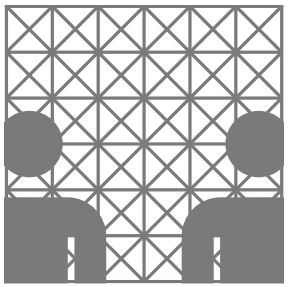
Specialization Module

Speech Technology

Timo Baumann
baumann@informatik.uni-hamburg.de

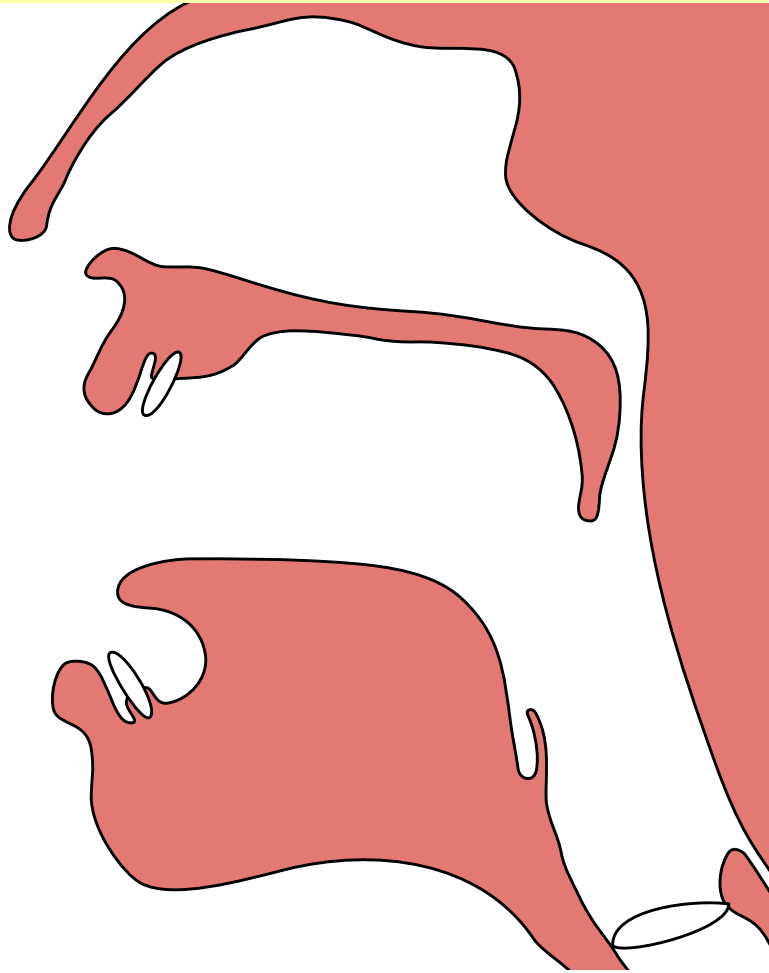
UNIVERSITÄT HAMBURG, DEPARTMENT OF INFORMATICS
NATURAL LANGUAGE SYSTEMS GROUP



A bit of Phonetics

Speech Production: Source-Filter Model

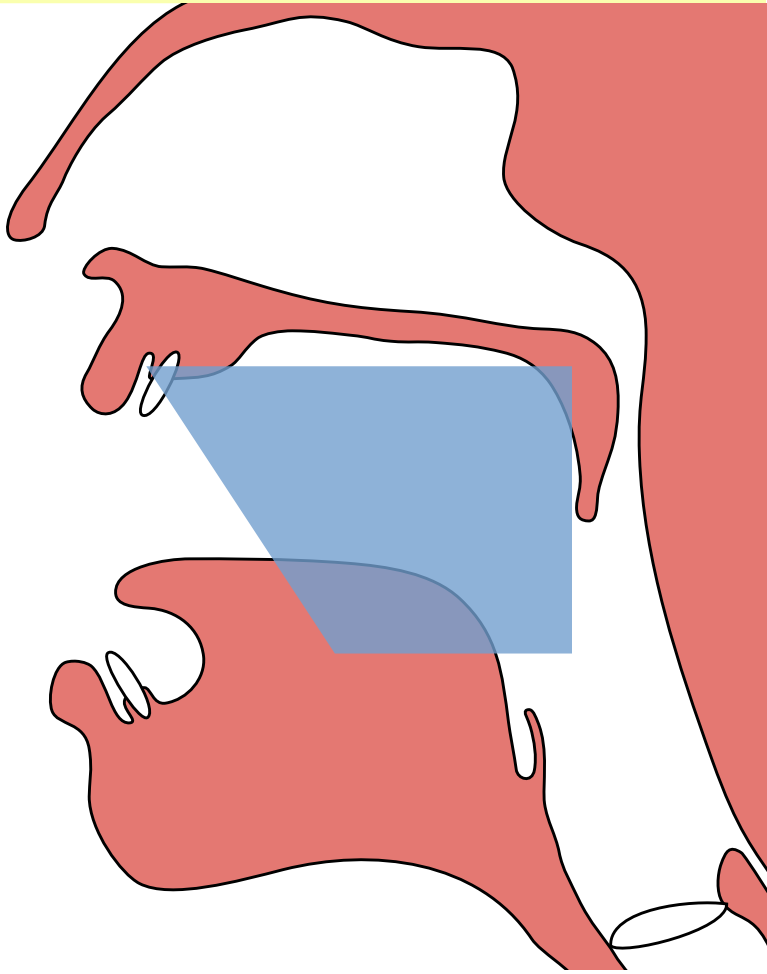
- glottal folds produce primary signal
- vocal tract acts as a filter



(slightly different for voiceless sounds)

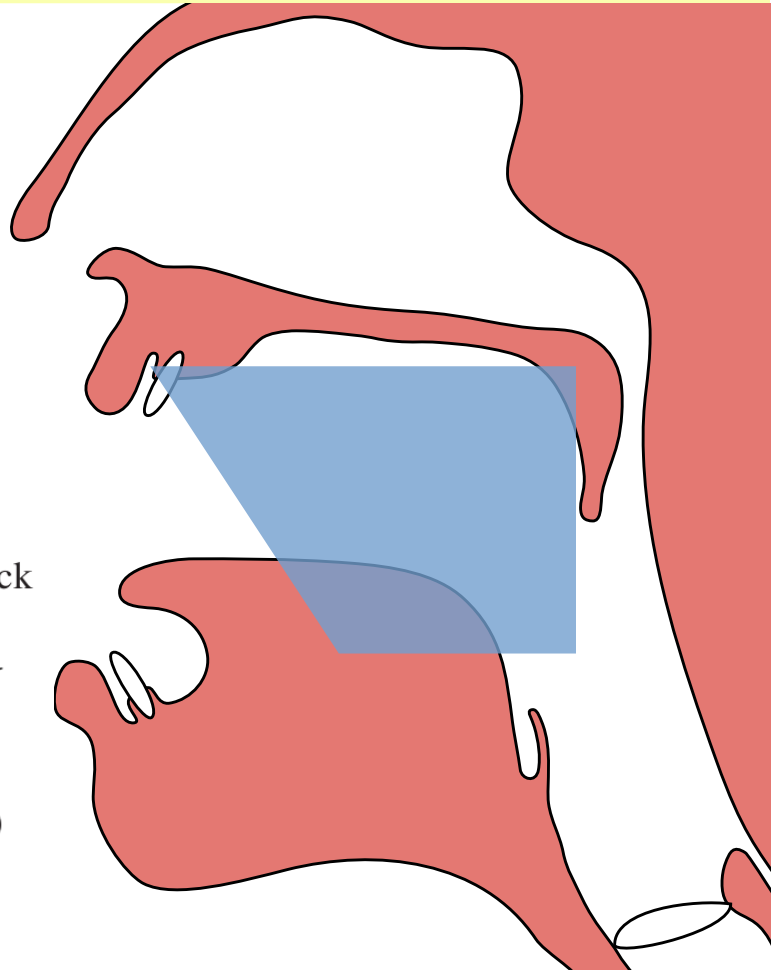
Speech Production: Vowels

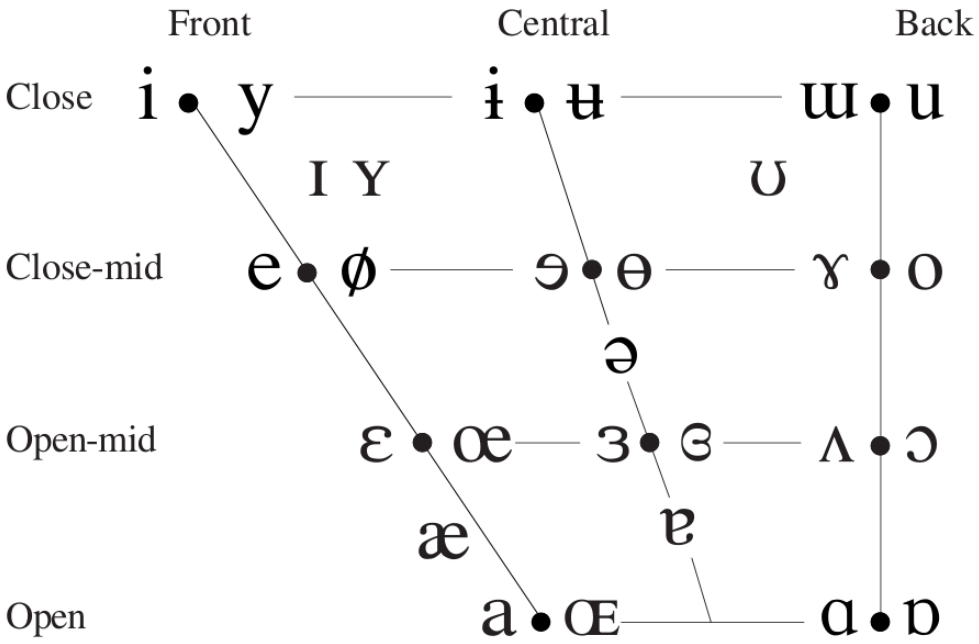
- glottal folds produce primary signal
- vocal tract acts as a filter
  - the field of movement for the tongue in oral cavity is idealized as a trapezoid
  - resonance of cavity determines vowel



Speech Production: Vowels

- glottal folds produce primary signal
- vocal tract acts as a filter
  - the field of movement for the tongue in oral cavity is idealized as a trapezoid
  - resonance of cavity determines vowel

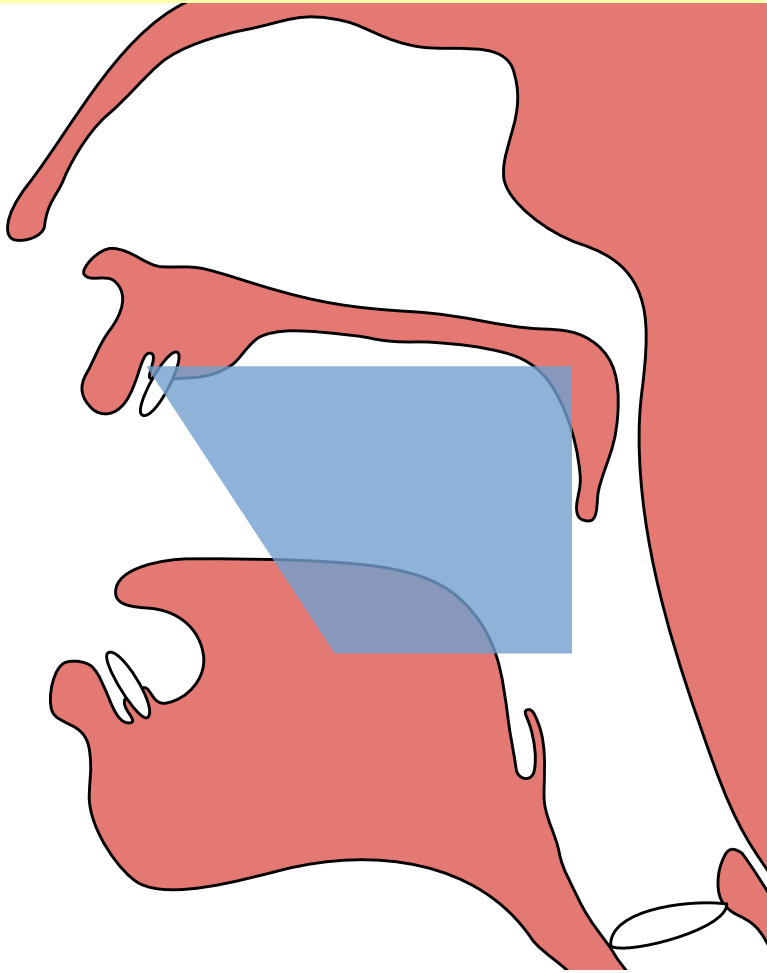


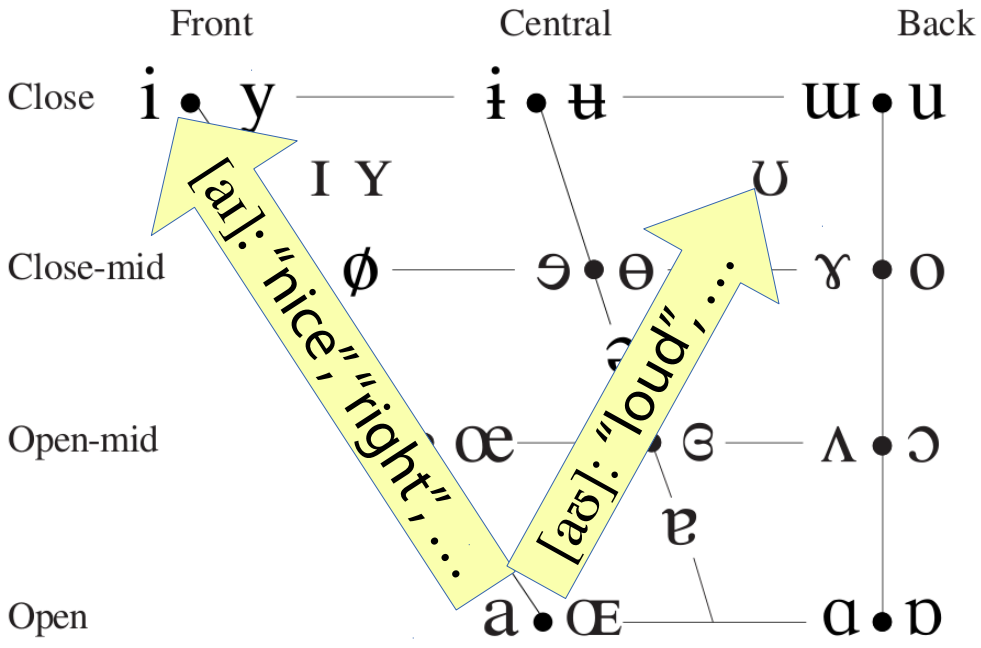


Where symbols appear in pairs, the one to the right represents a rounded vowel.

Vocalic sounds: Diphthongs

- of course, the tongue may move during the vowel, resulting in a changing sound

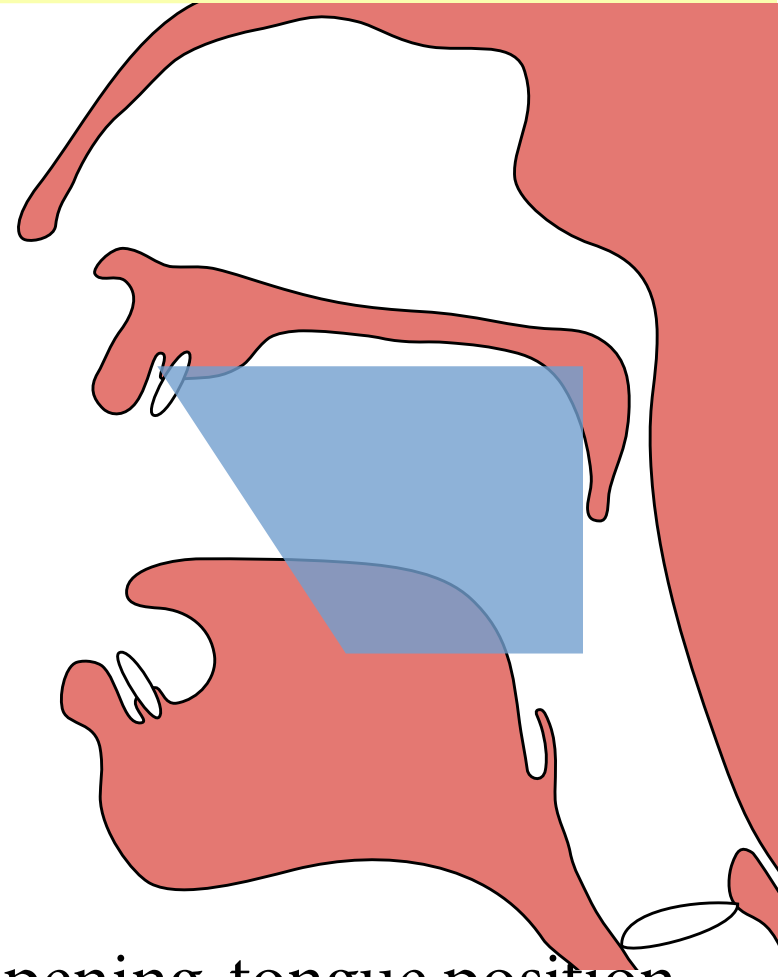




Where symbols appear in pairs, the one to the right represents a rounded vowel.

Speech Production: Consonants

- two types of phones:
  - ✔ vowels: air is exhaled „freely“
  - – consonants: obstruction perturbs air
    - although there's no clear definition of what is „still“ an [i:] or „already“ a [j]
    - vocal tract is not just a filter but also a source of additional sound
    - voiceless consonants: glottal folds are open, sound only from perturbation
- further classification criteria:
  - – means of articulation: voicing, mouth opening, tongue position, lip rounding, nasality, secondary obstructions, length, ...
- classification by International Phonetic Association



Consonants

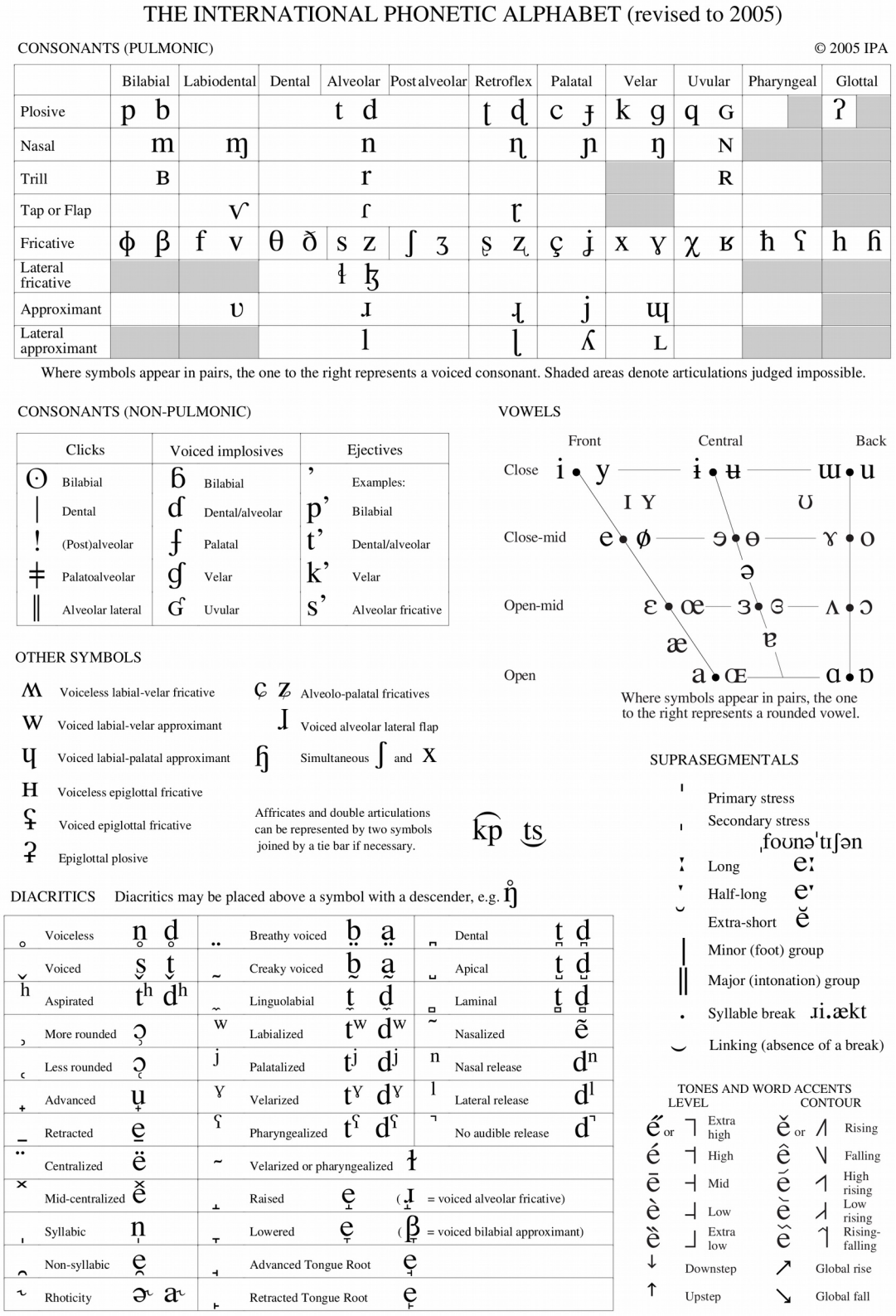
- *manner* of articulation (plosives, nasals, fricatives, …)
- place of constriction (lips, teeth, … glottis)

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

The International Phonetic Alphabet

- more symbols:
  - other sounds (clicks, …)
  - tones
  - stress marks
  - lengthening
  - more details
- used for *narrow transcription*, e.g. in dialectology
- languages often do not distinguish between all possible sounds



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
|	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
!	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
‖	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʧ ʤ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən	
ˌ	Secondary stress		
ː	Long	eː	
ˑ	Half-long	eˑ	
̆	Extra-short	ĕ	
|	Minor (foot) group		
‖	Major (intonation) group		
.	Syllable break	ɹi.ækt	
‿	Linking (absence of a break)		

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ə˞ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR	
e̋ or e᷅	˥	Extra high	ě or e᷄	˩˥ Rising
é	˦	High	ê	˥˩ Falling
ē	˧	Mid	e᷌	˩˥˦ High rising
è	˨	Low	e᷈	˨˩˦ Low rising
ȅ or e᷄	˩	Extra low	e᷉	˥˩˥ Rising-falling
↓		Downstep	↗	Global rise
↑		Upstep	↘	Global fall

Exercise (in small groups):

1. transcribe your name in the phonetic alphabet
2. transcribe some words (ideally: not English nor German) without speaking them aloud
3. exchange notes, listen carefully whether your partners correctly read out your transcript; check for errors

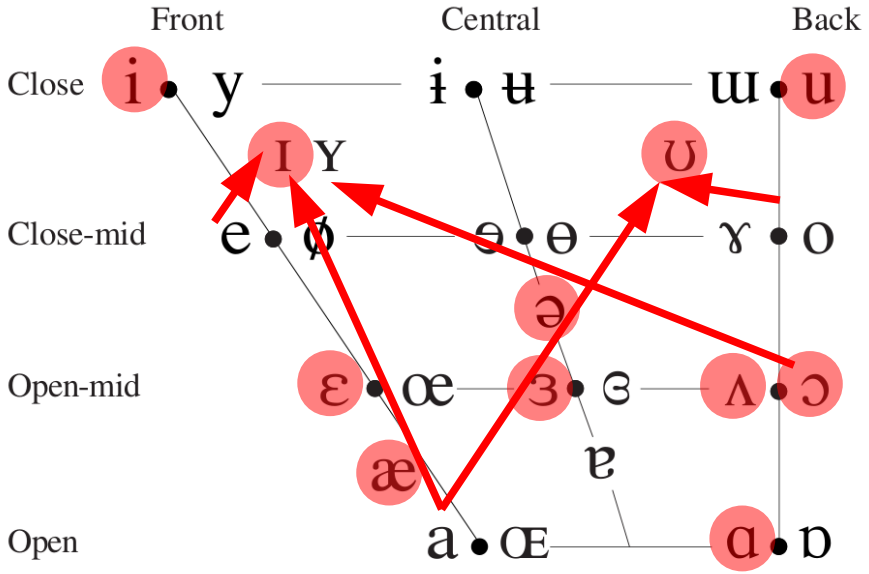
The Phonemic System of a Language

- only small subset of symbols in the IPA
- contextual rules determine phonetic realization
  - e.g. German [ç/x] (“ich”/“ach”) is a single phoneme /ç/
- context limitations (***Phonotactics***), often in combination with **syllabic structure**
  - syllable = onset + nucleus + coda
  - e.g. German nucleus must be a vowel; complex coda with up to 5 consonants (rules for consonant sequences)
  - e.g. Japanese: restrictions on coda and consonant clusters: „Arbeit“ → „arubaito“ – „baumukūhen“, „ryukkusakku“?
  - e.g. English: no /ŋ/ in onset, no /h/ in coda, …

N-American English Phoneme Set

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

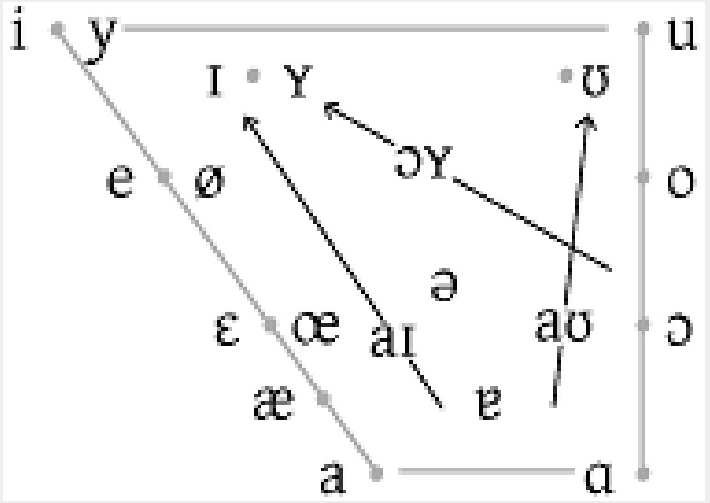
Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.



Where symbols appear in pairs, the one to the right represents a rounded vowel.

German Phoneme Set

Vokale



Konsonanten

		m				n								ŋ			
p	b				t	d						k	g				
		f	v	s	z	ʃ	ʒ	ç				x			h		
		p͡f		t͡s		t͡ʃ											
					l				j				ʀ				

- more vowels(/y/, /ʏ/, /œ/), fewer diphthongs
- similar consonants
  (but their realization differs, e.g. aspiration)

Units of Speech: Phones vs. Phonemes

- speech sounds (→ Phonetics)
- distinguishable units
- language independent
- **Signifiant**

- linguistic symbols (→ Phonology)
- distinctive units
- every language has its phoneme system
- **Signifié**

  - minimal pairs: "**b**at" – "**r**at" – "**c**at"
    - /**b**/, /**r**/, /**k**/ are phonemes in English, thus different phones
  - one's articulatory/perceptory capacities are shaped by the mother tongue(s)
    - different sounds may sound identical or be hard to pronounce

Units of Speech: Phones vs. Phonemes

- speech sounds (→ Phonetics)
- distinguishable units
- language independent
- **Signifiant**

Notational Convention:

"examples" in quotes
/phonemes/ in slashes
[phones] in brackets

- linguistic symbols (→ Phonology)
- distinctive units
- every language has its phoneme system
- **Signifié**

- minimal pairs: "**b**at" – "**r**at" – "**c**at"
  - /**b**/, /**r**/, /**k**/ are phonemes in English, thus different phones
- one's articulatory/perceptory capacities are shaped by the mother tongue(s)
  - different sounds may sound identical or be hard to pronounce

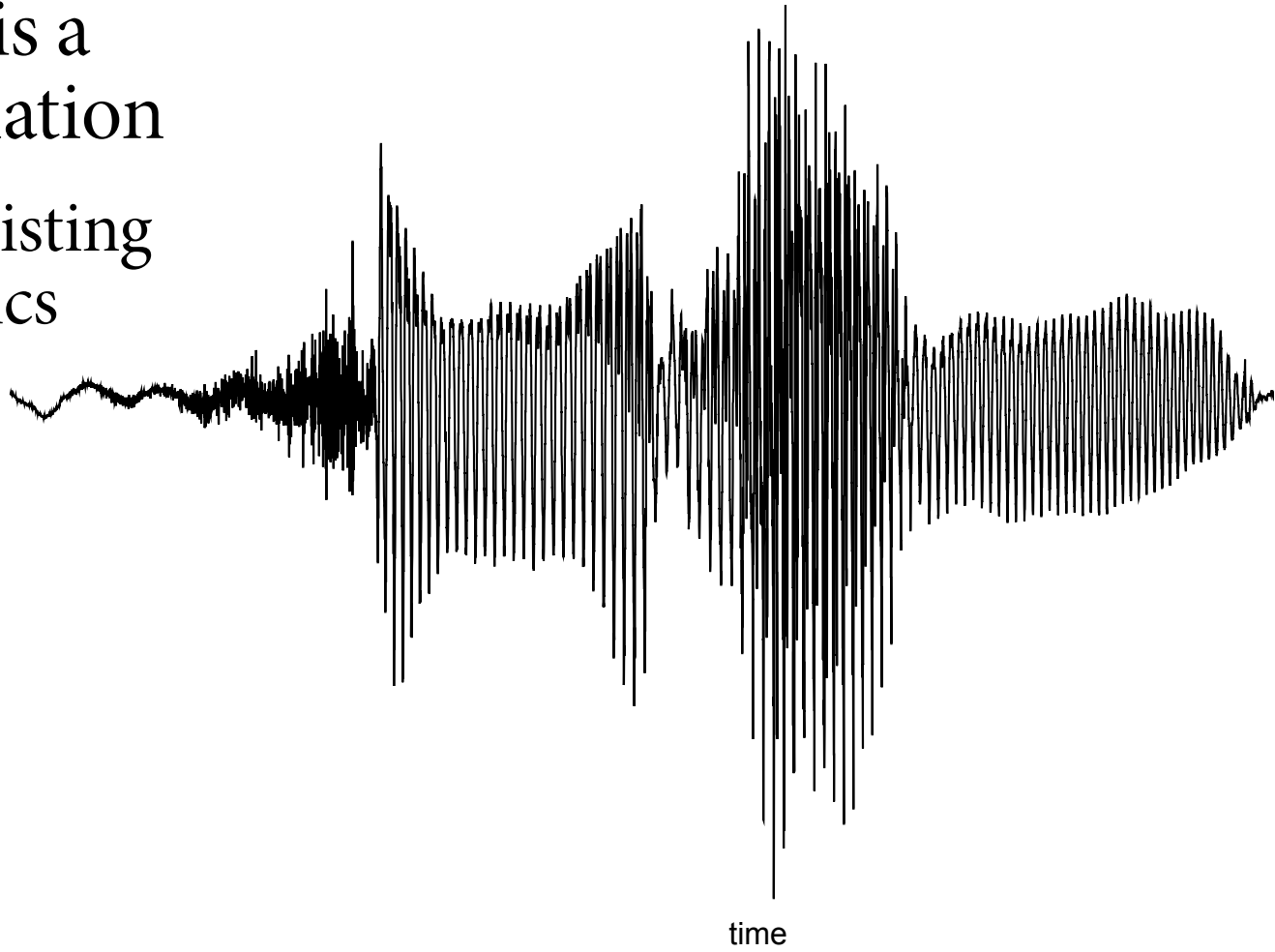
Phonotactics

- words have a phonemic representation in the mental lexicon:
  - "probably" → /ˈprabəbli/
- phonotactics determines realization
  - /ˈprabəbli/ → [prɑːbəbli]
- often material is left out in faster speech (**elision**)
  - "probably" → → [prɑːwliː]
  - this is also (partly) determined by phonotactics and highly context-dependent (speed, setting, …)

Speech: the continuous signal of a symbolic system (language).

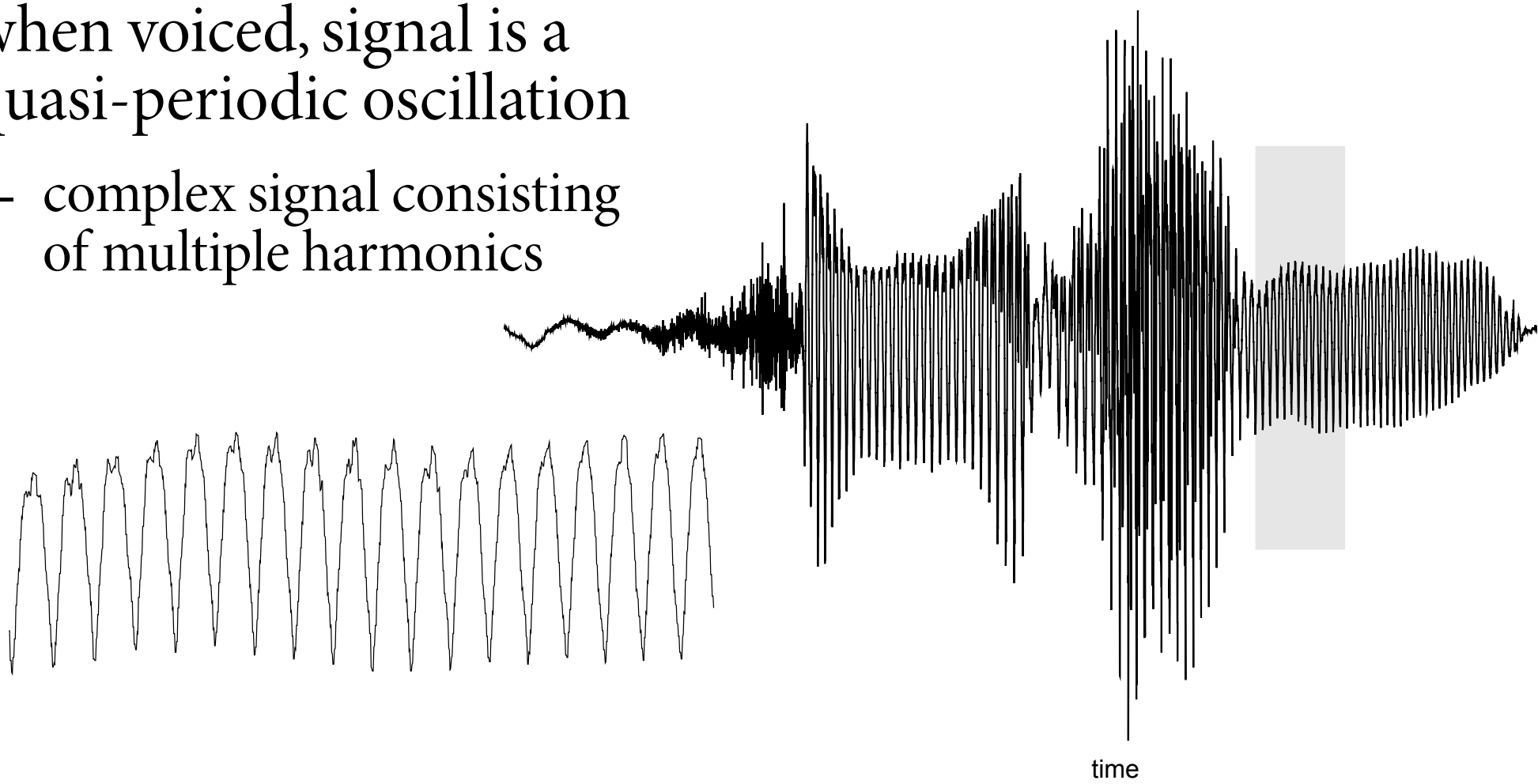
Acoustic (and other 1-dimensional) Signals

- x(t): pressure differential in air over time
- non-stationary: signal changes over time
- when voiced, signal is a quasi-periodic oscillation
  - complex signal consisting of multiple harmonics



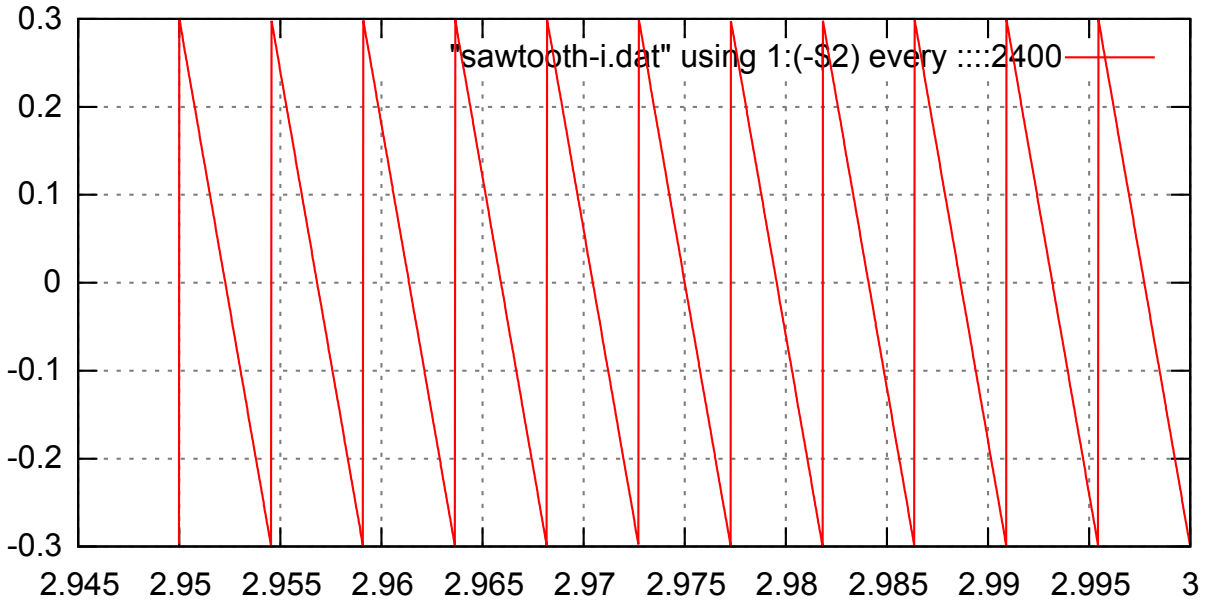
Acoustic (and other 1-dimensional) Signals

- x(t): pressure differential in air over time
- non-stationary: signal changes over time
- when voiced, signal is a quasi-periodic oscillation
  - complex signal consisting of multiple harmonics



Complex Periodic Signals

- simplest signal: sine wave
  - frequency (= 1/wavelength), amplitude, phase
- all periodic signals can be combined from (an infinite number) of sine waves
- e.g. the sawtooth signal:

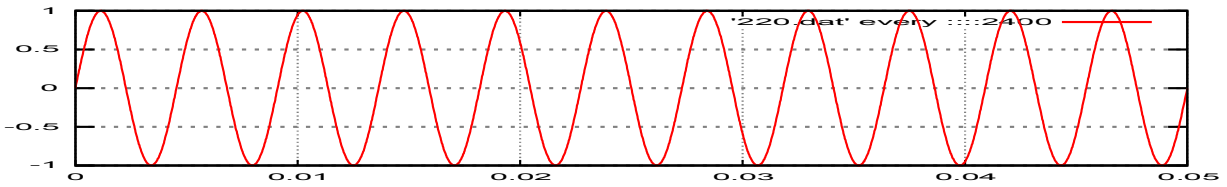


Fourier Synthesis

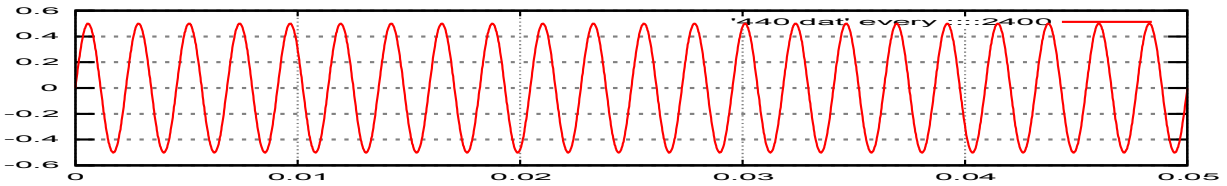
- sawtooth signal: $x(t) = \sum_{k=1}^{\infty} \sin \frac{(2\pi k f t)}{k}$

- approximate with fewer (than infinitely many) sine waves:

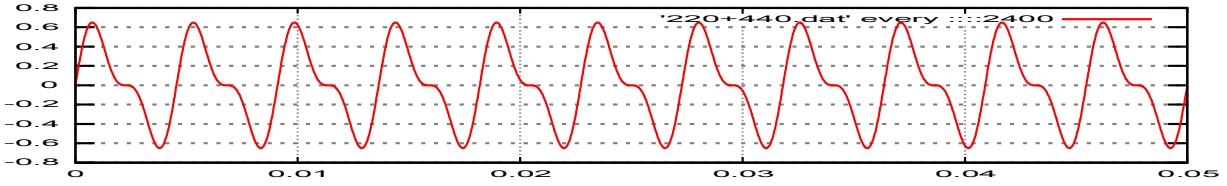
220Hz



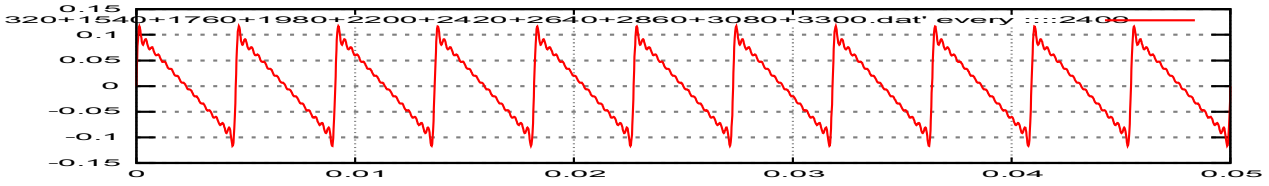
440Hz



220+440Hz

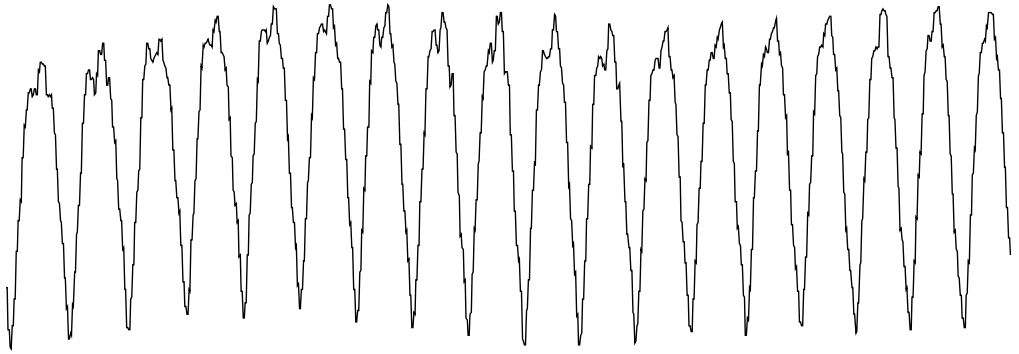


220+440+
...+3300 Hz



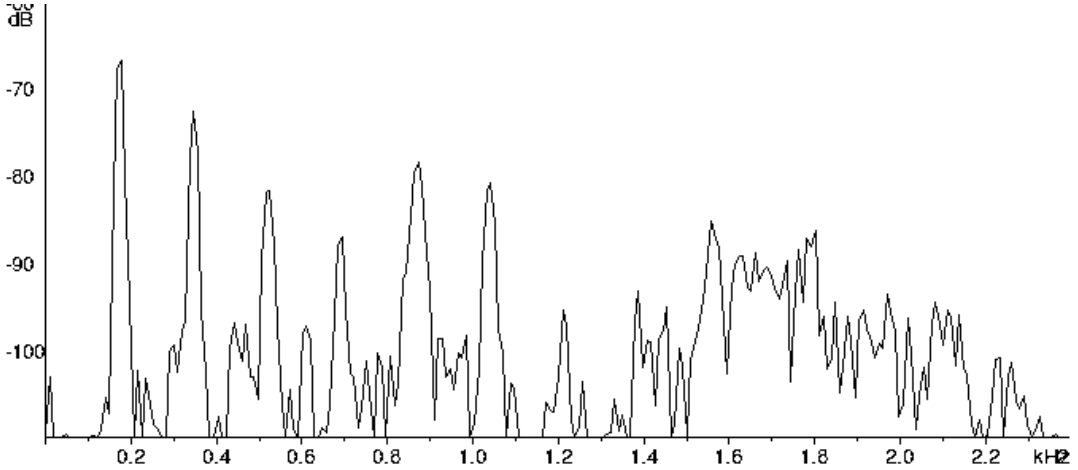
Fourier Analysis

- every complex signal can be analysed into their constituting sine waves (frequency, phase, amplitude)
  → Fourier's theorem



- speech signal
  x-axis: time
  y-axis: amplitude

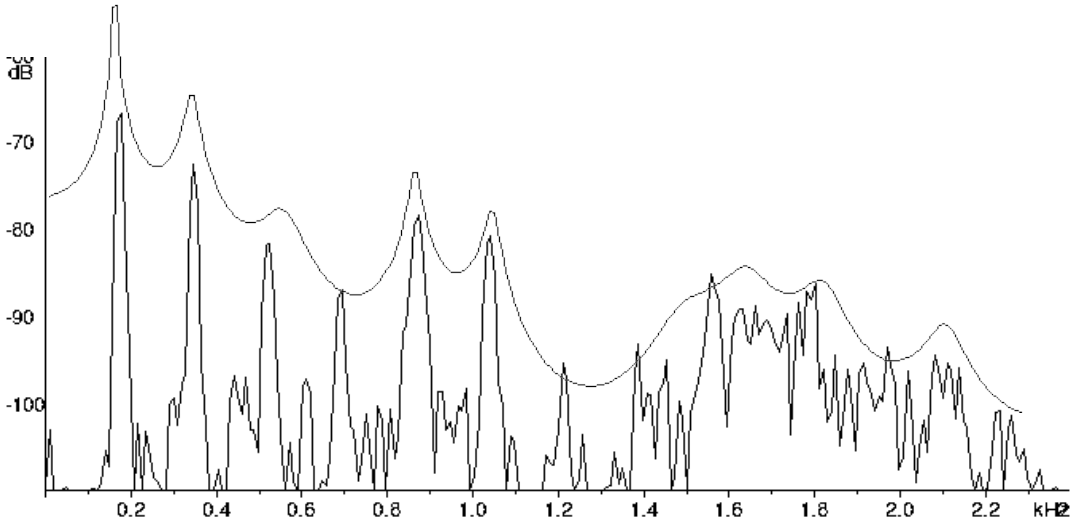
- FFT-spectrum
  x-axis: **frequency**
  y-axis: amplitude
- phase is often ignored



The human ear performs frequency analysis.

Auditory Processing

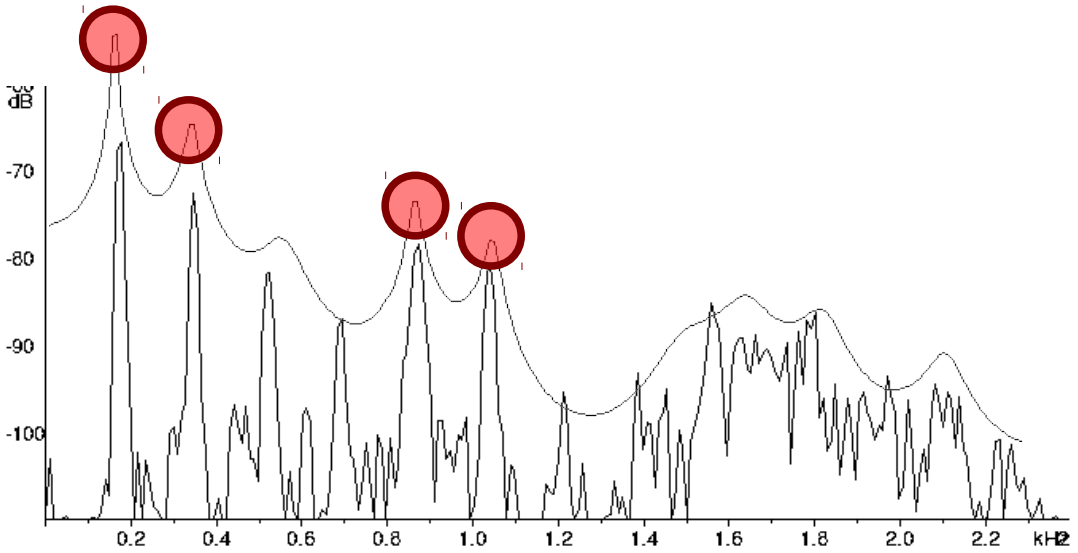
- large spikes from harmonics of **fundamental frequency**
- **signal envelope** is registered by the auditory organ
- speech sounds result in characteristic peaks in the signal envelope
- **formants**



- exception: non-harmonic sounds, such as plosives

Auditory Processing

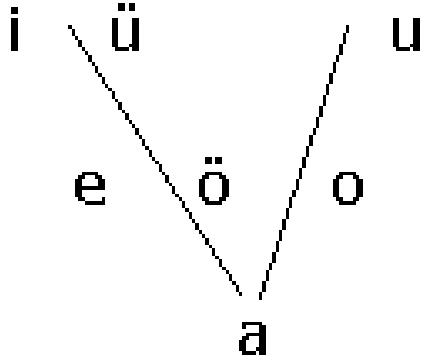
- large spikes from harmonics of **fundamental frequency**
- **signal envelope** is registered by the auditory organ
- speech sounds result in characteristic peaks in the signal envelope
- **formants**



- exception: non-harmonic sounds, such as plosives

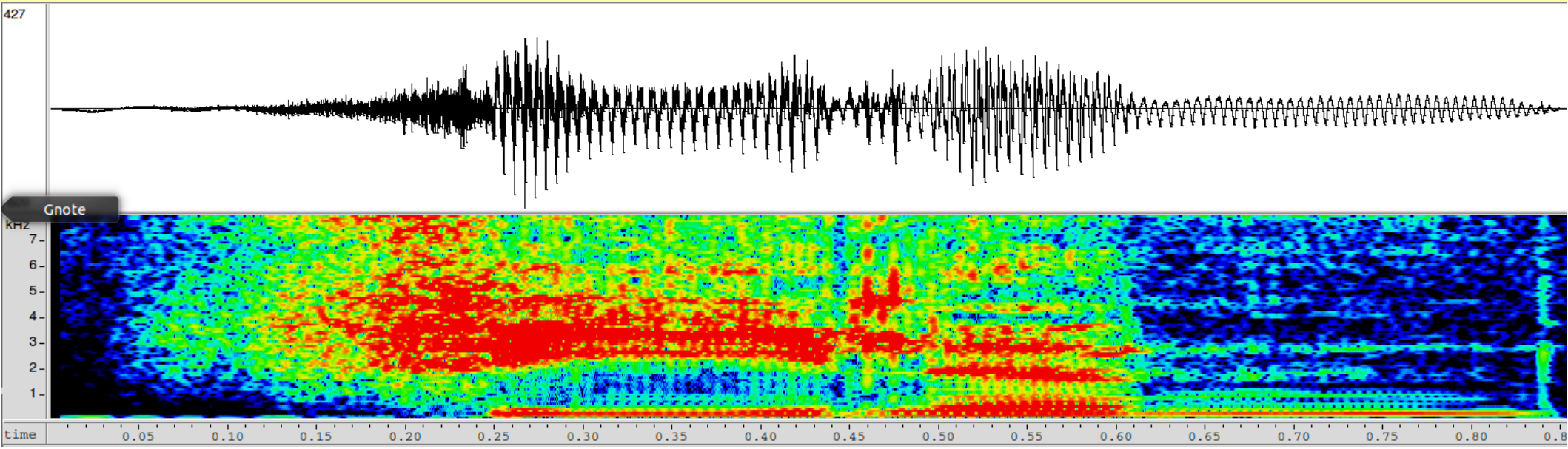
Formants

- the auditory organ performs frequency analysis
- peaks *mask* close-by but smaller peaks
- only largest peaks are tracked and amplified → formants
- *Schwa* sound (mid-central vowel):
  peaks ~ 500Hz, 1500Hz, 2500Hz
  (depends on length of vocal tract)
- vowel triangle: positions of vowels
  relative to 1st and 2nd formant



Speech varies over time.

Spectrogram



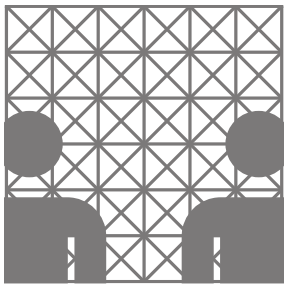
- display changing spectrum over time
  - slice the signal into (overlapping) windows
  - analyze windows individually (using Fourier analysis)
  - use colors to draw spectrum strength

Thank you.

baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SLP16

UNIVERSITÄT HAMBURG, DEPARTMENT OF INFORMATICS
NATURAL LANGUAGE SYSTEMS GROUP



Further Reading

- Speech Signal Representation:
  - P. Taylor (2009): *Text-to-Speech Synthesis*. Cambridge Univ Press. ISBN: 978-0521899277. InfBib: A TAY 43070
  - D. Jurafsky & J. Martin (2009): Speech and Language Processing. Pearson International. InfBib: A JUR 4204x
- Phonetics:
  - M. Pétursson & J. Neppert (1996): Elementarbuch der Phonetik. Buske.
  - J. Neppert (1999): Elemente einer akustischen Phonetik. Buske.
- Phonology/Phonotactics/Phonological Systems:
  - E. Ternes (1999): Einführung in die Phonologie. Wiss. Buchgesellschaft. ISBN: 978-3534138708.

Notizen

Desired Learning Outcomes

- understand the basics of phonetics:
  - voiced/unvoiced sounds, place and manner of articulation, ...
  - formants explain vowel perception
  - phonetics vs. phonology: (ir)relevance of variability
- understand Fourier synthesis
  - all waveforms can be synthesized from sine waves
  - correspondingly, all waveforms can be analyzed into constituting sine waves: frequency, phase, amplitude
  - speech varies over time, hence we use sliding windows