\pi\sigma_n^2} \exp\left(-\frac{|Y - S|^2}{\sigma_n^2}\right)$
- evidence model $p(Y) = \frac{1}{\pi(\sigma_n^2 + \sigma_s^2)} \exp\left(-\frac{|Y|^2}{(\sigma_n^2 + \sigma_s^2)}\right)$

Result

$$\hat{S} = E(S | Y) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_n^2} Y$$

- The Wiener filter is the optimal non-constrained filter, if speech and noise are Gaussian distributed!

- Different filter solutions appear if
 - We optimize for functions of the spectral coefficients $f(S)$
 - we assume different distributions.

Log spectral amplitude estimator (LSA)³

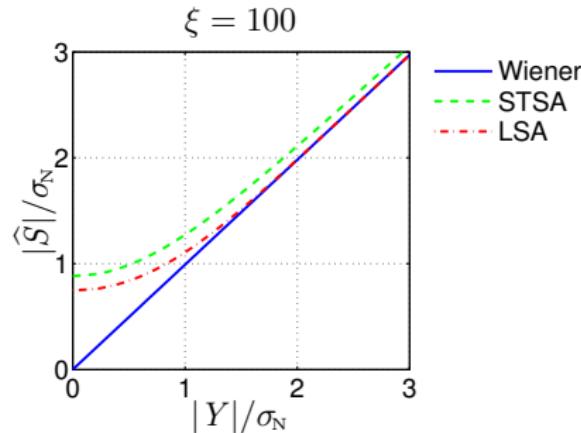
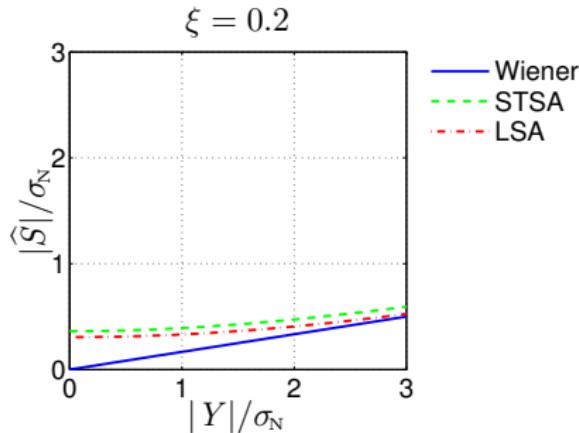
Motivation: a logarithmic distortion measure reflects speech perception better

- Optimal log-amplitude estimator for Gaussian distributed speech and noise.
- Nonlinear estimator, i.e. gain is a function of the input
-

$$\widehat{|S|} = \exp(E(\log |S| \mid Y)) = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2} \int_v^\infty \frac{e^{-t}}{t} dt\right) |Y|$$

- The integral is known as *exponential integral of v*.
- It can be shown that the output is always lower or equal to the STSA
→ more noise reduction than STSA.
- for large inputs v , integral tends to zero, and LSA → Wiener

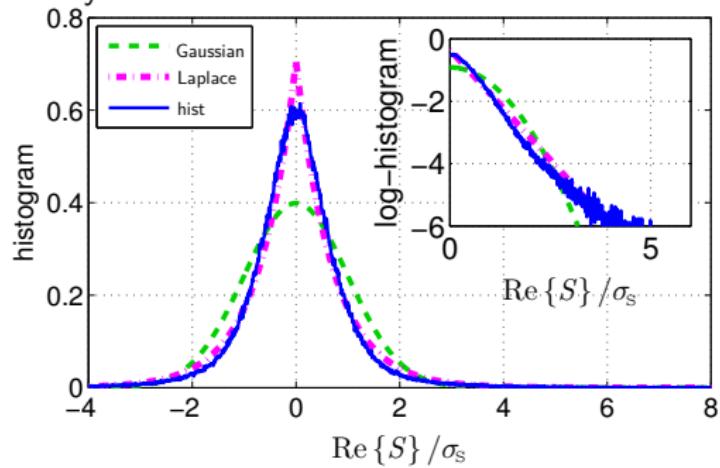
³ Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.



- Wiener filter has largest suppression,
- LSA and STSA have less suppression for low input values
→ masks musical noise.

4 5
,

- Computing the histogram of speech reveals that speech follows a more heavy-tailed distribution.

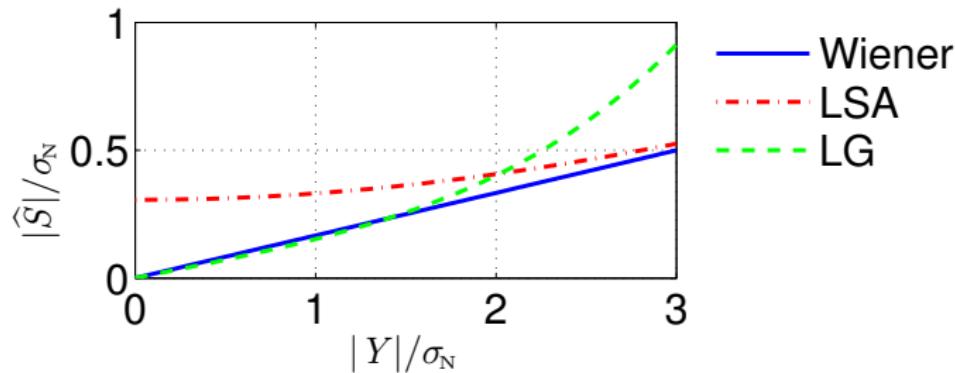


→ Laplace-Gauss (LG) Estimator

⁴ R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

⁵ T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

$$\xi = \sigma_s^2 / \sigma_n^2 = 0.2, G_{\text{Wiener}} = 1/6$$



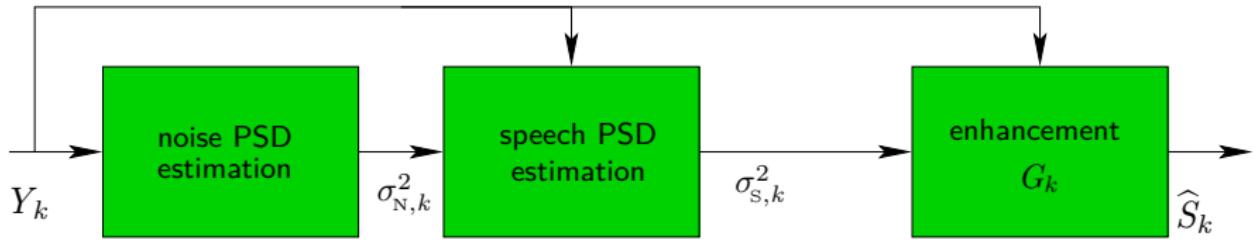
- Super-Gaussian estimator (LG) → less attenuation for large inputs.
 - ✓ less speech distortions,
 - ✗ more outliers / processing artifacts.

Wiener:

LSA:

LG:

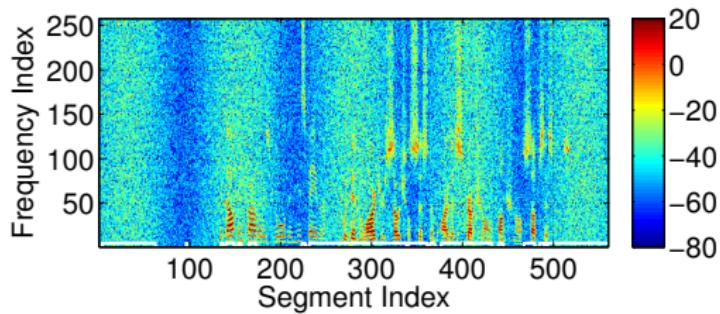
- The Wiener filter is the MMSE-optimal *linear* estimator for clean speech spectral coefficients
- The non-linearly constrained estimator is given by the conditional mean, e.g. $E(S | Y)$.
- The Wiener filter is the *non-linearly-constrained* MMSE-optimal estimator of complex speech coefficients *if* speech and noise are Gaussian distributed.
- Different non-linear estimators arise when
 - functions of the speech coefficients are estimated
 - $E(|S| | Y) \rightarrow$ short-time spectral amplitude estimator (STSA)
 - $E(\log |A| | Y) \rightarrow$ log-spectral amplitude estimator (LSA)
 - When different distributions for the speech or noise coefficients are chosen
 - so-called super-Gaussian estimators (e.g. Laplace-Gauss estimator)



- Where, e.g., $G_k = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2}$, the Wiener filter.

- Estimation of the noise PSD
 - Estimation in speech absence using voice activity detector
 - Minimum Statistics approach
 - Employ speech presence probability / MMSE-estimation

- Update noise PSD in signal segments where speech is absent.
 - Threshold on frame energy $\sum_{k=0}^{N-1} |Y_k|^2 \geq \tau$
 - ✖ Best threshold is also a function of σ_N^2
 - ✖ When the noise level changes during speech activity this leads to either
 - Noise overestimation \rightarrow speech distortions, or
 - Noise underestimation \rightarrow musical noise.



Update noise in segments where speech is absent.

- Even for a perfect VAD: Not robust in nonstationary noise.
- In reality the VAD has to be estimated making performance even worse.

Clean:

Ideal VAD:

Noisy:

VAD:

- Difference to VAD:
 - Soft decision instead of hard decision
 - Applied per frequency bin
- \mathcal{H}_1 denotes the hypothesis that speech is present in a time-frequency bin

$$Y = \begin{cases} S + N, & \text{given } \mathcal{H}_1 \\ N, & \text{given } \mathcal{H}_0 \end{cases}$$

- $P(\mathcal{H}_1 \mid Y)$: a posteriori probability that speech is present.
- SPP-based estimation of the noise periodogram

$$\widehat{|N|^2} = P(\mathcal{H}_0 \mid Y) |Y|^2 + P(\mathcal{H}_1 \mid Y) \widehat{\sigma_N^2}. \quad (25)$$

- Apply subsequent recursive smoothing to approximate $\widehat{\sigma_N^2} = E(|N|^2)$

$$\widehat{\sigma_N^2}(\ell) = \alpha \widehat{\sigma_N^2}(\ell - 1) + (1 - \alpha) \widehat{|N(\ell)|^2}. \quad (26)$$

- Apply Bayes' theorem

$$P(\mathcal{H}_1 \mid Y) = \frac{p(Y \mid \mathcal{H}_1)P(\mathcal{H}_1)}{p(Y \mid \mathcal{H}_1)P(\mathcal{H}_1) + p(Y \mid \mathcal{H}_0)P(\mathcal{H}_0)}$$

- Model assumptions

- Apply Bayes' theorem

$$P(\mathcal{H}_1 \mid Y) = \frac{p(Y \mid \mathcal{H}_1)P(\mathcal{H}_1)}{p(Y \mid \mathcal{H}_1)P(\mathcal{H}_1) + p(Y \mid \mathcal{H}_0)P(\mathcal{H}_0)}$$

- Model assumptions
 - Prior: $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$

- Apply Bayes' theorem

$$P(\mathcal{H}_1 \mid Y) = \frac{p(Y \mid \mathcal{H}_1) P(\mathcal{H}_1)}{p(Y \mid \mathcal{H}_1) P(\mathcal{H}_1) + p(Y \mid \mathcal{H}_0) P(\mathcal{H}_0)}$$

- Model assumptions
 - Prior: $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$
 - Likelihoods:
 - $p(Y \mid \mathcal{H}_0) = \mathcal{N}(0; \sigma_{\text{N}}^2)$
 - $p(Y \mid \mathcal{H}_1) = \mathcal{N}(0; \sigma_{\text{N}}^2 + \sigma_{\text{s}}^2) = \mathcal{N}(0; \sigma_{\text{N}}^2(1 + \xi_{\mathcal{H}_1}))$
 - $\xi_{\mathcal{H}_1}$ is the SNR that can be expected in speech presence, e.g. $\xi_{\mathcal{H}_1} = 15 \text{ dB}$.

- Apply Bayes' theorem

$$P(\mathcal{H}_1 \mid Y) = \frac{p(Y \mid \mathcal{H}_1) P(\mathcal{H}_1)}{p(Y \mid \mathcal{H}_1) P(\mathcal{H}_1) + p(Y \mid \mathcal{H}_0) P(\mathcal{H}_0)}$$

- Model assumptions

- Prior: $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$

- Likelihoods:

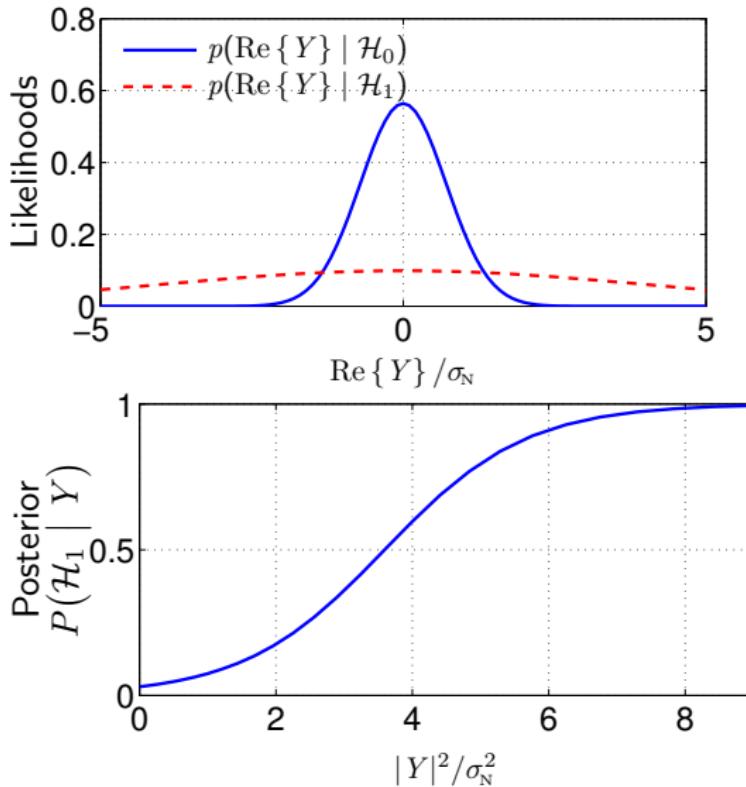
- $p(Y \mid \mathcal{H}_0) = \mathcal{N}(0; \sigma_N^2)$

- $p(Y \mid \mathcal{H}_1) = \mathcal{N}(0; \sigma_N^2 + \sigma_s^2) = \mathcal{N}(0; \sigma_N^2(1 + \xi_{\mathcal{H}_1}))$

- $\xi_{\mathcal{H}_1}$ is the SNR that can be expected in speech presence, e.g.
 $\xi_{\mathcal{H}_1} = 15 \text{ dB}$.



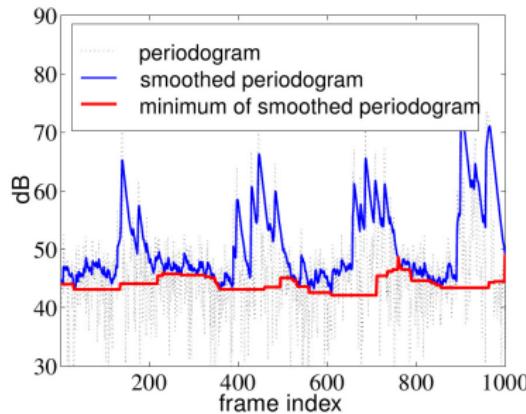
$$P(\mathcal{H}_1 \mid Y) = \left(1 + \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} (1 + \xi_{\mathcal{H}_1}) \exp\left(\frac{|Y|^2}{\sigma_N^2} \frac{\xi_{\mathcal{H}_1}}{1 + \xi_{\mathcal{H}_1}}\right) \right)^{-1}$$



- $P(\mathcal{H}_1 | Y) = \frac{p(Y|\mathcal{H}_1)}{p(Y|\mathcal{H}_1)+p(Y|\mathcal{H}_0)}$
- Here $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 0.5$
- Here $\xi_{\mathcal{H}_1} = 15 \text{ dB}$.
- $\xi_{\mathcal{H}_1}$ is part of the model and represents the SNR typical for speech presence.

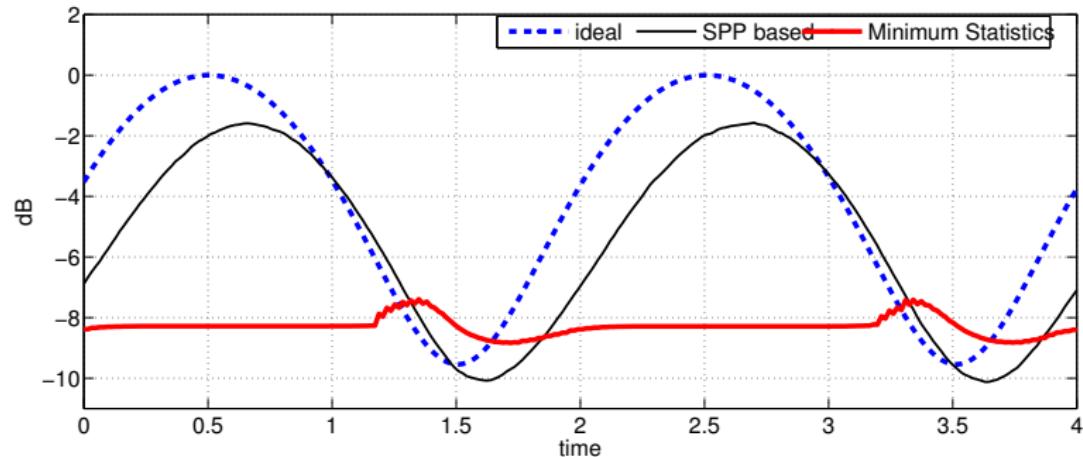
Minimum Statistics⁶

→ Search for a minimum of $|Y_k|^2$ within ≈ 1.5 s.



- Bias compensation required.
- ✓ Good estimation properties in stationary noise.
- ✓ Decreasing noise powers can be tracked quickly.
- ✗ Abruptly increasing noise powers are only tracked with a delay of 1.5 s.

⁶ R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.



noisy speech



Minimum Statistics



Based on speech presence probability (SPP)



- ✓ SPP-based method allows for fast noise tracking,
- ✓ lower computational complexity.

- Estimation of the speech PSD

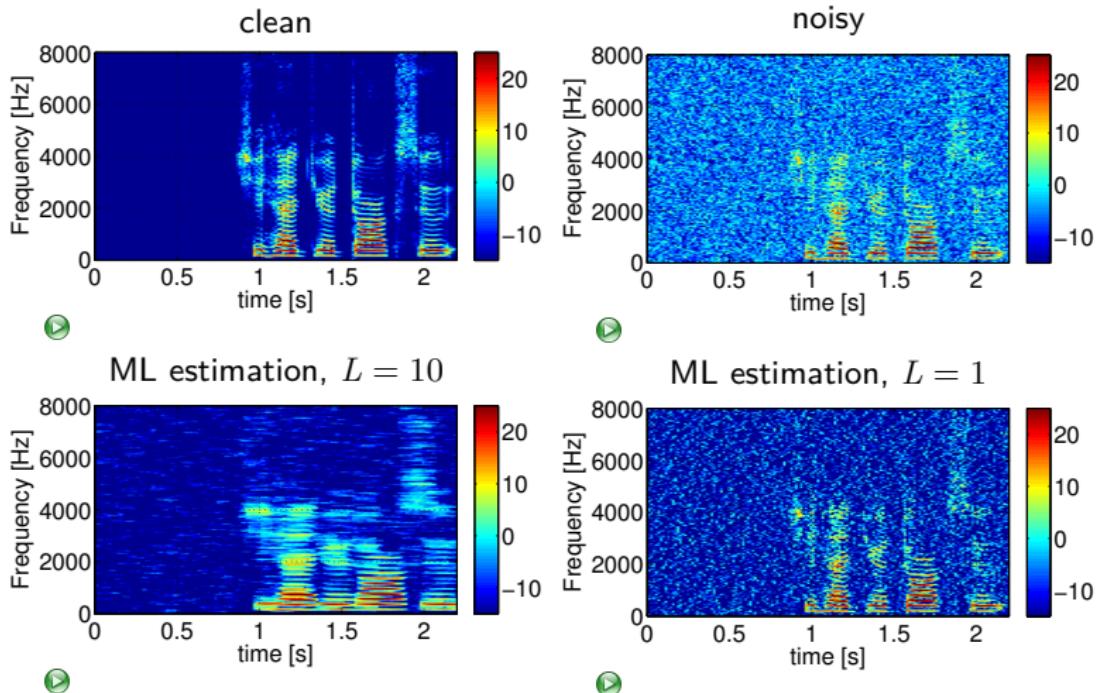
- Maximum Likelihood:

$$\begin{aligned}\sigma_{S,ML}^2 &= \arg \max_{\sigma_S^2} \prod_{n=0}^{L-1} p(Y(\ell - n) | \sigma_S^2) \\ &= \frac{1}{L} \sum_{n=0}^{L-1} |Y(\ell - n)|^2 - \sigma_N^2(\ell)\end{aligned}$$

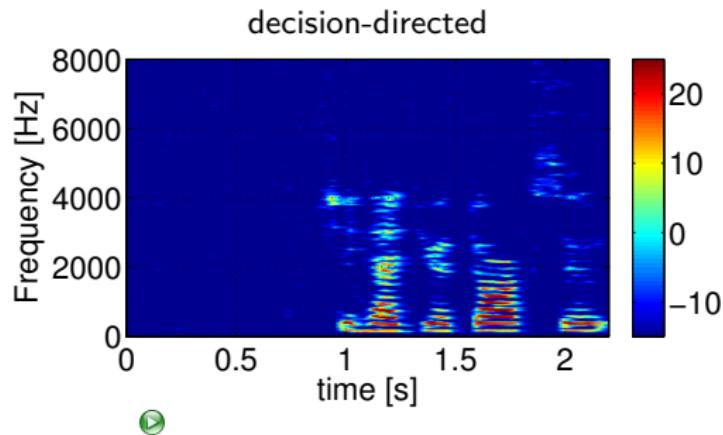
- Decision-directed approach

$$\widehat{\sigma}_S^2(\ell) = \alpha |\widehat{S}(\ell - 1)|^2 + (1 - \alpha) \max(0, |Y(\ell)|^2 - \sigma_N^2(\ell))$$

- Temporal Cepstrum Smoothing



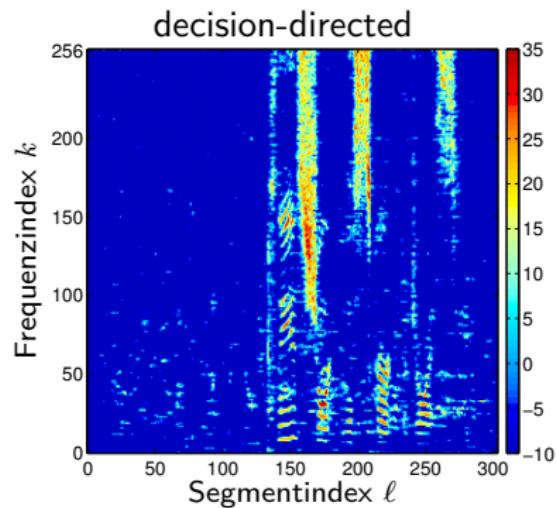
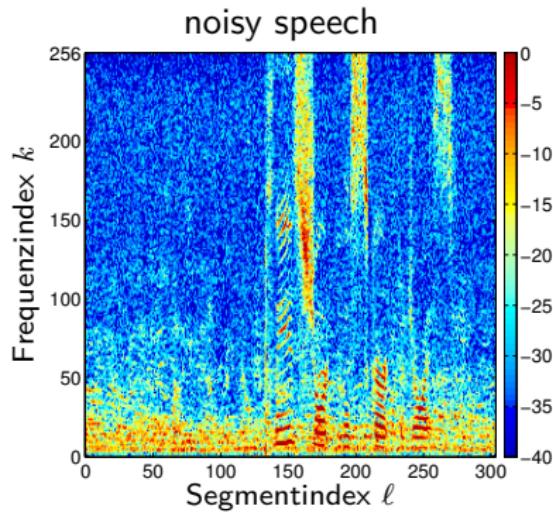
$$\widehat{\sigma}_s^2(\ell) = \alpha |\widehat{S}(\ell - 1)|^2 + (1 - \alpha) (|Y(\ell)|^2 - \sigma_n^2(\ell))$$



- ✓ Fast tracking of speech PSD.

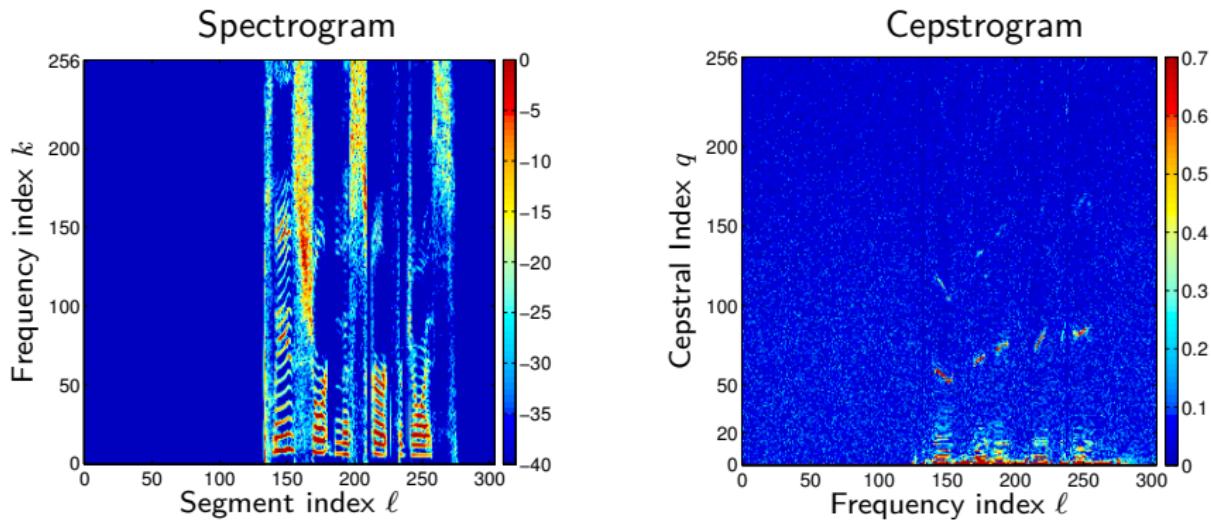
⁷ Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

Babble noise



- ✖ Outliers and speech distortions in nonstationary noise.

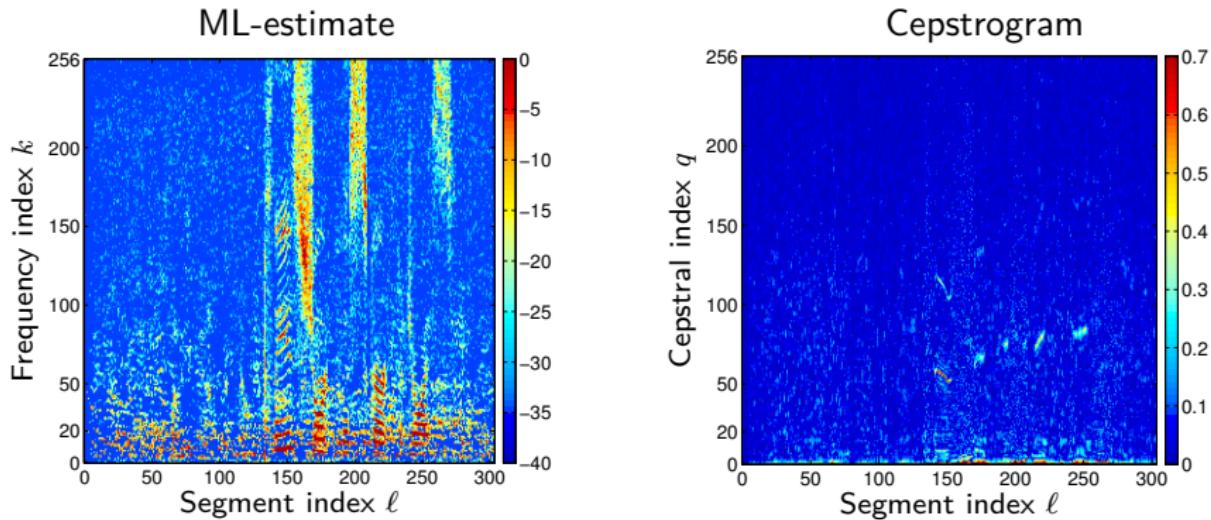
[Breithaupt, Gerkmann, Martin, 2008],⁸



Transformation to the Cepstral domain (Fourier transform of log-spectrum)

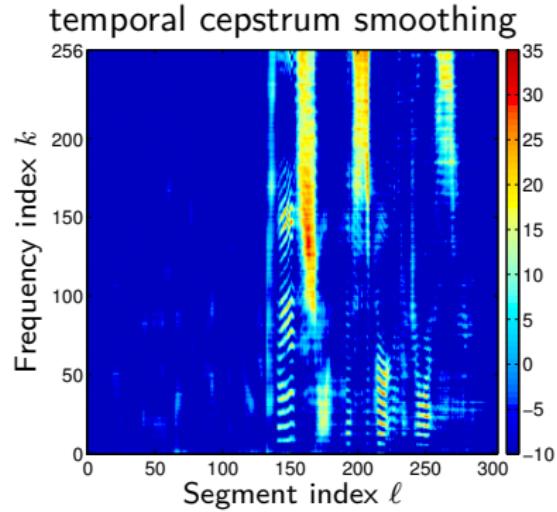
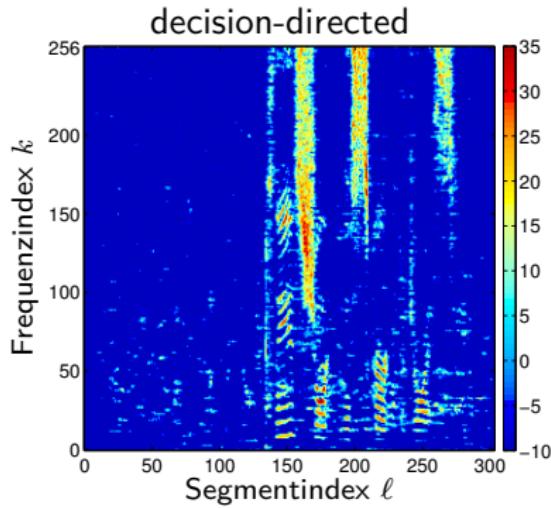
- Compact representation of speech spectral structures,
- Selective smoothing of speech and artifacts, respectively.

⁸ T. Gerkmann, "Statistical analysis of cepstral coefficients and applications in speech enhancement," Ph.D. dissertation, Ruhr-Universität Bochum, Bochum, Germany, 2010.



Selective temporal smoothing:

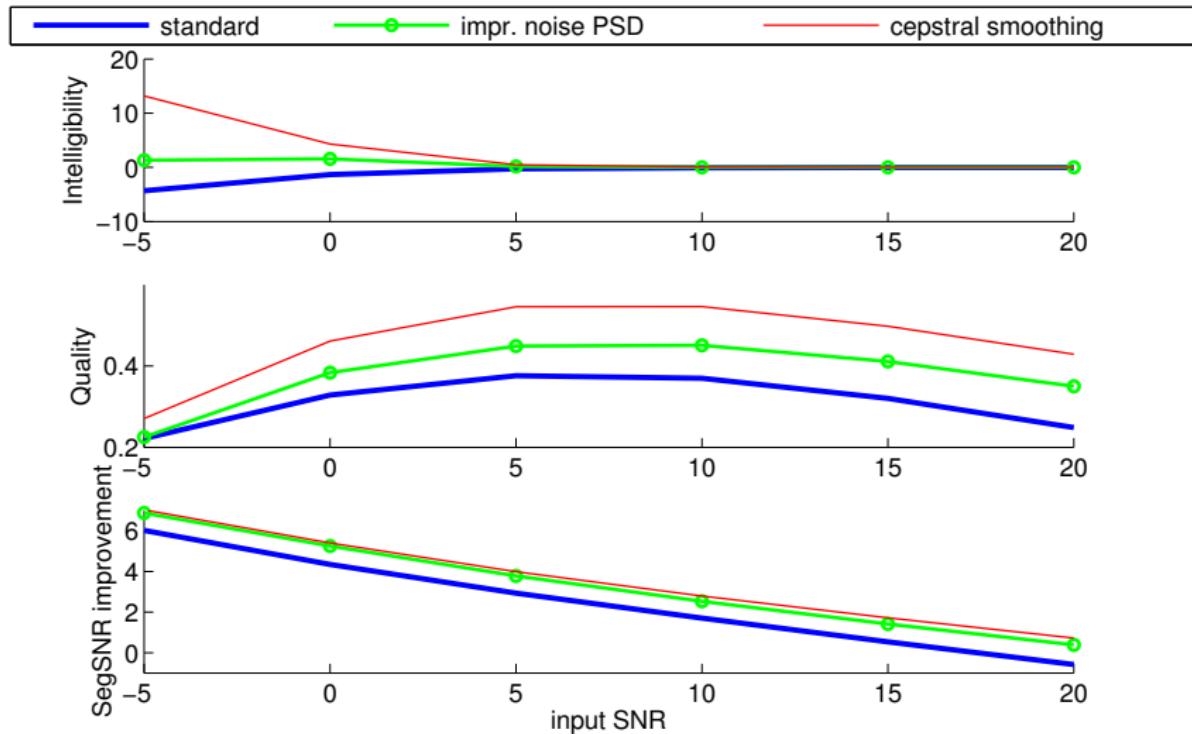
- Preserve speech related coefficients ($q \in \{0, \dots, 19, q_0\}$),
- Smooth remaining coefficients.



- ✓ Less outliers,
- ✓ more naturally sounding output,
- ✓ better preservation of speech spectral structures.

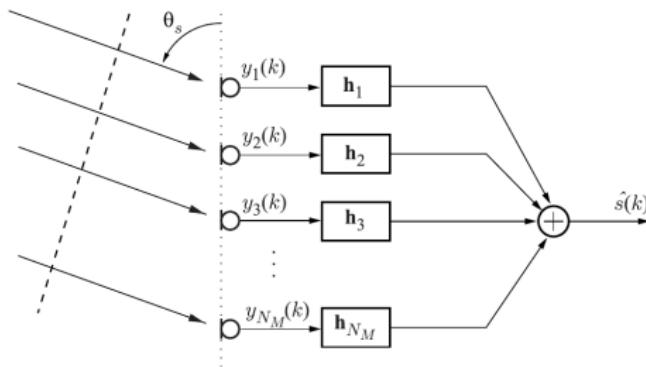
- Wiener filtering may yield annoying artifacts perceived as *musical noise*
- Musical noise can be mended by
 - Noise overestimation (at the cost of speech distortions)
 - Applying a lower limit on the gain function (at the cost of less noise reduction)
 - Smoothing techniques (e.g. temporal cepstrum smoothing)

- Less outliers in babble noise
 - noisy speech
 - ▶
 - decision directed approach / minimum statistics
 - ▶
 - SPP-based noise PSD tracking / temporal cepstrum smoothing
 - ▶
- Improved tracking of increasing noise levels
 - noisy speech
 - ▶
 - decision directed approach / minimum statistics
 - ▶
 - SPP-based noise PSD tracking / temporal cepstrum smoothing
 - ▶



1. Introduction
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
 - Single Channel Speech Enhancement
 - Multichannel Speech Enhancement
7. Machine Learning Basics
8. Source Separation

- Often, we are dealing with multidimensional or even multimodal signals
 - Usage of multiple microphones (see smartphone and laptop)
 - images and videos
 - audiovisual signal processing, ...
- We can stack these different channels into one vector to obtain the **Multichannel Wiener filter**



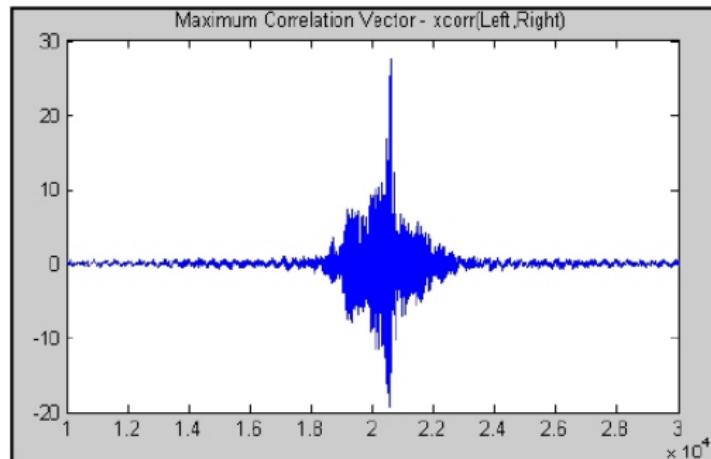
© 2006 John Wiley & Sons, Ltd
Vary, Martin - Digital Speech Transmission

Figure 12.6: Filter-and-sum beamformer

Cross correlation for source localization

- The time-delay of arrival can be estimated from a peak in the cross correlation of the microphone signals

$$\tau_0 = \arg \max_{\tau} \gamma_{y_2 y_1}(\tau) = \arg \max_{\tau} E(y_2(t)y_1(t - \tau))$$



Delay-and-sum

- The simplest beamformer design is the **delay-and-sum** beamformer
- Account for the signal delay τ_0 and average signals

$$\hat{s}(t) = \frac{1}{2} (y_1(t - \tau_0) + y_2(t))$$

- Target signal constructively added
- Interference destructively added
- Extension to M microphones straight forward
- ✓ Optimal beamformer for spatially uncorrelated noise:
→ SNR improves with $\log(M)$
- ✗ Limited noise reduction for directional (correlated) noise sources

Minimum Variance Distortionless Response (MVDR)

- Microphone observations $\mathbf{y} = [Y_1, \dots, Y_M]^T = \underbrace{\mathbf{a}S}_{\mathbf{s}} + \mathbf{n}$
- Let \mathbf{h} be a beamforming filter applied as $\widehat{S} = \mathbf{h}^H \mathbf{y}$
- Minimize the noise power without distorting the target

$$\mathbf{h}_{\text{mvdr}} = \arg \min_{\mathbf{h}} E(|\mathbf{h}^H \mathbf{n}|^2) \quad \text{s.t. } \mathbf{h}^H \mathbf{s} = S$$

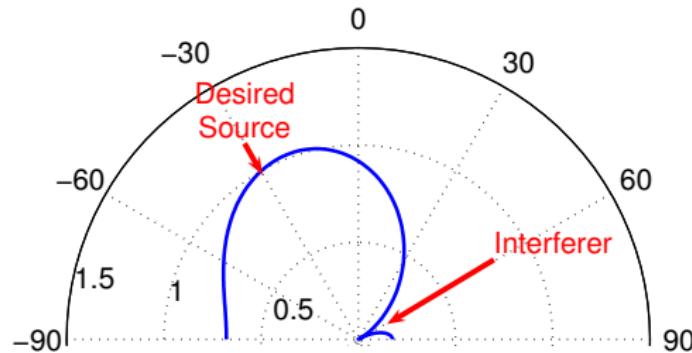
$$\rightarrow \widehat{s}_{\text{mvdr}} = \frac{\mathbf{a}^H \boldsymbol{\Phi}_{\text{NN}}^{-1} \mathbf{y}}{\underbrace{\mathbf{a}^H \boldsymbol{\Phi}_{\text{NN}}^{-1} \mathbf{a}}_{\mathbf{h}_{\text{mvdr}}^H}}$$

with $\boldsymbol{\Phi}_{\text{NN}} = E(\mathbf{n}\mathbf{n}^H)$ the noise covariance matrix

→ Noise statistics included

Minimum Variance Distortionless Response (MVDR)

- Steers main lobe on desired speaker
- Steers null on interferer(s)
 - With M sensors $M - 1$ interferers can be canceled



$$\hat{s}(n) = \sum_{m=1}^M \sum_{\nu=-\infty}^{\infty} h_m(\nu) y_m(n - \nu) \quad M : \text{number of channels}$$

Goal:

$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$

$$\hat{s}(n) = \sum_{m=1}^M \sum_{\nu=-\infty}^{\infty} h_m(\nu) y_m(n - \nu) \quad M : \text{number of channels}$$

Goal:

$$\min \left(E \left((s(n) - \hat{s}(n))^2 \right) \right)$$

Solution:

$$\sum_{m=1}^M \sum_{\nu=-\infty}^{\infty} h_m(\nu) \varphi_{y_m, y_m}(i - \nu) = \varphi_{y_m, s}(i)$$

$$\hat{s}(n) = \sum_{m=1}^M \sum_{\nu=-\infty}^{\infty} h_m(\nu) y_m(n-\nu) \quad M : \text{number of channels}$$

Goal:

$$\min(E((s(n) - \hat{s}(n))^2))$$

Solution:

$$\sum_{m=1}^M \sum_{\nu=-\infty}^{\infty} h_m(\nu) \varphi_{y_m, y_m}(i-\nu) = \varphi_{y_m, s}(i)$$

Transformed into Fourier domain and using vector notation:

$$\sum_{m=1}^M H_m(e^{j\Omega}) \Phi_{Y_m Y_{m'}} = \Phi_{Y_{m'}, S} \quad \text{for } m' = 1, \dots, M$$

$$\Phi_{YY} \mathbf{h} = \Phi_{YS}$$

Multichannel Wiener

$$\mathbf{h} = \Phi_{YY}^{-1} \Phi_{YS}$$

- For the multichannel Wiener filter we have

$$\mathbf{h} = \boldsymbol{\Phi}_{\mathbf{YY}}^{-1} \boldsymbol{\Phi}_{\mathbf{YS}}$$

- Assuming speech and noise are uncorrelated, $\boldsymbol{\Phi}_{\mathbf{YS}} = \boldsymbol{\Phi}_{\mathbf{SS}}$
- Given the propagation vector \mathbf{a} , we can write $\boldsymbol{\Phi}_{\mathbf{SS}} = \mathbf{a}\boldsymbol{\Phi}_{\mathbf{SS}}\mathbf{a}^H$

$$\mathbf{h} = (\mathbf{a}\mathbf{a}^H \boldsymbol{\Phi}_{\mathbf{SS}} + \boldsymbol{\Phi}_{\mathbf{NN}})^{-1} \boldsymbol{\Phi}_{\mathbf{SS}}\mathbf{a}$$

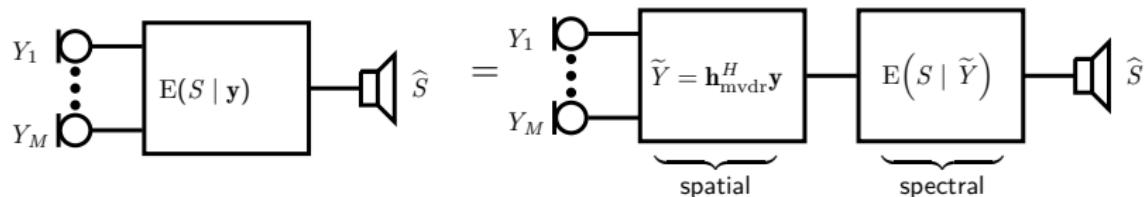
- Woodbury's matrix identity, $\mathbf{A} = \boldsymbol{\Phi}_{\mathbf{NN}}$, $\mathbf{U} = \boldsymbol{\Phi}_{\mathbf{SS}}\mathbf{a}$, $\mathbf{C} = 1$, $\mathbf{V} = \mathbf{a}^H$

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

■ 9

$$\mathbf{h} = \underbrace{\frac{\Phi_{NN}^{-1}\mathbf{a}}{\mathbf{a}^H\Phi_{NN}^{-1}\mathbf{a}}}_{\text{spatial filtering (MVDR beamformer)}}$$

$$\underbrace{\frac{\Phi_{SS}}{(\mathbf{a}^H\Phi_{NN}^{-1}\mathbf{a})^{-1} + \Phi_{SS}}}_{\text{single channel NR (Wiener filter)}}$$



→ Improving single channel speech enhancement techniques also improves multi-channel approaches!

⁹ K. Uwe Simmer, Joerg Bitzer, and Claude Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., Berlin, Heidelberg, New York: Springer-Verlag, 2001, pp. 39–60.

1. Introduction
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. Machine Learning Basics**
8. Source Separation
9. Automatic Speech Recognition



Universität Hamburg

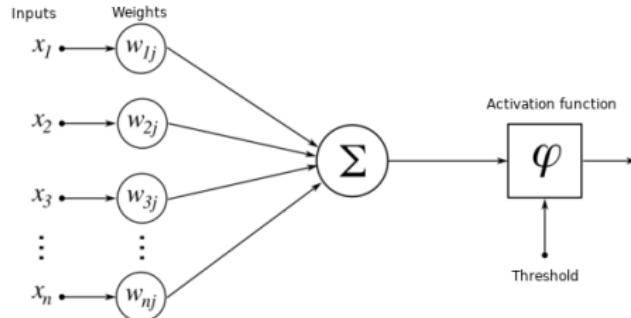
DER FORSCHUNG | DER LEHRE | DER BILDUNG



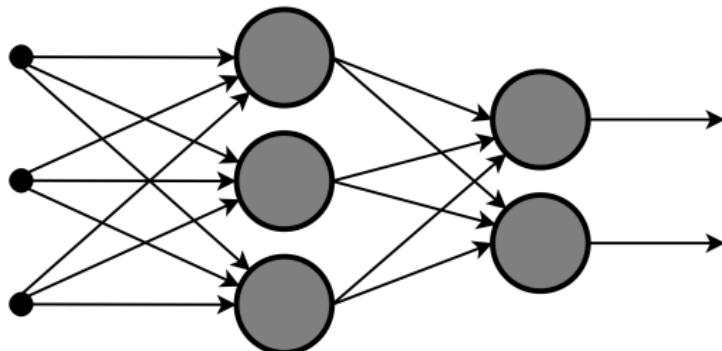
Machine Learning Basics

- Machine Learning (ML) algorithms identify structure and patterns in data and use these to infer solutions
- Solutions are data-driven
- ML techniques are essential for many challenging tasks in speech processing
 - Speech separation of two speaker with one microphone only became tractable with these methods
 - In general, ML techniques can exploit tempo-spectral properties for improved speech processing

- A certain class of machine learning methods, deep neural networks, are of particular relevance for the recent success of modern machine learning methods
- Deep neural networks consists of multiple artificial neurons
- Artificial neurons are remotely inspired by the neurons in the human brain
- From a signal processing perspective the consist of a weighted addition of input signals followed by a nonlinear *activation function*



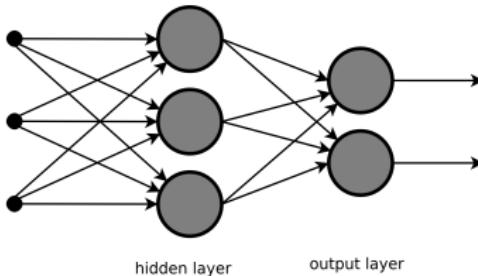
- If multiple artificial neurons (nodes) are connected, we call this a neural network
- If several hidden layers are present: **Deep Neural Network**
- If between all layers all neurons are connected: **Fully connected** neural network
- Signal processing perspective: Signal flow from input to output with concatenated linear (weights, biases) and nonlinear (activation) functions, all of which are differentiable
- The free parameters (e.g. weights, biases) can be learned/optimized given training data



- Neural Networks can be used to approximate any nonlinear function
- The more complex the function is, the more learnable parameters do we need
- This could in principle be achieved with only one hidden layer but many nodes
- Using a deep neural network with many layers has been shown to be more easy to train and handle

- While Artificial Neural Networks have been proposed decades ago, their breakthrough by means of Deep Neural Networks really came in the mid 2010's
- This is mainly because training neural networks is not trivial
- Their success is based on
 - More training data
 - More computational power for parallelized training (GPUs)
 - Efficient training mechanisms (e.g. ADAM)
 - Simplified activation functions (e.g. ReLu)
 - Efficient and accessible software tools (Pytorch, Tensorflow, Keras, ...)

- Training involves a forward and a backward pass
- The *forward pass* means to compute the output for a given input
- The *backward pass* propagates the error between the output and a target backwards through the network and adjusts the weights so that the network output for this example is closer to the target

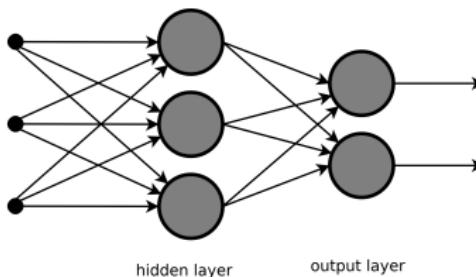


Forward pass

- In the forward pass, the weights and thresholds determined in the previous training pass are used
- At each neuron, the weighted sum of the inputs is computed, the bias is added and the activation function is applied

$$h(\mathbf{w}^T \mathbf{x} + b)$$

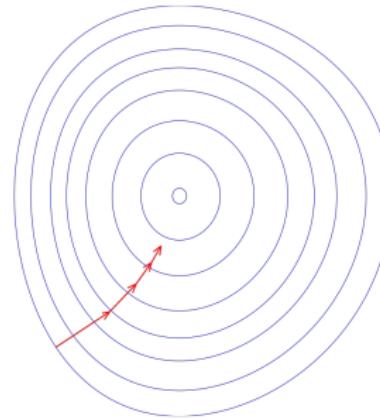
- The outputs of each layer are then fed to the next layer



Backward pass

- In the backward pass, our goal is to change the network parameters (weights and biases) such that the network output would have been closer to the target
- If we consider one layer, the weights can be updated using gradient descent by taking the partial derivatives of the loss function \mathcal{L} with respect to the parameter that is to be adapted
- We may start e.g. with random weights and then adapt
- E.g. for two weights we have

$$\begin{aligned} w_1 &:= w_1 - \alpha \frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_1} \\ w_2 &:= w_2 - \alpha \frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_2} \end{aligned}$$



with

Backward pass

- **Backpropagation** is used to propagate the loss/error backwards through the network
- This allows us to compute the partial derivative of the loss function with respect to each individual weight in the networks
- Here the chain rule of derivation is used to take derivatives of the composite function given by the neural network structure
- E.g. for two composite functions the chain rule is given by

$$(h(x))' = (g(f(x)))' = g'(f(x)) \cdot f'(x)$$

- In backpropagation, the chain rule is recursively applied for updating the weights in a neural network

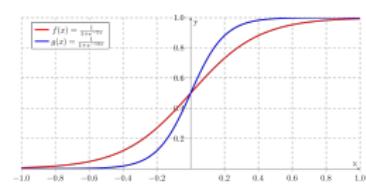
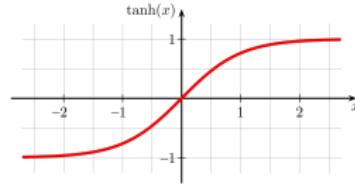
Summary

1. Initialize weights and biases
2. Input a training vector, and compute the output and loss function (forward pass)
3. Backpropagate to compute the partial derivatives with respect to each parameter of the network (weights and biases)
4. Update parameters using the partial derivatives and learning rate α

- If the entire training dataset was used to compute the average loss and gradients *per parameter update* this would often be computationally infeasible for large amounts of training data
- Instead, in stochastic gradient descent, the training data is randomly grouped into *minibatches*
- Parameter updates are computed for each minibatch until the whole dataset has been processed (one epoch)
- This method allows to use large amounts of training data while remaining computationally feasible
- Networks are usually trained for many epochs

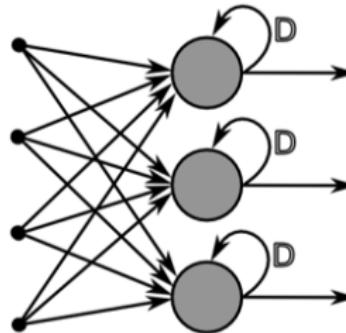
- The loss function (error function) is used to measure how close the output of the network is to a given target
- The mean squared error we used for speech enhancement can be interpreted as a loss function
- The performance and behavior of a network strongly depends on the chosen loss function
- The loss function has to be wisely chosen, just as the network architecture and activation functions
- Besides MMSE, also relative and cross entropies may serve as loss functions.
- Also regularization using e.g. the ℓ_1 -norm are common

- The activation function is what makes the DNN nonlinear, and thus enables approximating arbitrary nonlinear functions
- The choice depends again on the problem at hand (e.g. classification versus regression)
- The rectified linear unit is a common choice that has been shown to result in robust training results.
- Other choices are tanh or generalized sigmoidal functions



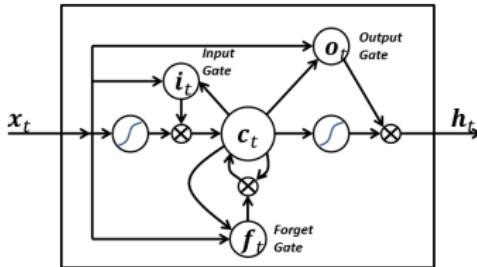
Recurrent Neural Networks

- The network architecture is an important choice and its optimization an often subject to trial and error
- For speech signals it is of particular interest to take into account temporal correlations
- Similar to an Autoregressive model, recursive structure can be employed to model temporal dependencies using **Recurrent Neural Networks (RNN)**



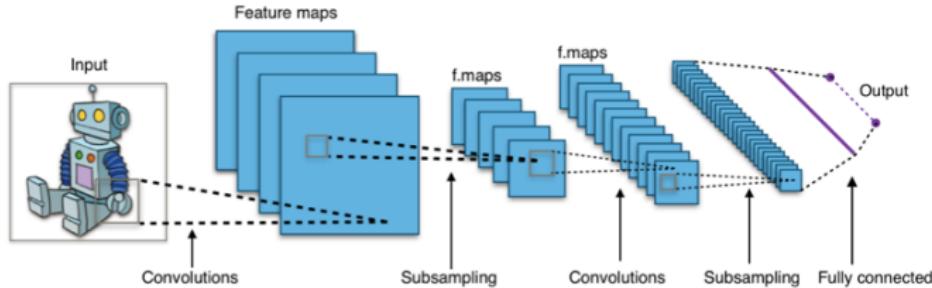
Long short-term memory (LSTM)

- When an error is propagated through an RNN, the error term may get either smaller and smaller or higher and higher
- This may yield either vanishing or exploding gradients, which are problematic when training DNNs
- For instance, for vanishing gradients, the respective neurons are not/hardly updated anymore
- The LSTM solves this by introducing so-called forget gates, allow to pass or erase information, which helps to avoid vanishing/exploding gradients



Convolutional Neural Network (CNN)

- In the CNN network architecture convolutional filters are learned
- The convolutional layers are frequency followed by pooling layers for dimension reduction



- Without activations functions, these filters may resemble filterbanks, such as the STFT

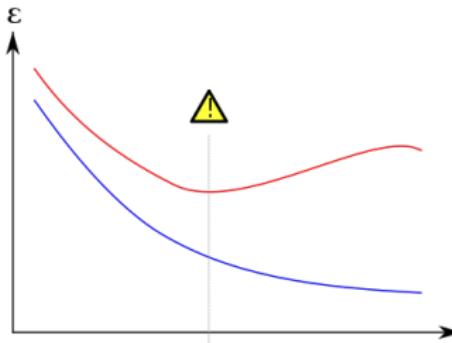
- When designing a network, one has to make sure that the amount of available training data is sufficient for the amount of parameters that have to be trained
- If the number of parameters is too high, or the amount of available data is too low, **overfitting** is likely to occur
- Overfitting means that the networks learns to yield good results for the given training set, but fails for unseen data
- Overfitting may be reduced e.g. by
 - More training data
 - Reducing model capacity
 - Early stopping
 - Regularization

- The data used to train and evaluate is split into three strictly non-overlapping parts

Training This is used to train the model

Validation Per epoch, the performance is validated using the validation set. If the loss on the validation set increases, this can be a sign for overfitting. Training should be stopped at this point (early stopping)

Test This dataset is never seen during training or validation. It is used to evaluate the performance of the trained model after the hyperparameter selection stage.



1. Introduction
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. Machine Learning Basics
8. Source Separation
9. Automatic Speech Recognition

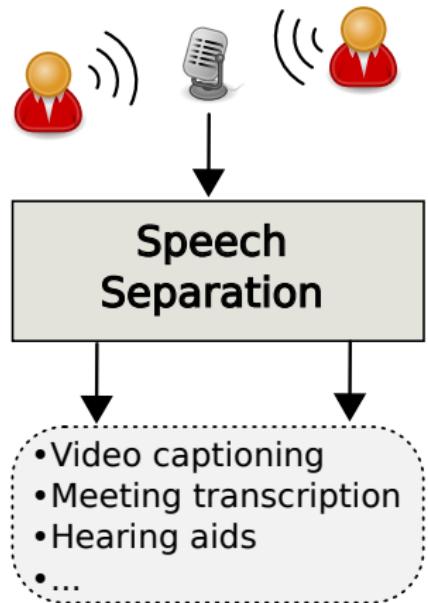


Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Source Separation

How to solve the cocktail party problem

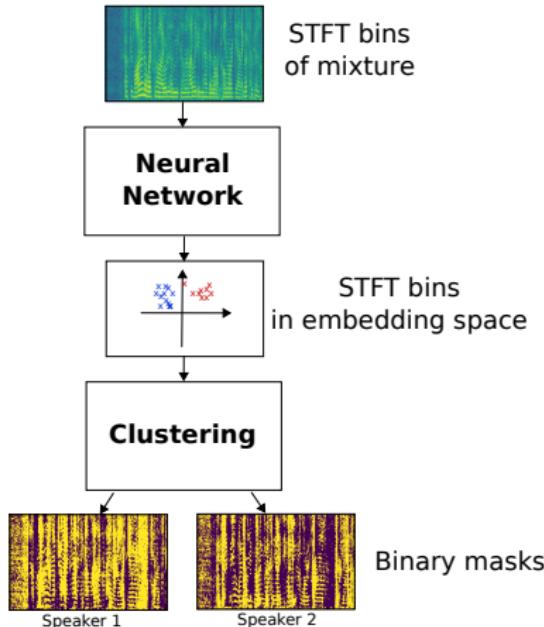


- In source separation, the target and interference have similar properties
- Machine learning algorithms are required

Conditions:

- Undefined number of speakers
- Unknown speakers
- Single microphone

Deep Clustering Approach^[12]:



Technical details:

- Recursive network architecture
- Usually 2 to 4 LSTM layers
- 200 to 400 units per LSTM layer
- Fully connected layer into embedding space
- Algorithmic latency defined by STFT windowing (32 ms usually)

Performance:

- Allows for at least 6 dB SDR average improvement (online processing)

[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative Embeddings for Segmentation and Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35.

- $X_i = X_{t,f}$, with $i \in \{1, \dots, N\}$ is the considered input STFT magnitude spectrum with N STFT bins
- Sparsity assumption: TF-bins can be partitioned into sets of TF bins dominated by one speaker, respectively
- If the corresponding sets are found, individual speakers can be recovered by taking the ISTFT of the TF-bins of the corresponding set, while setting the remaining STFT-bins to zero
- The target partition in a given mixture is represented by the indicator $Y = \{y_{i,c}\}$, i.e. $y_{i,c} = 1$ if element i is in cluster c (i.e. TF bin i belongs to speaker c)
- From this indicator function an affinity matrix $A = YY^T$ can be obtained with

$$A_{i,j} = \begin{cases} 1 & \text{if elements } i \text{ and } j \text{ belong to the same cluster/speaker} \\ 0 & \text{otherwise} \end{cases}$$

- Example: The TF-bins in X contains speakers a and b in order $\{aabab\}$
 - The indicator would result in ($y_{i,c} = 1$ if element i is in cluster c)

$$Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad Y^T = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

- For the affinity matrix ($A_{i,j} = 1$ if i, j belong to same cluster)

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

- Goal: find a D -dimensional embedding $V_\theta = f(x) \in \mathbb{R}^{N \times D}$ with DNN parameters θ
- The embeddings have unit norm $|v_i|^2 = 1$ and are implicitly assumed to approximate the $N \times N$ affinity matrix $\hat{A} = VV^T$
- The loss function is designed such that the ideal affinities $A = YY^T$ are approximated by the learned embeddings $\hat{A} = VV^T$ by means of the Frobenius norm as

$$\mathcal{L}_Y(V) = \|\hat{A} - A\|_F^2 = \|VV^T - YY^T\|_F^2$$

- However, as the affinity matrices are of size $N \times N$ with N the number of STFT bins, this loss function is computationally very heavy
- With matrix calculus, due to the low rank structure of the affinity matrices (e.g. $A = YY^T$), we can rearrange to obtain the efficient implementation

$$\mathcal{L}_Y(V) = \|V^T V\|_F^2 - 2\|V^T V\|_F^2 + \|Y^T Y\|_F^2$$

- Thus the Frobenius norms are not taken over all N^2 pairs of STFT bins, but only over D^2 bins
- The derivatives with respect to V are also efficiently obtained due to the low-rank structure

$$\frac{\partial \mathcal{L}_Y(V)}{\partial V^T} = 4V(V^T V) - 4Y(Y^T V)$$

- The derivation follows using the following properties
see Petersen, Pedersen, "The Matrix Cookbook", Version Nov 15, 2012

$$(9) \quad (AB)^H = B^H A^H$$

$$(14) \quad \text{Tr}(AB) = \text{Tr}(BA)$$

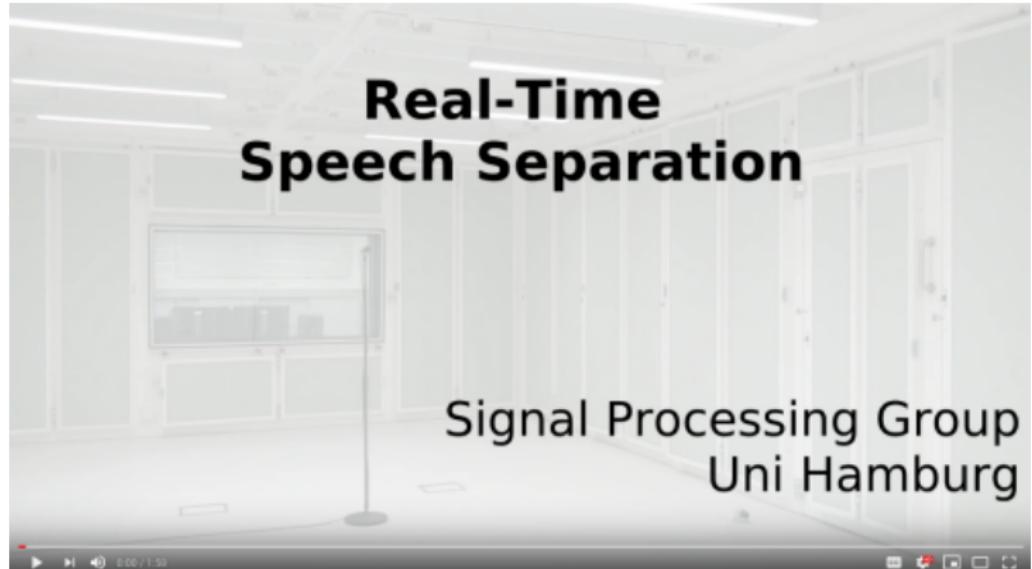
$$(15) \quad \text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$$

$$(16) \quad \text{Tr}(ABC) = \text{Tr}(CAB)$$

$$(132) \quad \frac{\partial}{\partial X} \|X\|_F^2 = \frac{\partial}{\partial X} \text{Tr}(XX^H) = 2X$$

$$(541) \quad \|A\|_F^2 = \text{Tr}(AA^H)$$

$$\begin{aligned} \mathcal{L}_Y(V) &= \|VV^T - YY^T\|_F^2 \\ &\stackrel{(541)}{=} \text{Tr}((VV^T - YY^T)(VV^T - YY^T)^T) \\ &\stackrel{(9)}{=} \text{Tr}(VV^T VV^T + YY^T YY^T - YY^T VV^T - VV^T YY^T) \\ &\stackrel{(14,15,16)}{=} \text{Tr}(VV^T VV^T) + \text{Tr}(YY^T YY^T) - 2\text{Tr}(V^T YY^T V) \\ &\stackrel{(541)}{=} \|V^T V\|_F^2 + \|Y^T Y\|_F^2 - 2\|V^T Y\|_F^2 \end{aligned}$$

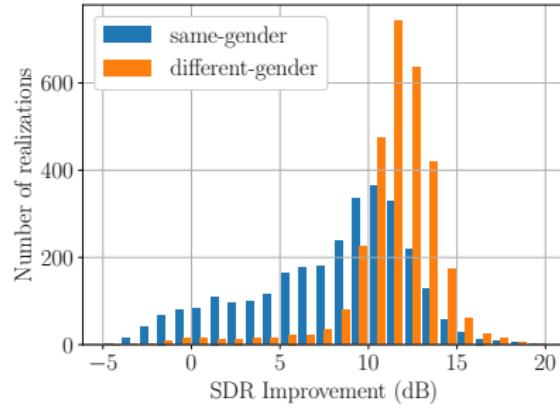


[13] [14]

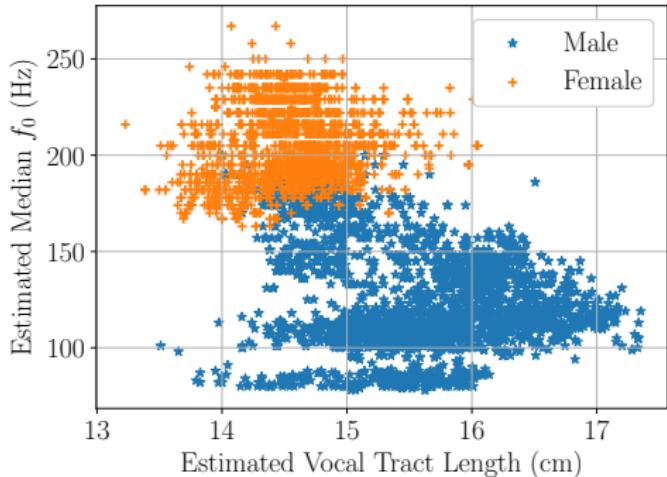
[13] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

[14] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," en, in *ISCA Interspeech*, Graz, Austria, Sep. 2019, pp. 4584–4588. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2459.html (visited on 09/16/2019).

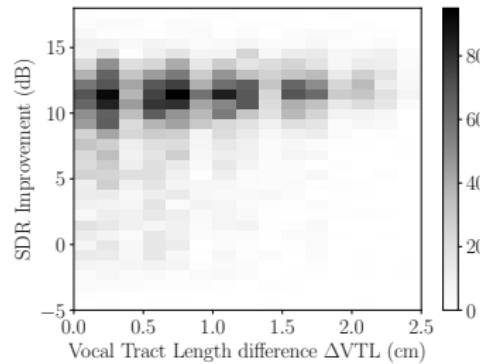
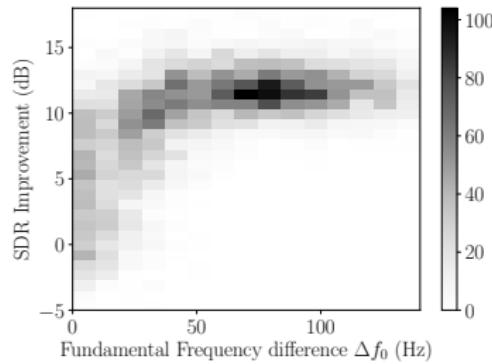
- ✓ Shown system^[15] shows good average performance
- ✗ But: Performance varies for certain speaker constellations



[15] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 686–690.



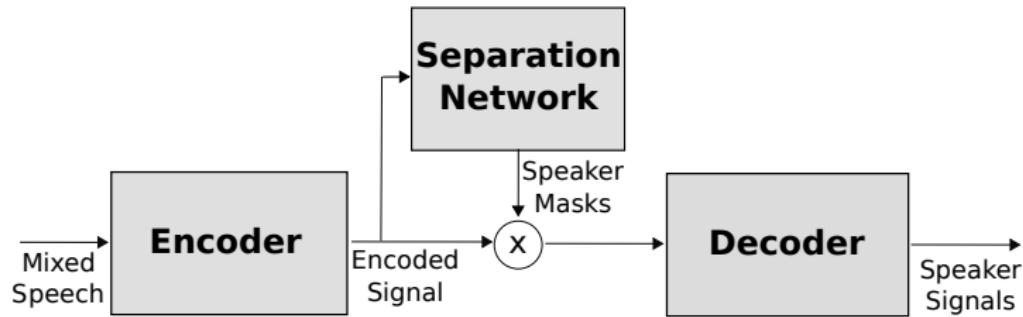
- On average, females have shorter vocal tract lengths and higher fundamental frequencies
- Vocal tract length: Changes spectral envelope
- Fundamental frequency: Changes spectral fine structure



- Measure performance for speaker pairs as a function of the difference in pitch and vocal tract length
 - Fundamental frequency difference Δf_0 is the dominant factor to predict separation quality^[14]
 - For speaker pairs with close f_0 , source separation may be harmful

[14] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," en, in *ISCA Interspeech*, Graz, Austria, Sep. 2019, pp. 4584–4588. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2459.html (visited on 09/16/2019).

Conv-TasNet approach^{[16][13]}

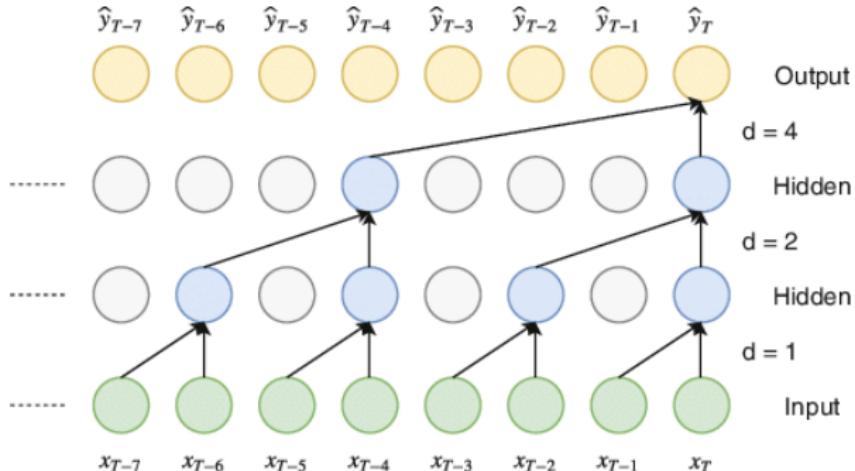


- Encoder and decoder are learned convolutional layers (i.e. filterbanks)
- Algorithmic latency defined by encoder window size
- Filterbank windows can be very small (e.g. ≤ 2 ms)
- Receptive field around 1 to 2 s

[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[13] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

Separation Network

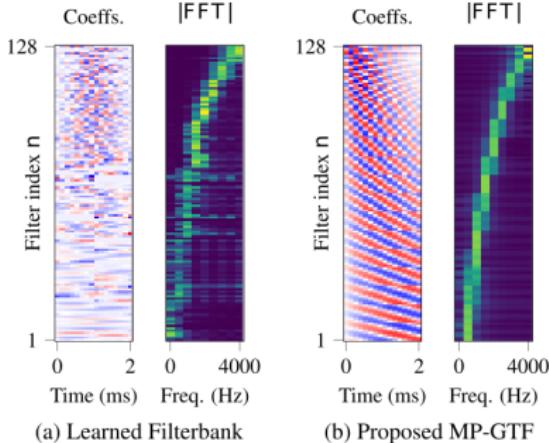


- Separation network is fully convolutional including non-linearities
- Use of dilated convolutions to enlarge receptive field
- Use of skip connections for easier training^[16]

[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

Learned Filterbank: Key to TasNet's Success?

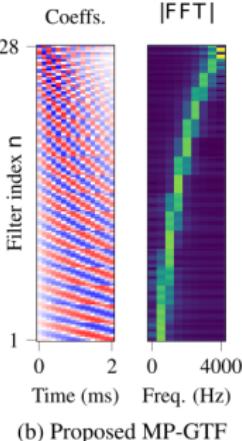
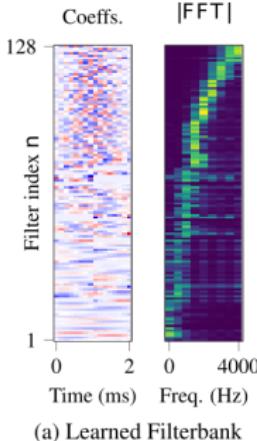
Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)^[13]:



[13] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

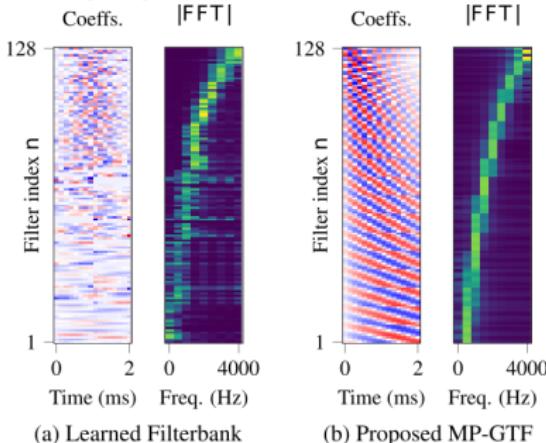
Learned Filterbank: Key to TasNet's Success?

Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)^[13]:



Encoder / Filterbank	N	SI-SNRI (dB)
Learned	512	15.4
Learned	128	15.2
MP-GTF	512	15.9
MP-GTF	128	16.1

Learned Filterbank: Key to TasNet's Success?

Our proposal: Multi-Phase Gammatone Filterbank (MP-GTF)^[13]:

Encoder / Filterbank	N	SI-SNRi (dB)
Learned	512	15.4
Learned	128	15.2
MP-GTF	512	15.9
MP-GTF	128	16.1

- Motivation: Resembles human auditory system and structure of fully-learned encoder.
- ✓ Speeds up training time (less parameters)
- ✓ Slightly outperforms fully learned filterbanks

[13] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 36–40.

1. Introduction
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. Machine Learning Basics
8. Source Separation
9. Automatic Speech Recognition



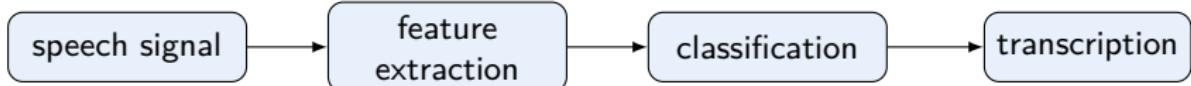
Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Automatic Speech Recognition

- | | |
|---------------------|---|
| Applications | <ul style="list-style-type: none">■ Hands-free control (GPS-system, smart-phone, ...)■ Dictation (e.g. SMS, note, letters, ...)■ Speech dialog systems (e.g. train schedule)■ Content analysis (e.g. labeling or subtitles for YouTube videos)■ Surveillance. |
| Principle | <ul style="list-style-type: none">■ Compare input speech to pre-trained models (e.g. templates, statistical models)■ Best matching model (template) is the recognized phoneme/word/sentence. |
| Challenges | <ul style="list-style-type: none">■ For the comparison, adequate features and models of speech are required that are not sensitive to<ul style="list-style-type: none">■ Variations in pronunciation (also: dialects, accents, ...)■ Variations in tempo (fast vs. slowly spoken speech)■ Different speakers■ Robustness to noise and reverberation. |



- Features:**
- Contain relevant information to discriminate phonemes
 - Should be compact, i.e. irrelevant information should be discarded.
 - Often: Mel-frequency cepstral coefficients (MFCCs)

- Classification:**
- Comparison of observed features to pre-trained models.
 - Simplest way: Euclidean distance between observed features and templates
 - Often: statistical representation of feature variations and sequential structure using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) or Deep Neural Networks (DNNs).

Continuity In spoken language, often there is no clear separation between words.

Ambiguities "How to recognize speech" versus "How to wreck a nice beach".

Intrinsic variabilities Same speaker, different speaker, gender, age (child, grown-up, elderly), emotions, speed, pronunciation (accents, dialects)

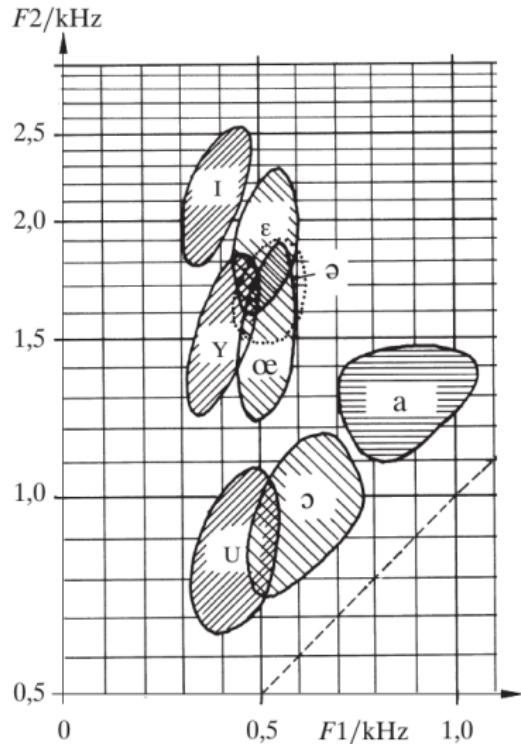
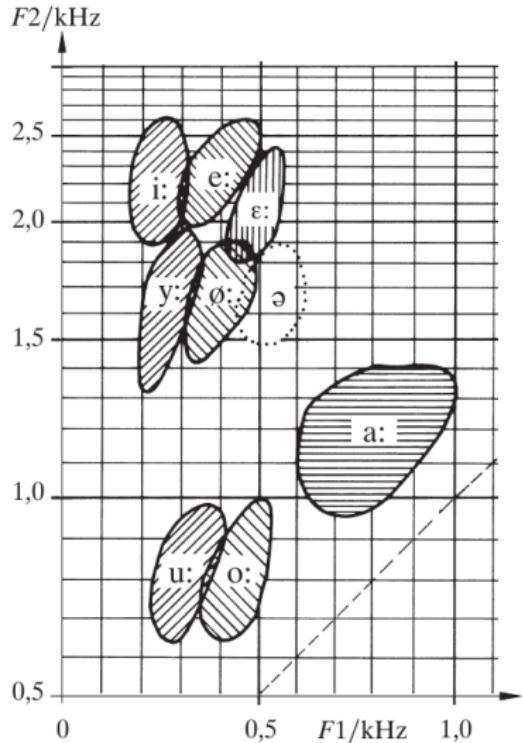
Extrinsic variabilities Noise, reverberation, transmission channel (e.g. telephone).

Ideal features are

- **Sensitive** to discriminating different phonemes
- **Insensitive** to

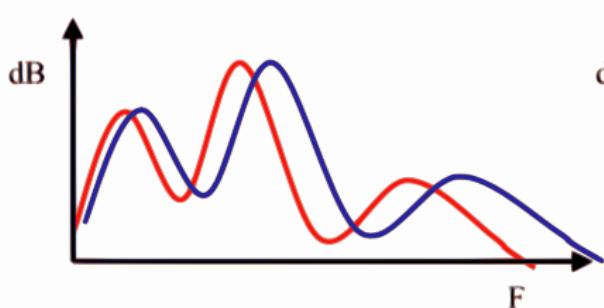
Ideal features are

- **Sensitive** to discriminating different phonemes
- **Insensitive** to
 - Discriminating allophones
 - Intrinsic variabilities
 - Extrinsic variabilities



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

- For different speakers spectral shapes of phonemes are modified (e.g. stretched) due to different length and shape of the vocal tract.
- Differences in recording level
- Noise and reverberance make spectra less similar

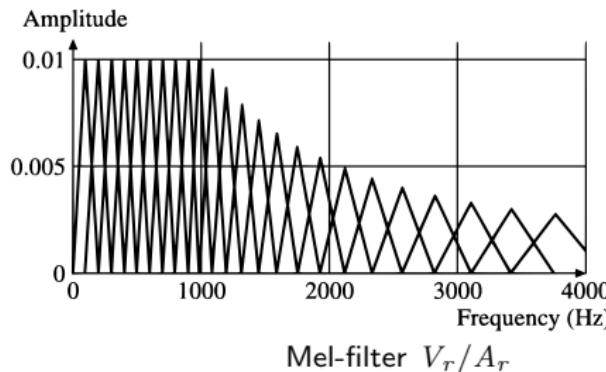


Motivation: model spectral resolution of the human auditory system

- After taking the STFT, comprise frequency bands within *mel-bands*

$$M_r(\ell) = \frac{1}{A_r} \sum_{k=L_r}^{U_r} V_{r,k} |X_k(\ell)|^2$$

where $A_r = \sum_{k=L_r}^{U_r} V_{r,k}$ is a normalization factor. Often: 24 bands.



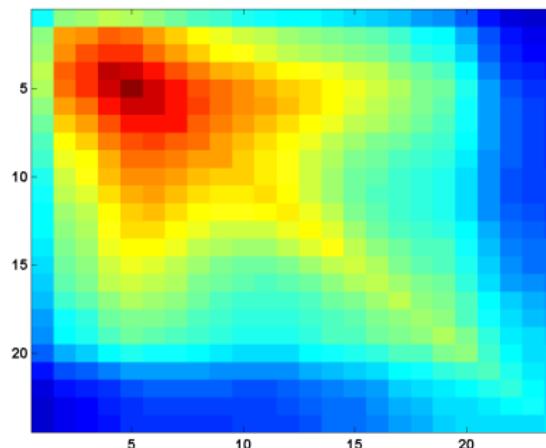
- Take discrete cosine transform of the log-magnitude of the mel-filter outputs.

$$\text{MFCC}_q(\ell) = \text{IDCT} \{ \log M_r(\ell) \}$$

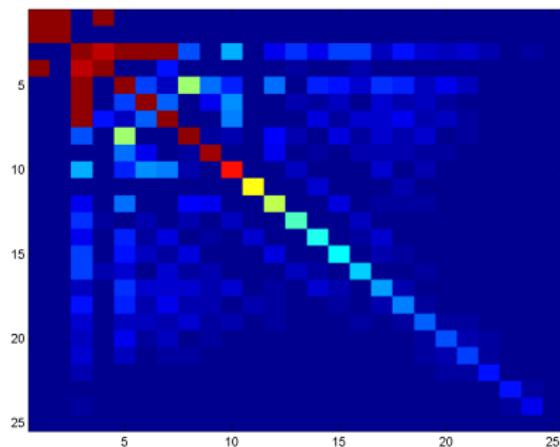
- Typically only the lowest 13 MFCCs are employed (spectral envelope).
 - In addition: ΔMFCC and $\Delta\Delta\text{MFCCs}$ (temporal differences).
- The MFCCs (except the zeroth) are level-independent.
- The DCT decorrelates the feature coefficients. Decorrelation simplifies the statistical modelling (diagonal covariance matrices)



Correlation of Mel Filterbank Coefficients

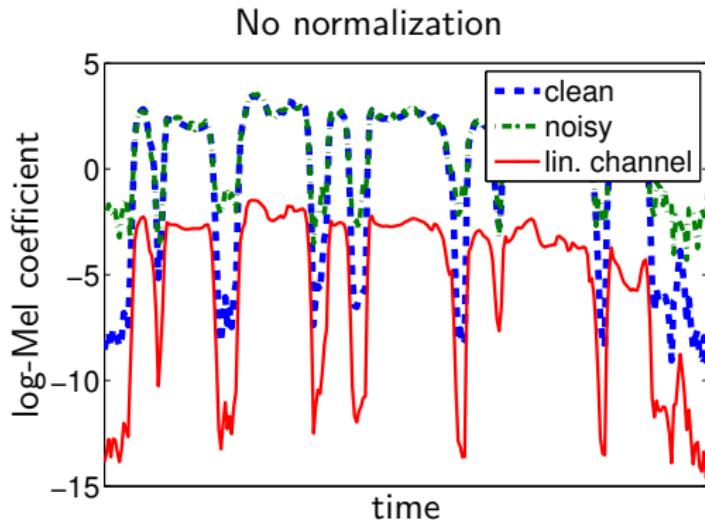


Correlation of MFCCs

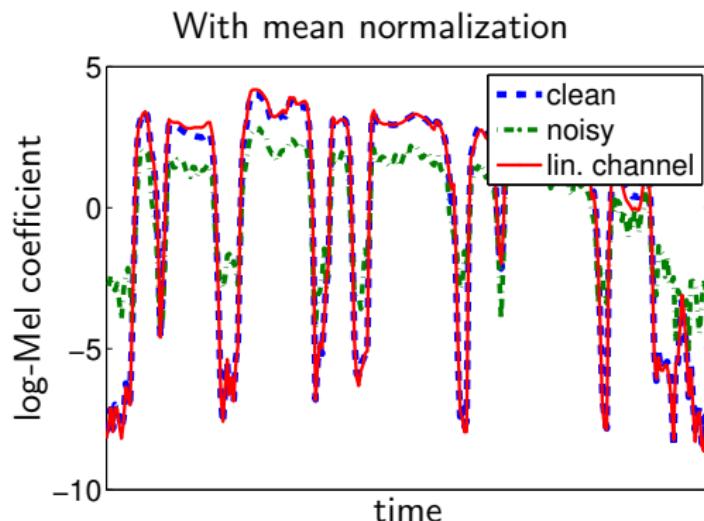


- MFCCs are less correlated than Mel filterbank coefficients.
- The DCT decorrelates the feature coefficients. Decorrelation simplifies the statistical modelling (diagonal covariance matrices)

- Noise reduction prior to computing features
- Feature enhancement (e.g. cepstral mean and variance normalization)
- Multi-condition training
 - When creating reference models, also employ noisy and reverberant data
 - Training and/or testing is computationally more complex
 - Robustness can be greatly improved.

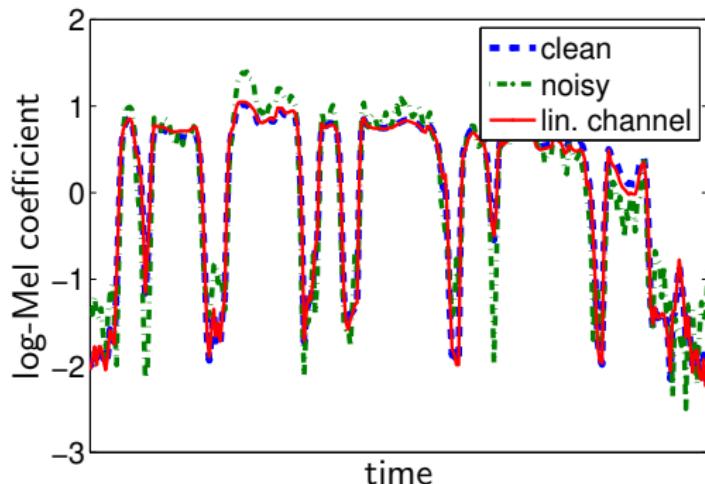


- Noise reduces the empirical variance of the feature.
- A linear channel results in a linear offset in the log-mel features.



- Mean normalization reduces the effect of a linear channel.
- The noisy feature sequence is still quite different from the clean feature sequence.

With both mean and variance normalization



- The variance normalization also makes the feature sequence for the noisy sequence more similar to the clean feature sequence.
- Increased robustness for recognizing noisy and filtered speech.

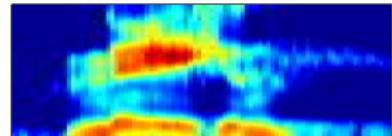
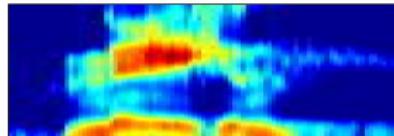
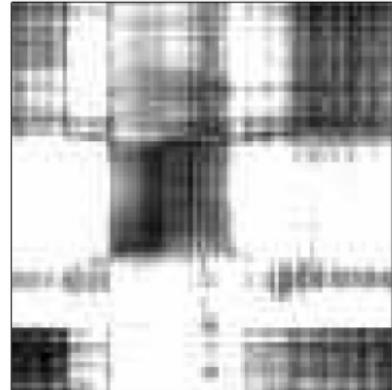
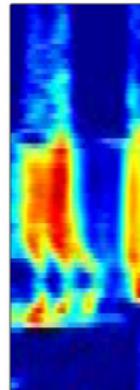
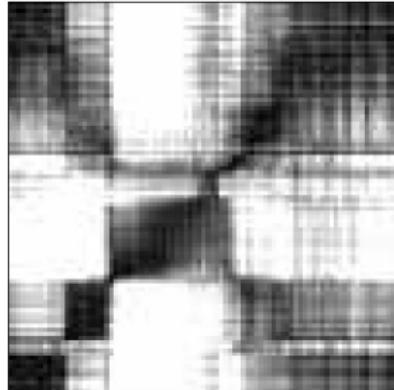
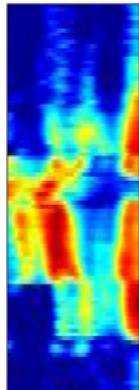
Compare features of observed input (testing signal) against trained (reference) models/features

- Example: recognition of the digits 0...9
 - One model for each number (i.e. 10 models)
 - Compare observed (test) data against each of the reference models
 - Most similar reference model is chosen as the recognized digit.

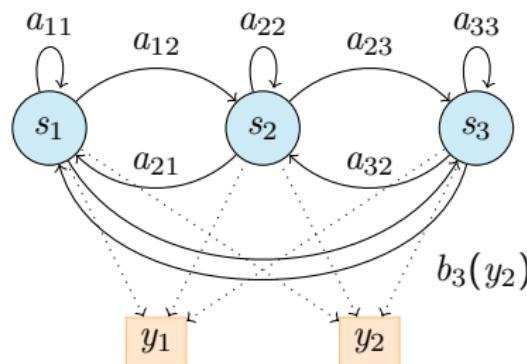
- Simple approach: for a single time-frame: compare each feature vector to all reference vectors, e.g. by computing the Euclidean distance between observed features and the reference features
- Gaussian Mixture Model (GMM)
 - Training: learn the probability distribution of each phoneme as the superposition (mixture) of several multidimensional Gaussians.
 - Recognition: for each (phoneme-)model find the probability that the observed feature vector was generated by that model.

- Find best match between the observed *sequence* of feature vectors and the reference models.
- Simple approach: Dynamic time warping (DTW)
 - Temporal warping of test and reference feature sequence to increase their similarity,
 - Best match corresponds to recognized word.
- Hidden Markov Models (HMMs)
 - Reference model typically consists of few states representing e.g. phonemes of speech,
 - The transition probabilities between states (phonemes) are obtained using training data.
- Deep Neural Networks (DNNs), Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs)

- The local match is known (e.g. Euclidean distance between test and reference)
- Goal: find optimal path to maximize the global similarity



- A sequence of states are connected by transition probabilities
- $a_{ij} > 0 \quad \forall i, j.$
- For an ergodic HMM, every state can be reached from every state
- Markov property: probability of a future state in a sequence depends only on current state.



Example: Weather

States:

s_1 : cloudy

s_2 : rainy

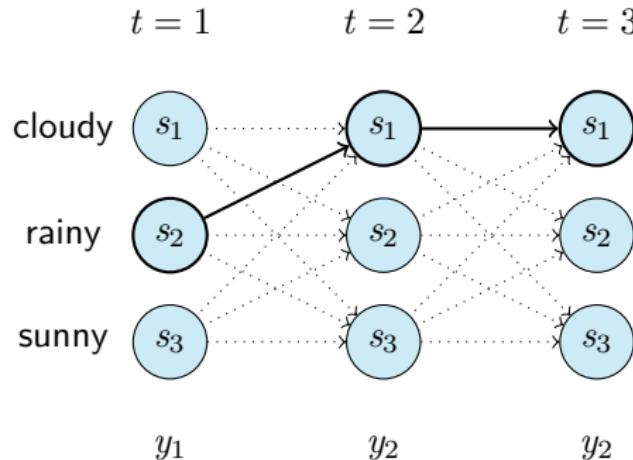
s_3 : sunny

Observations:

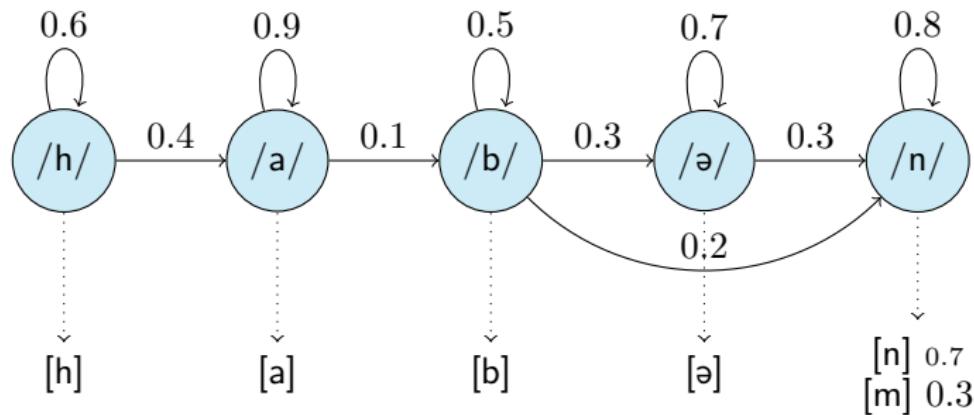
y_1 : man with umbrella

y_2 : man without umbrella

- Indicate the most likely path for a given observation
 - The thick arrows indicate the most probable transitions.
- state s_i at time t.



- In speech recognition, the states commonly represent phonemes modeled by GMMs
- Usually, in speech no backwards transitions are modelled (left-to-right HMM)
- Each HMM represents the item to be recognized (e.g. 10 HMMs for digit recognition)



- [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.
- [2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
- [3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
- [4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.
- [5] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.
- [6] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
- [7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
- [8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [9] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.
- [10] T. Gerkmann, "Statistical analysis of cepstral coefficients and applications in speech