Speech Science

WiSe 2023

Acoustic Phonetics

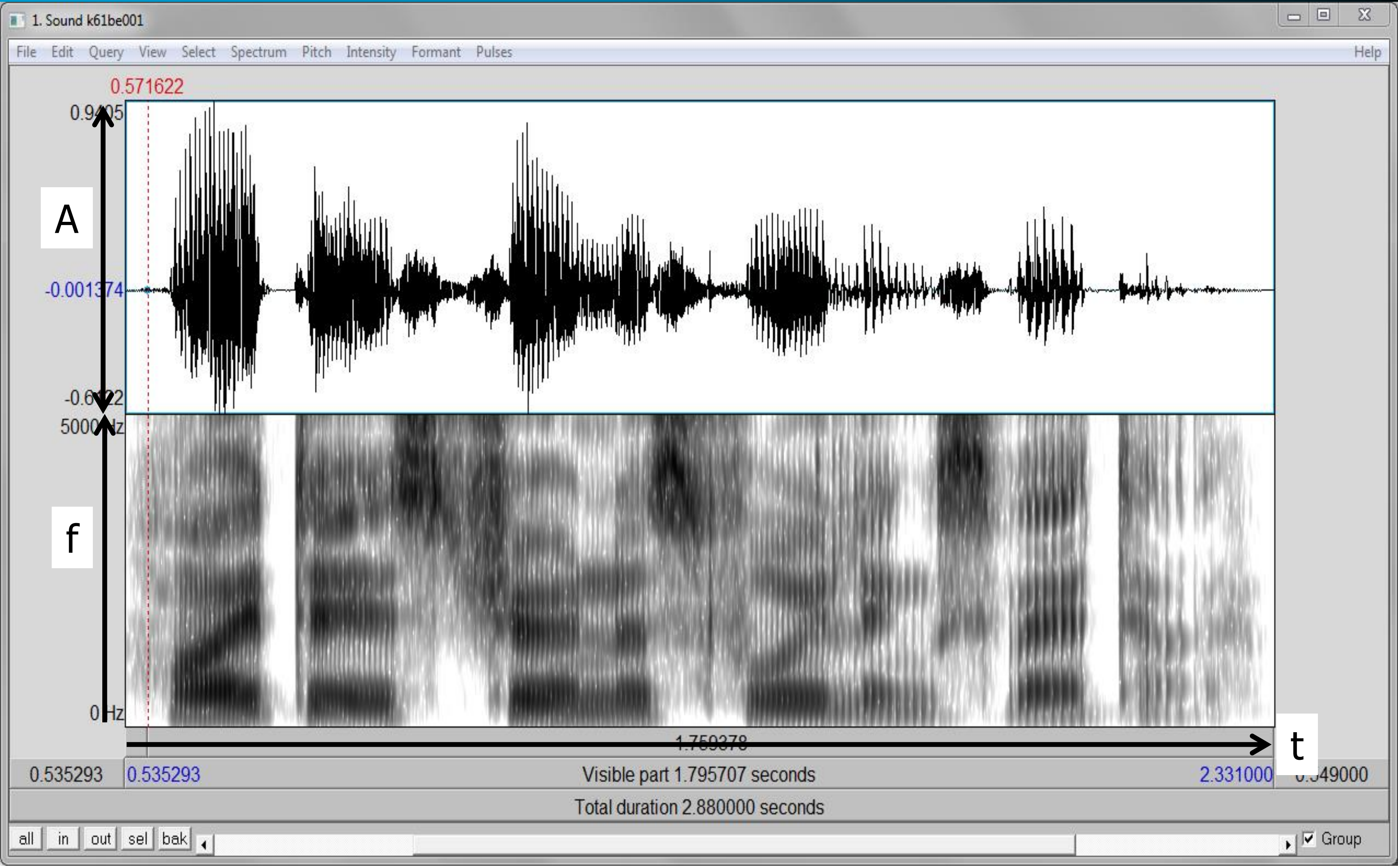
Nov 23, Nov 30, Dec 7, Dec 14, 2023

Bernd Möbius & Omnia Ibrahim

Language Science and Technology
Saarland University

UNIVERSITÄT DES SAARLANDES

Speech waveforms and spectrograms



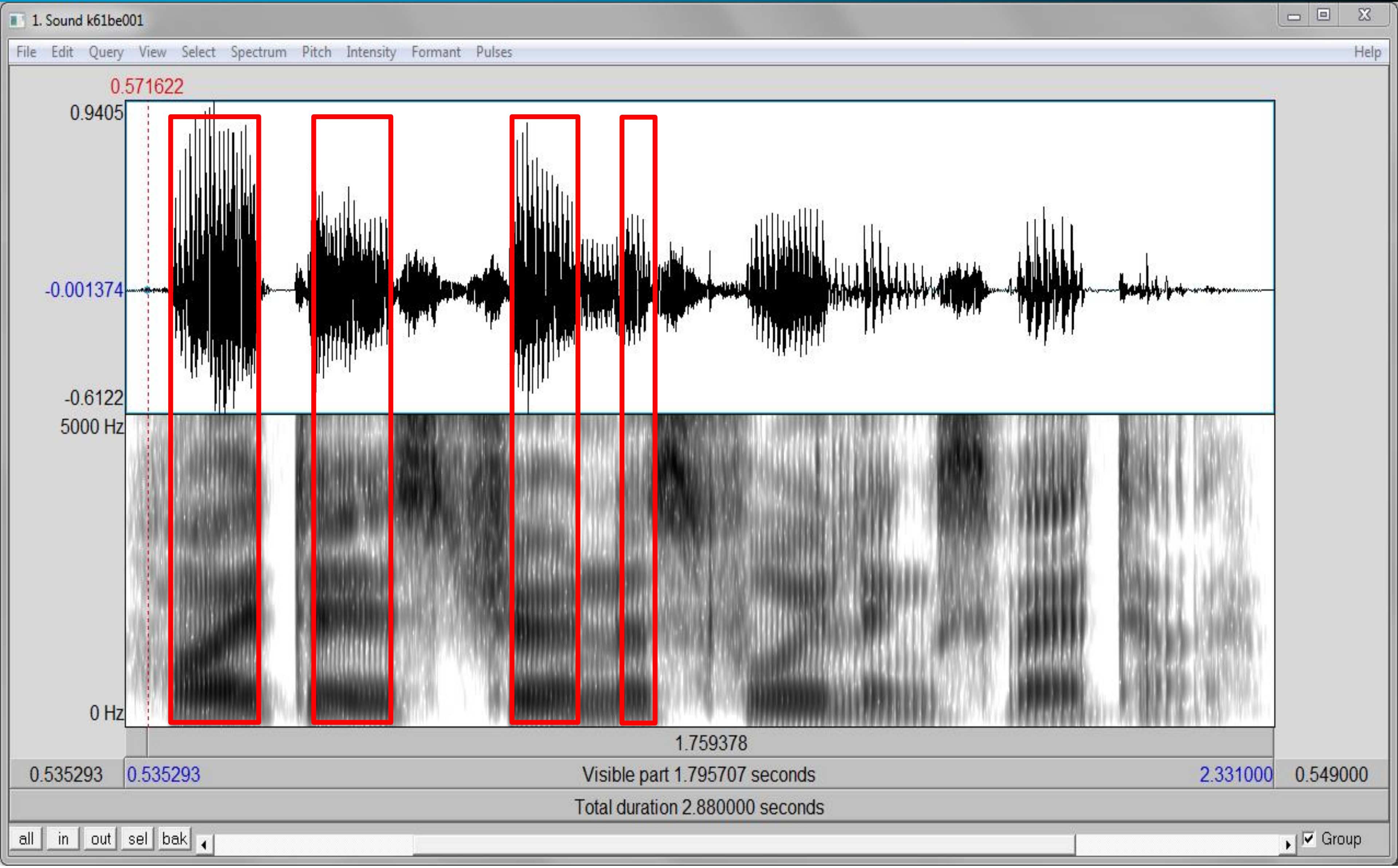
"Heute ist schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

Speech sounds and speech signals

- Basic types of speech signals
  - quasi-periodic signals: sonority
    - vowels
    - sonorants (approximants, glides, nasals, liquids)
  - stochastic signals: frication noise
    - fricatives
    - plosive aspirations
  - transient signals – impulse
    - plosive releases
  - mixed excitation – voiced frication noise
    - voiced fricatives

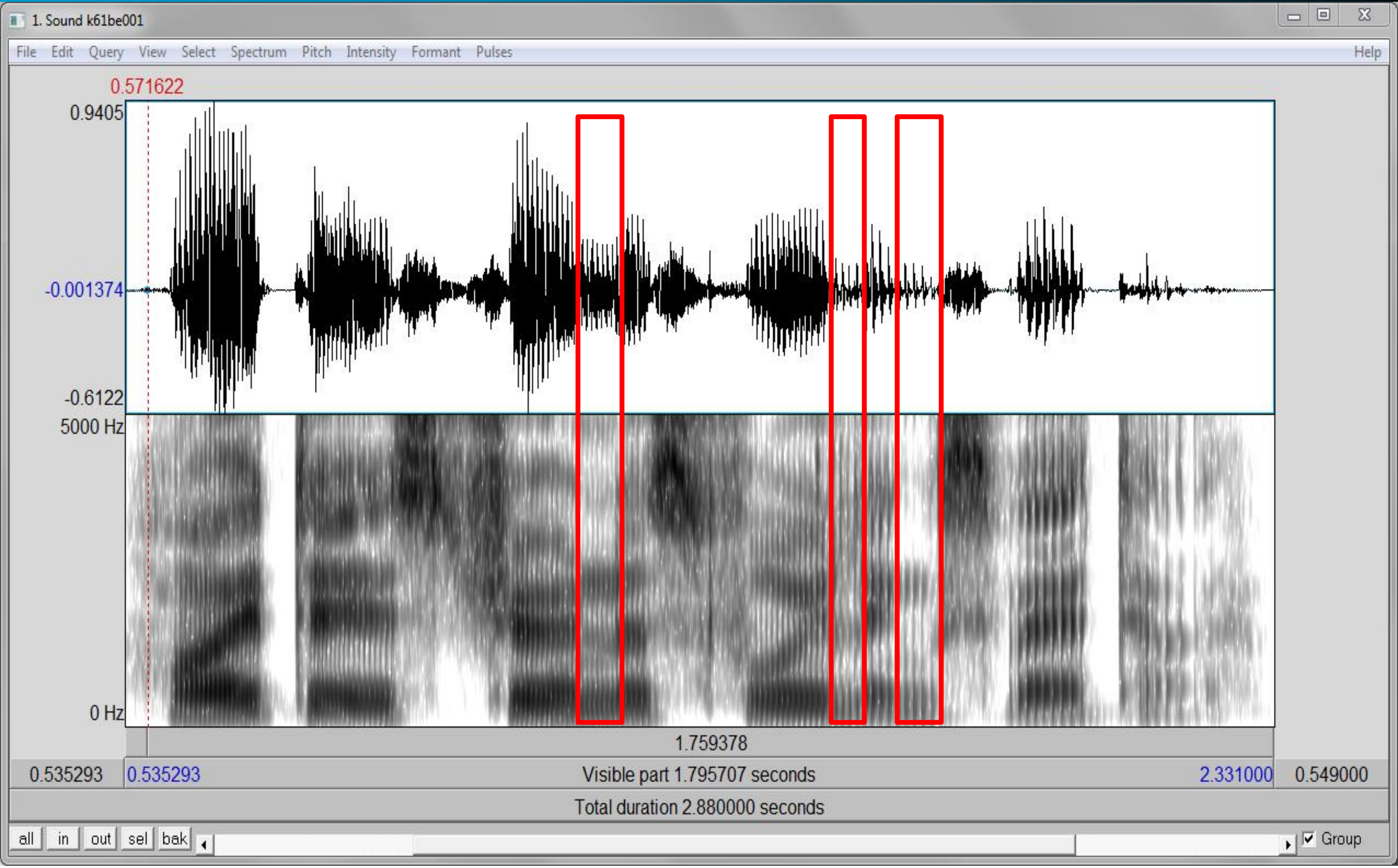
Speech sounds and speech signals: vowels



"Heute ist schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

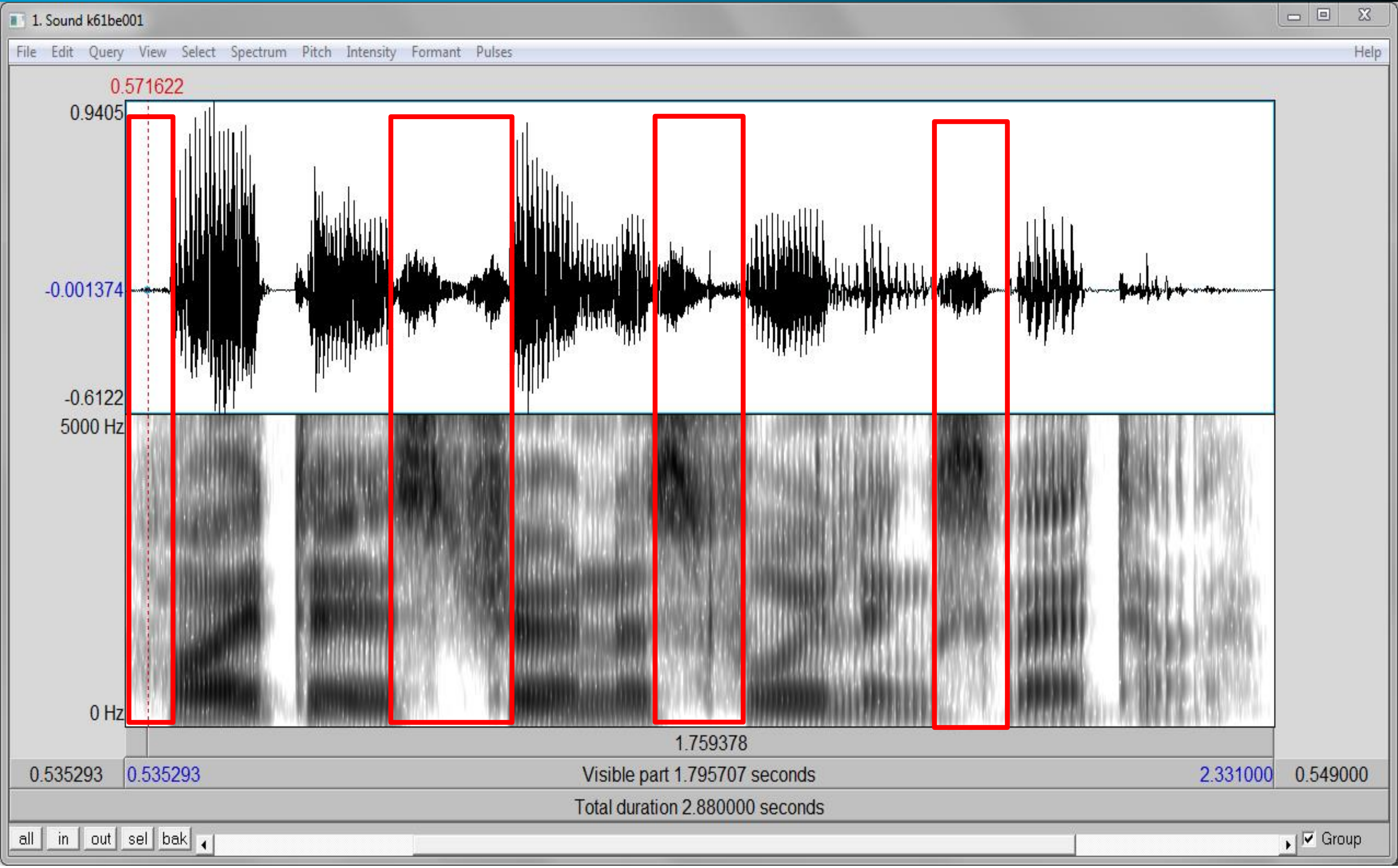
Speech sounds and speech signals: sonorants



"Heute ist schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

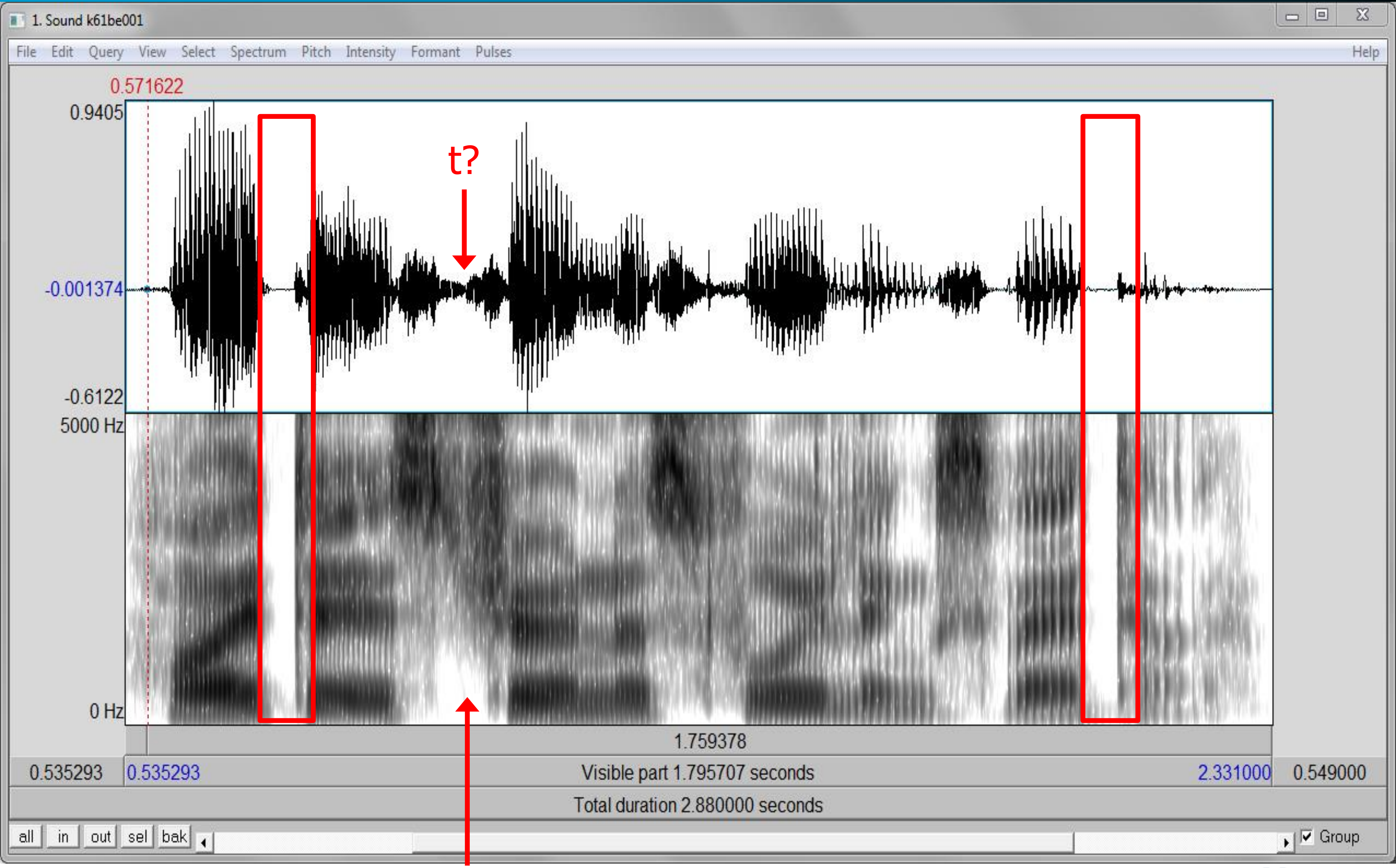
Speech sounds and speech signals: fricatives



"Heute is schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

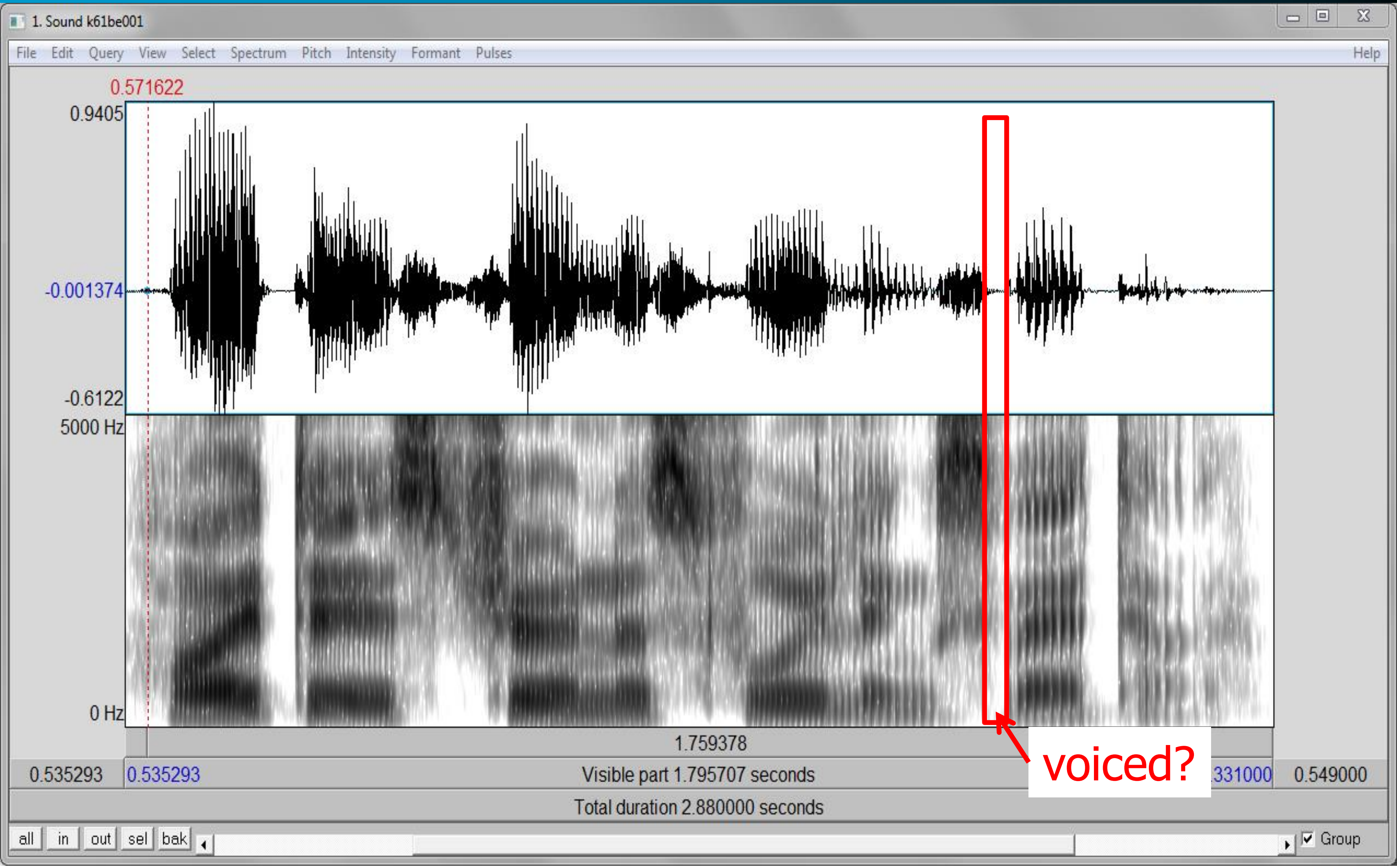
Speech sounds and speech signals: plosives



"Heute is(t) schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

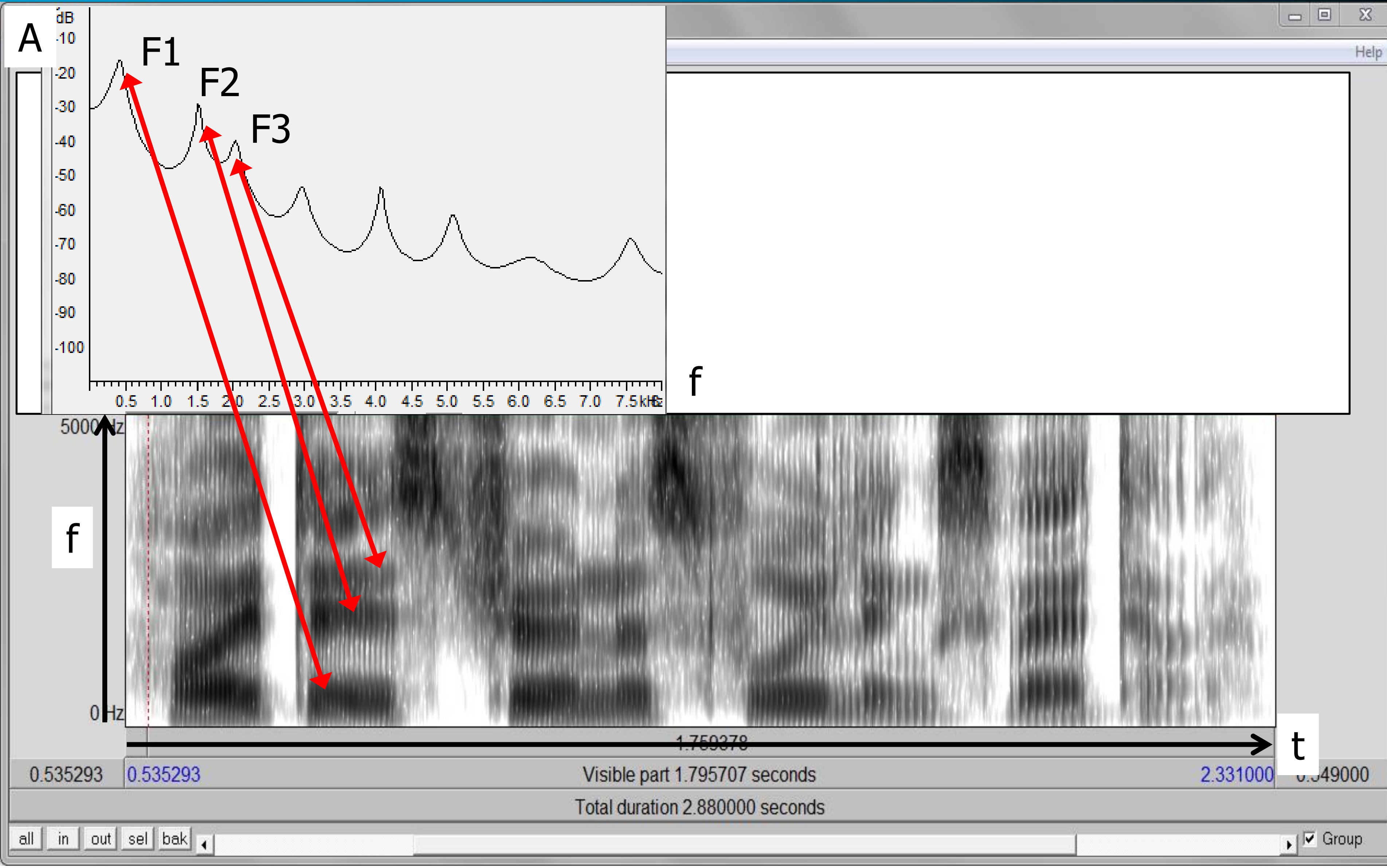
Speech sounds...: voiced fricatives



"Heute ist schönes Frühlingswetter."

UNIVERSITÄT DES SAARLANDES

Speech waveforms and spectrograms

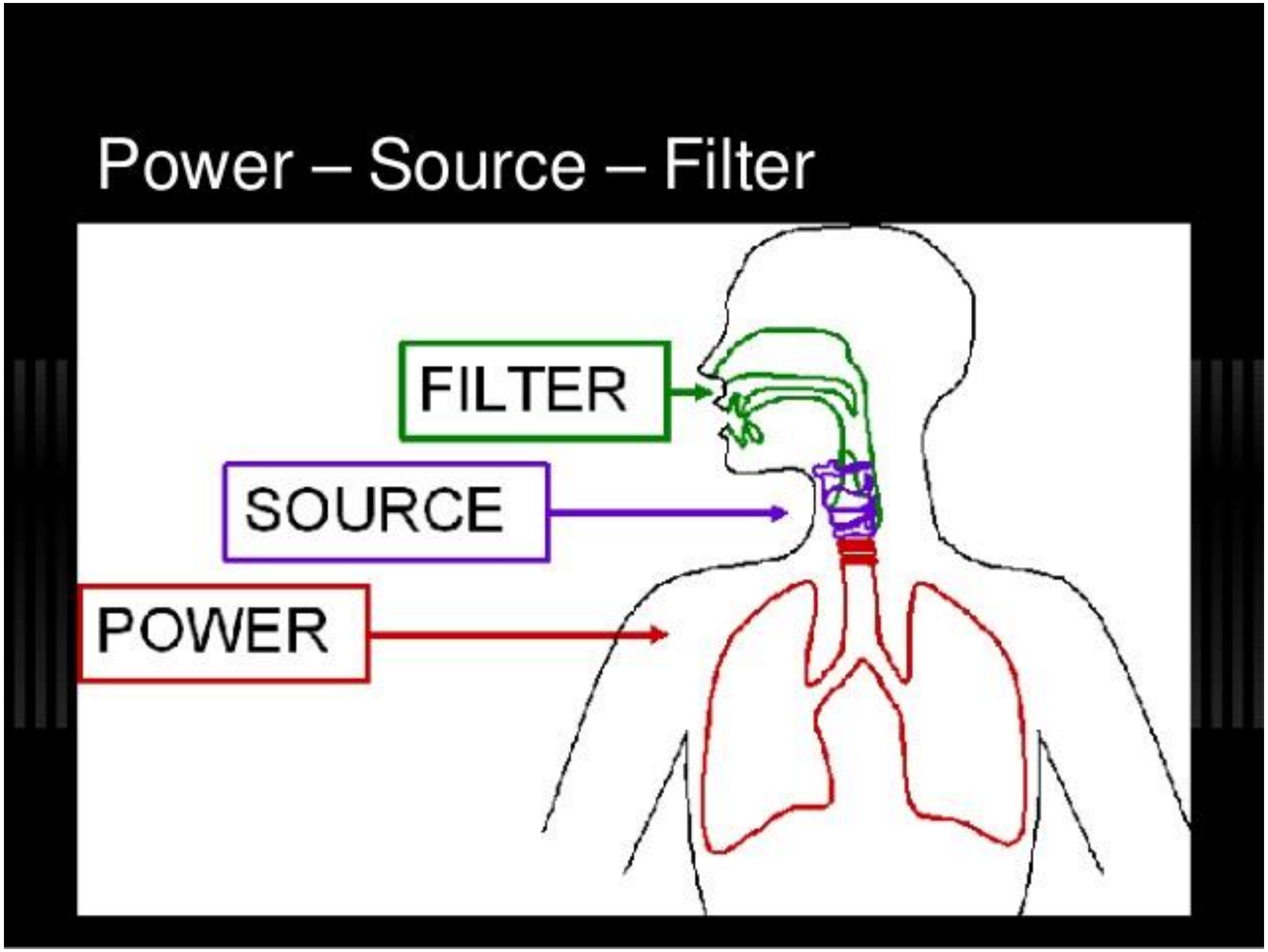


UNIVERSITÄT DES SAARLANDES

Formants

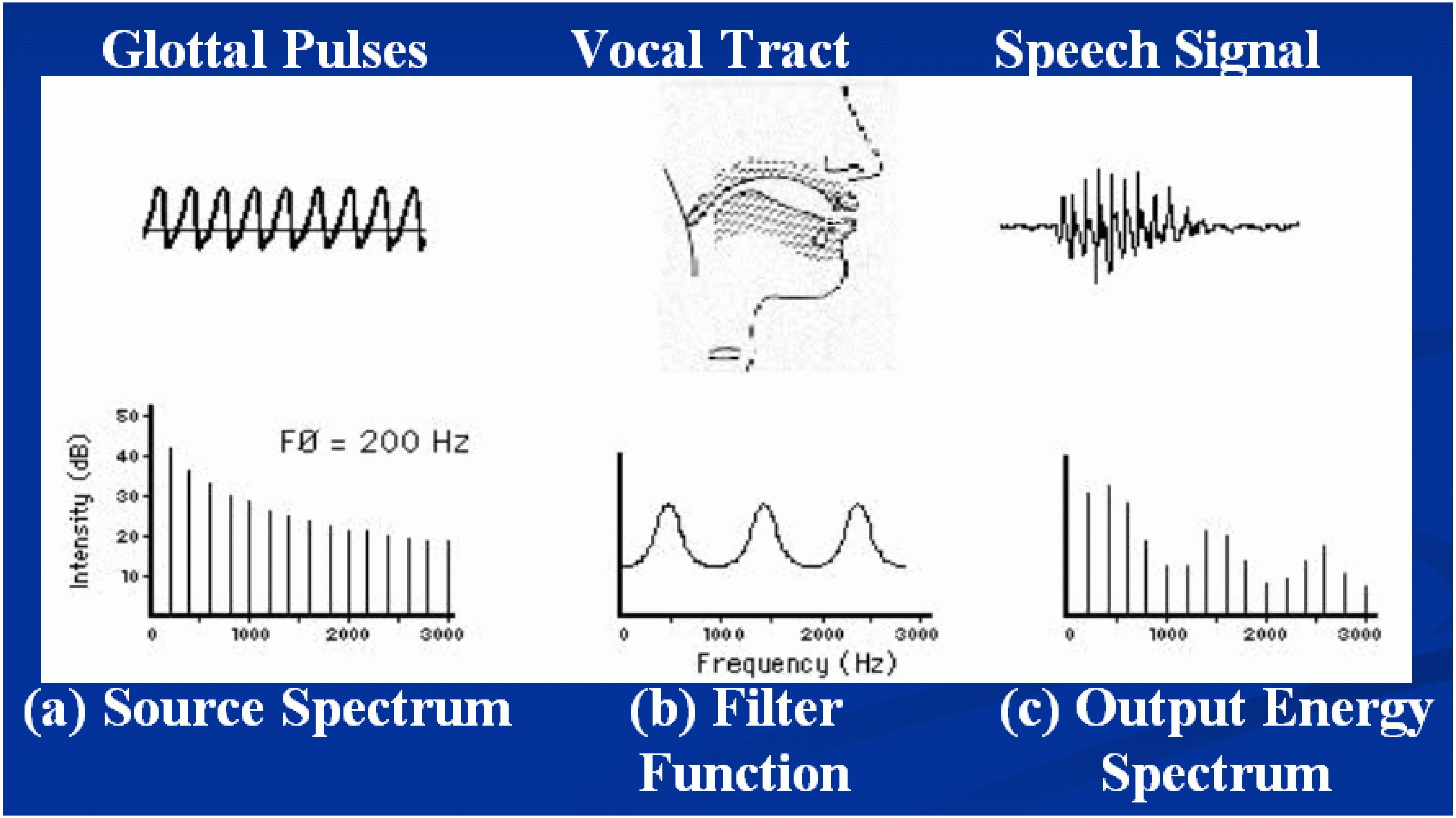
- Spectral peaks (energy maxima) of the sound spectrum: **formants (F1, F2, ..., Fn)**
- Formants emerge as a consequence of selective reinforcement of certain frequency ranges, corresponding to **resonance** characteristics of the vocal tract.
- Distinguishing between **voice source** (*excitation*) and *sound formation* in the vocal tract (**acoustic filter**) motivates the **source-and-filter model** of speech production.
- References:
  - Gunnar Fant (1960): Acoustic theory of speech production
  - Gerold Ungeheuer (1962): Elemente einer akustischen Theorie der Vokalartikulation

Source-filter model of speech production

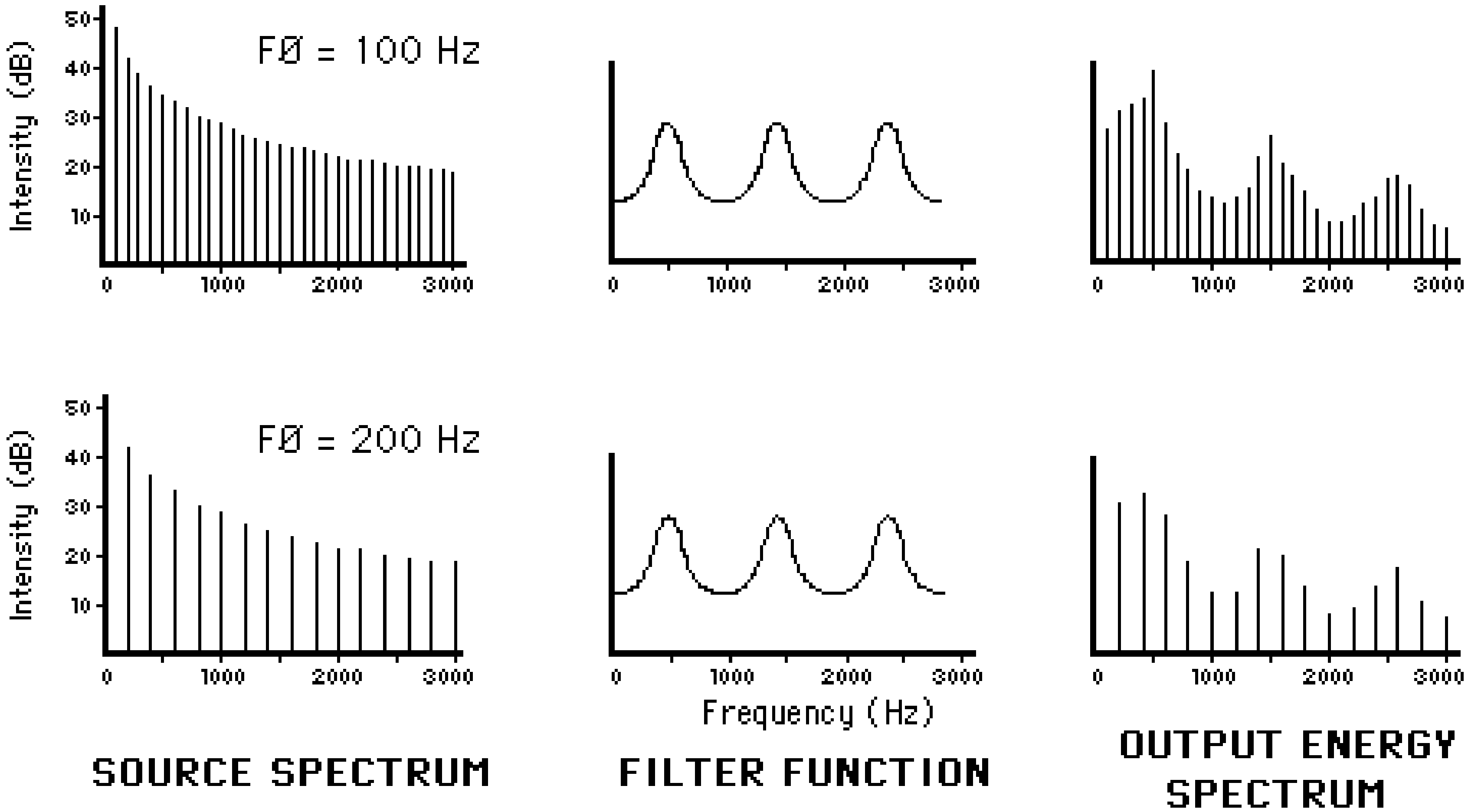


[https://www.vocalsonstage.com]

Source-filter model of speech production



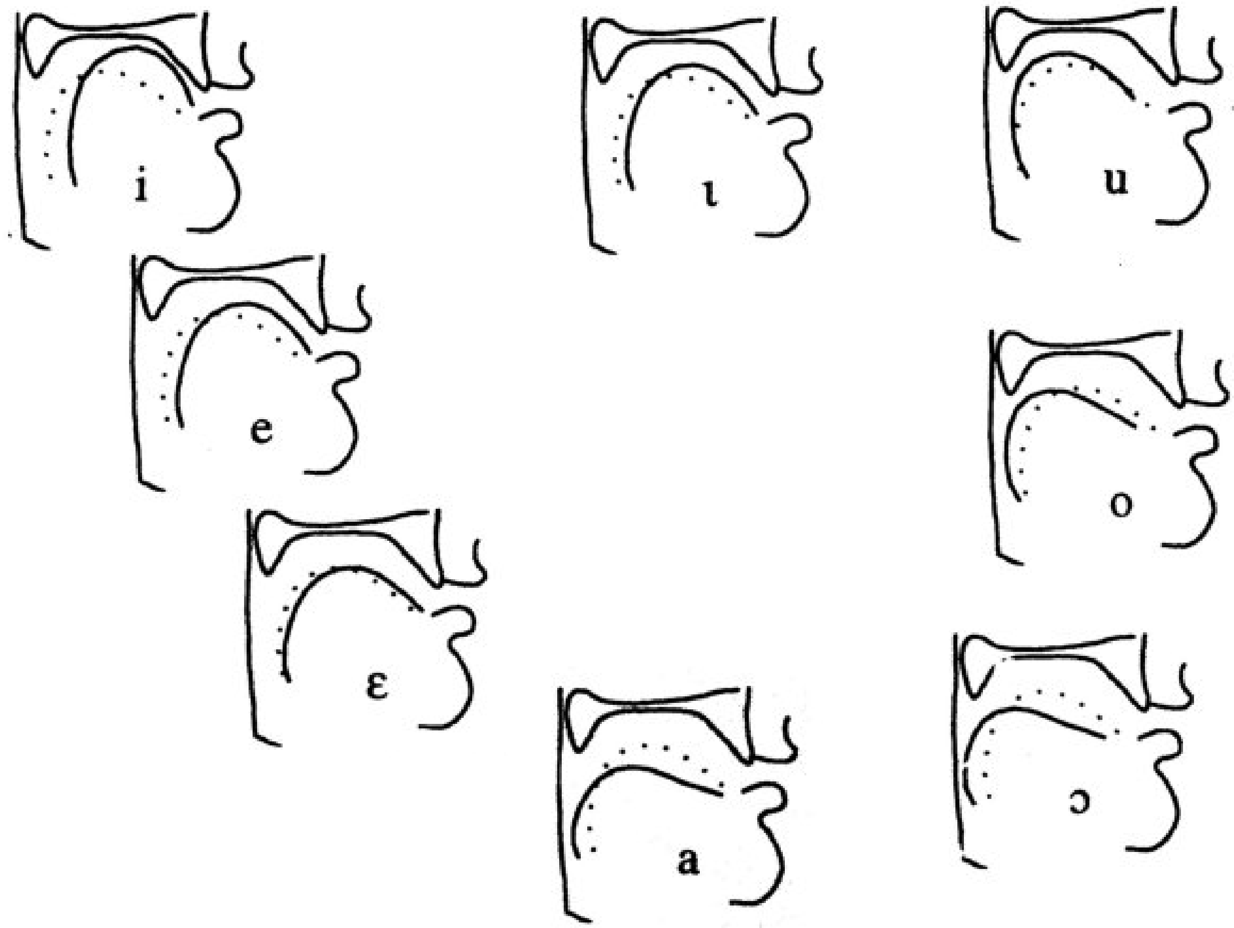
Source-filter model of speech production



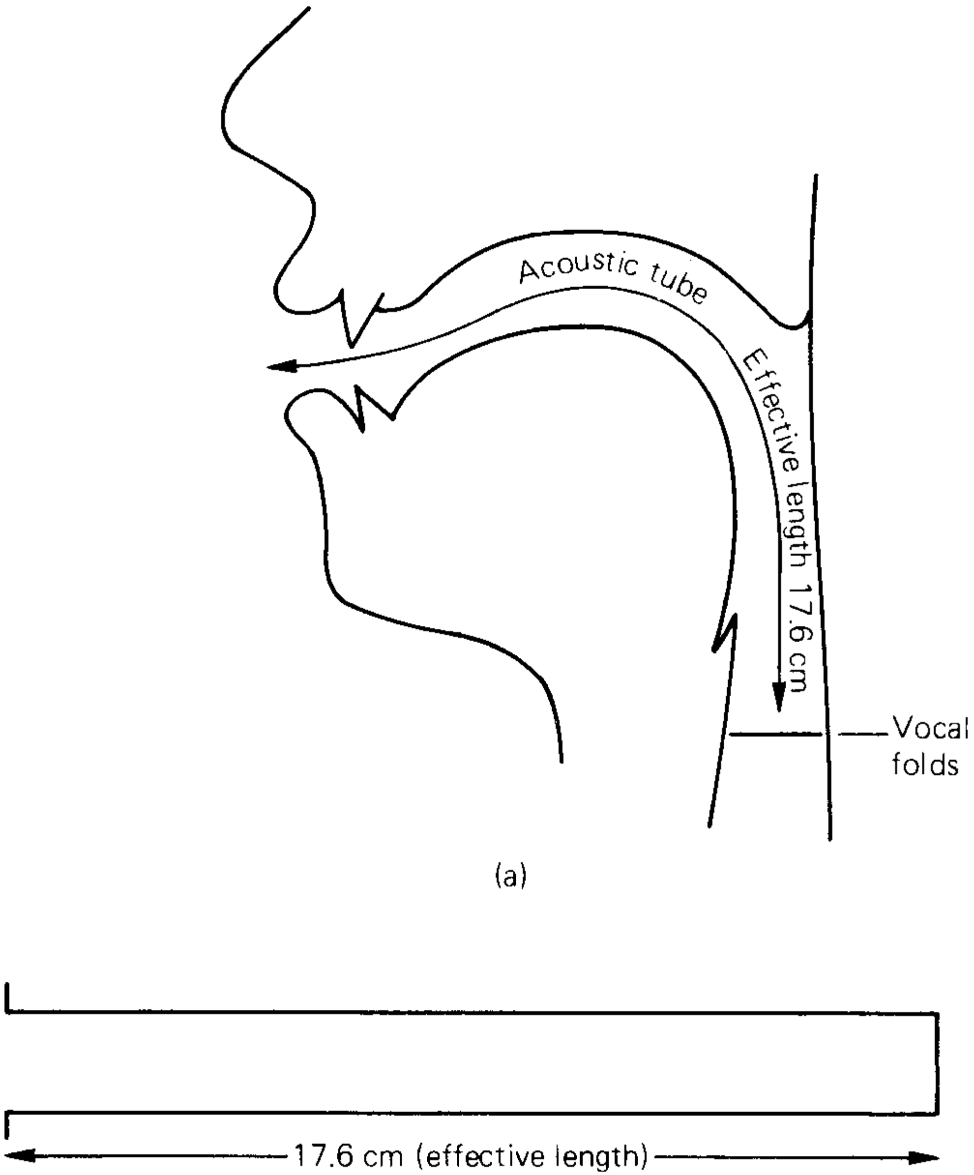
Glottal excitation   Vocal tract frequency response   Sound spectrum

Vocal tract as acoustic filter

- Vocal tract geometry, determined by tongue position (and jaw opening and lip protrusion, not shown)

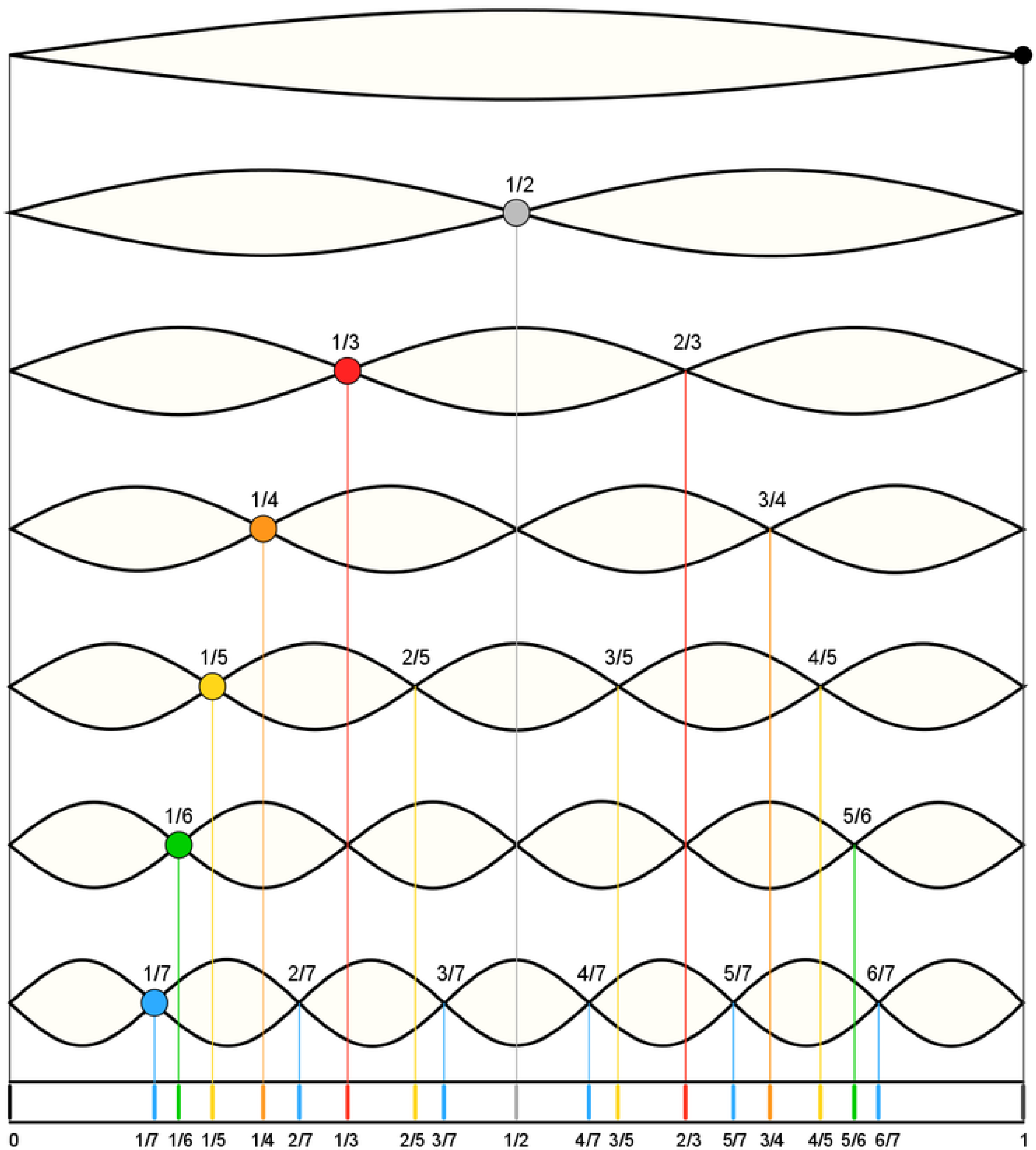


Vocal tract: acoustic tube model

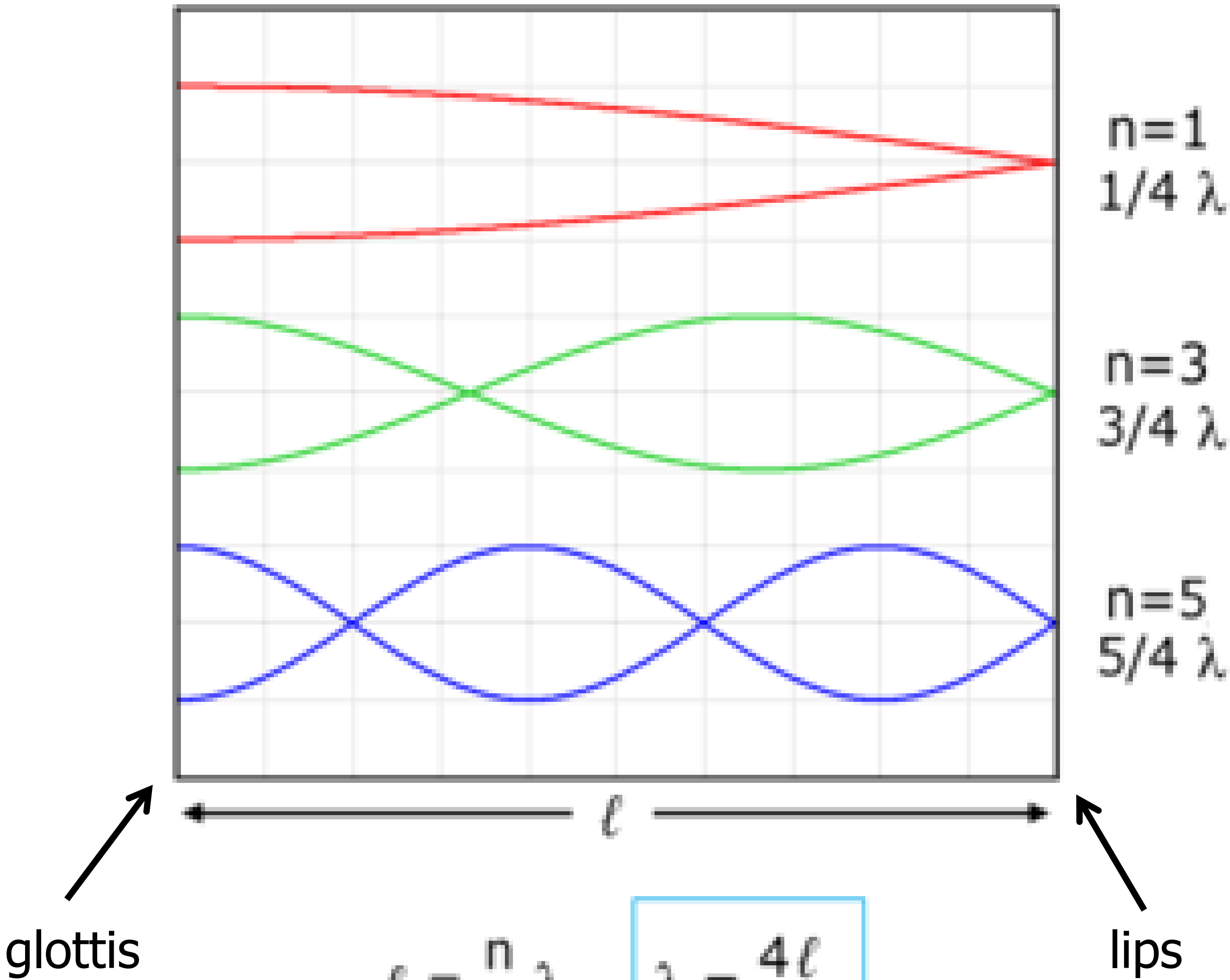


[Clark et al., 2007a, p.241]

Vibration modes: string



Vibration modes: vocal tract

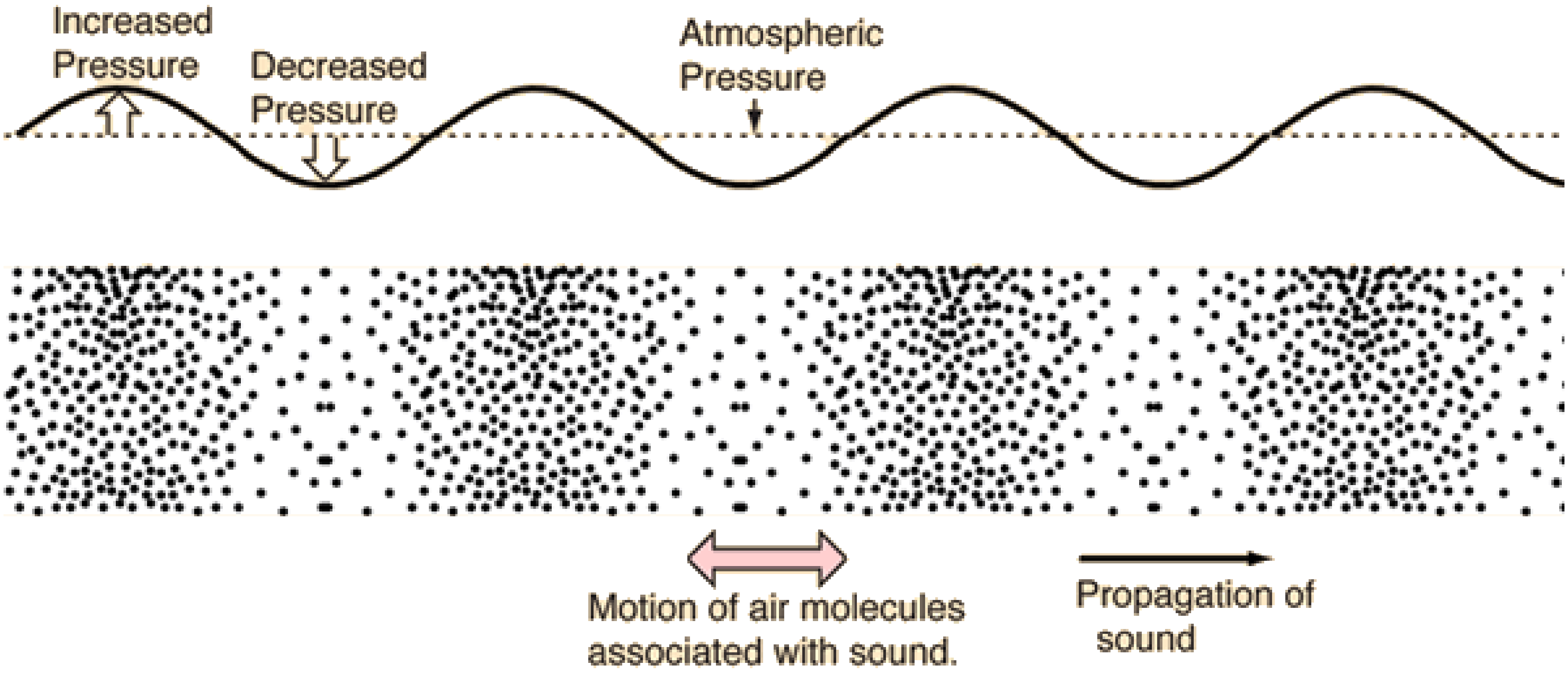


$$\ell = \frac{n}{4}\lambda \qquad \boxed{\lambda = \frac{4\ell}{n}}$$

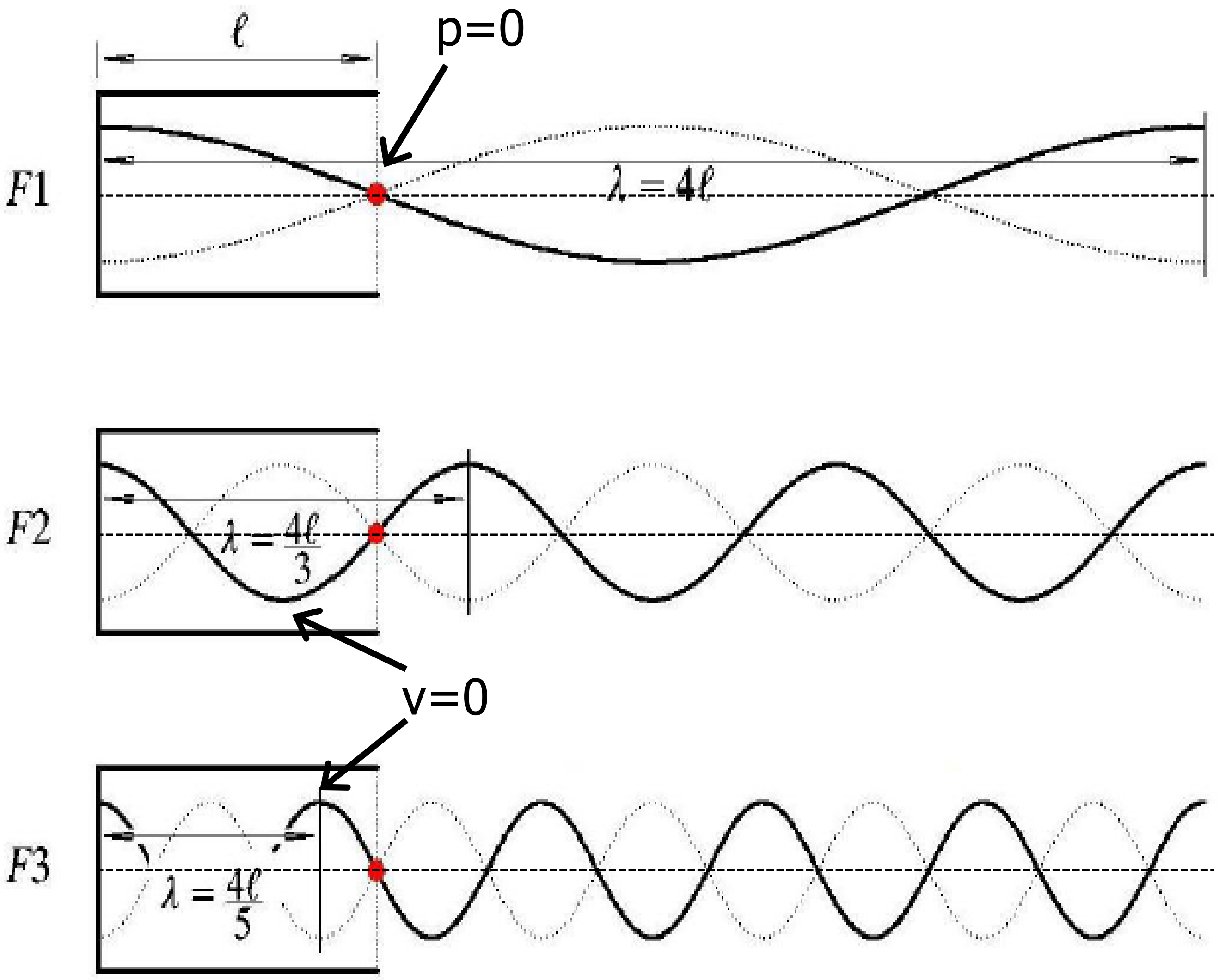
Longitudinal waves

- Acoustic signals evolve as longitudinal waves in vocal tract
- Physical parameters of acoustic waves
  - sound pressure p : change of air pressure caused by sound event, local deviation from average ambient pressure
  - sound/particle velocity v : particle velocity caused by sound event, oscillation of particle around resting position
  - speed of sound c : speed of sound waves in air (or other material), particle-to-particle interaction, distance of travel per unit of time (e.g. 340 m/s in air)

Sound propagation



Sound pressure waves in vocal tract



[Hess, ms.]

Vocal tract: acoustic tube model

- Perfect reflexion at sound-hard (lossless) walls of tube
  - $v = 0$ at place of reflexion
- (Lossy) reflexion at sound-soft transition from vocal tract to free acoustic field (i.e. from lips to air)
  - $p = 0$ at place of radiation

Computing formant frequencies

- Resonance frequencies of neutral vocal tract computed as speed of sound divided by wave length: $f_i = c / \lambda_i$

- Frequencies of resonances/formants:

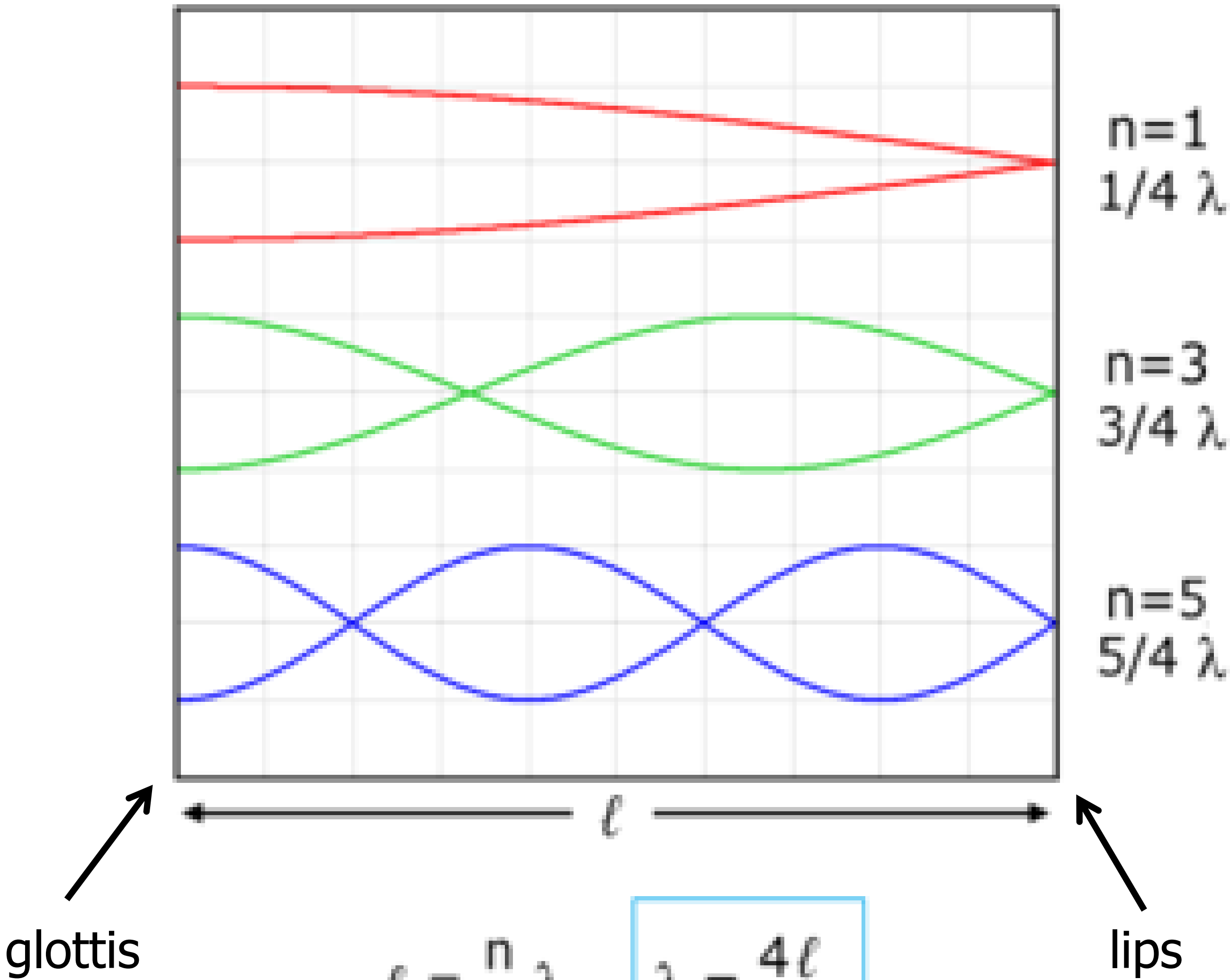
  F1 = 340 / (4 * 0.17) = 340 / 0.68 = 500 Hz

  F2 = 340 / (4/3 * 0.17) = 3 * 340 / (4 * 0.17) = 1500 Hz

  F3 = 340 / (4/5 * 0.17) = 5 * 340 / (4 * 0.17) = 2500 Hz

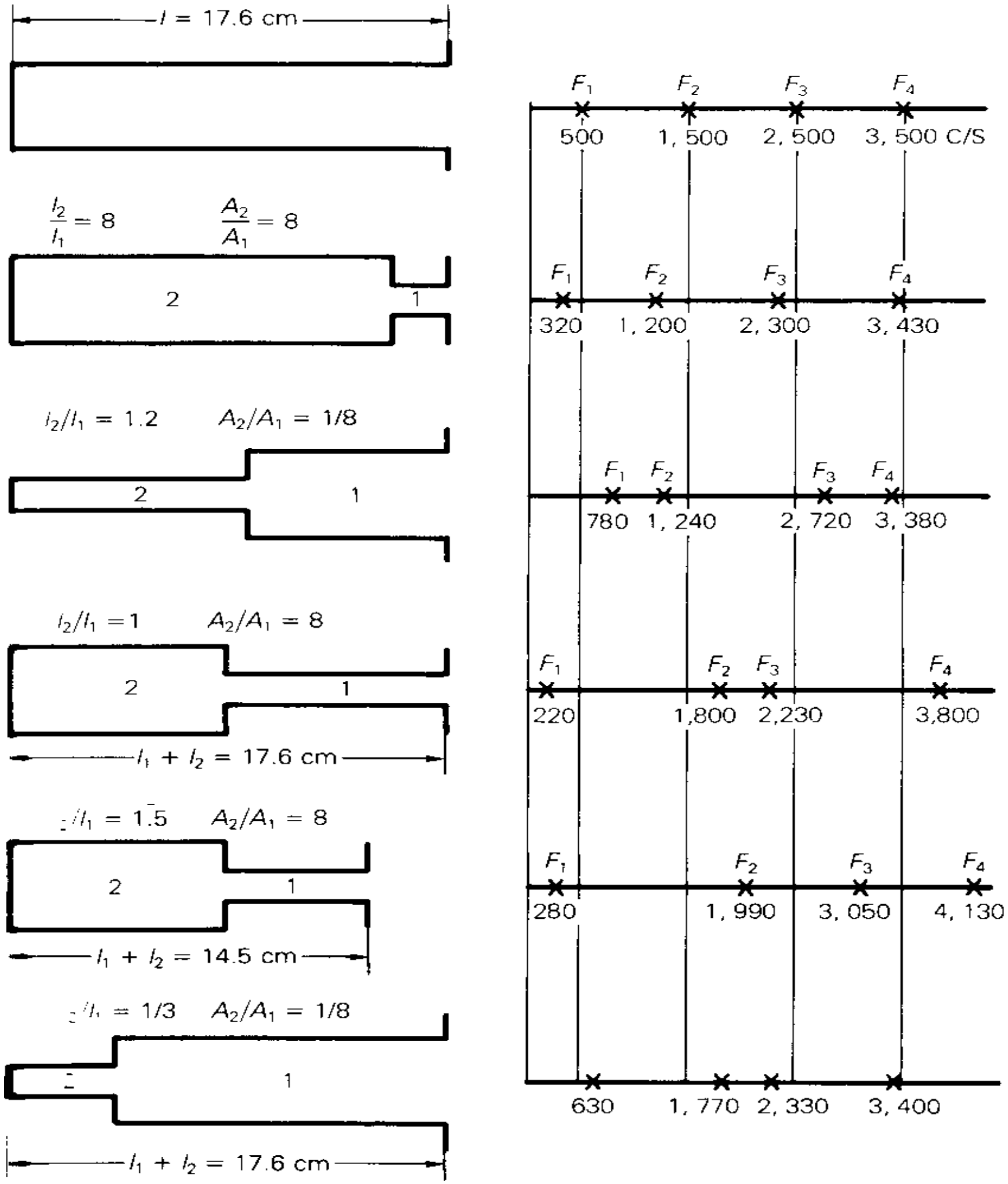
- Distribution of formant frequencies in neutral vocal tract corresponds to formants of central vowel [ə]
- Simple tube model, with constant area, is inadequate for computing formants of other vowels (cf. acoustic theory of vowel articulation [Ungeheuer 1962])

Vibration modes: vocal tract (repeated)



$$\ell = \frac{n}{4}\lambda \qquad \boxed{\lambda = \frac{4\ell}{n}}$$

Tube model with variable area



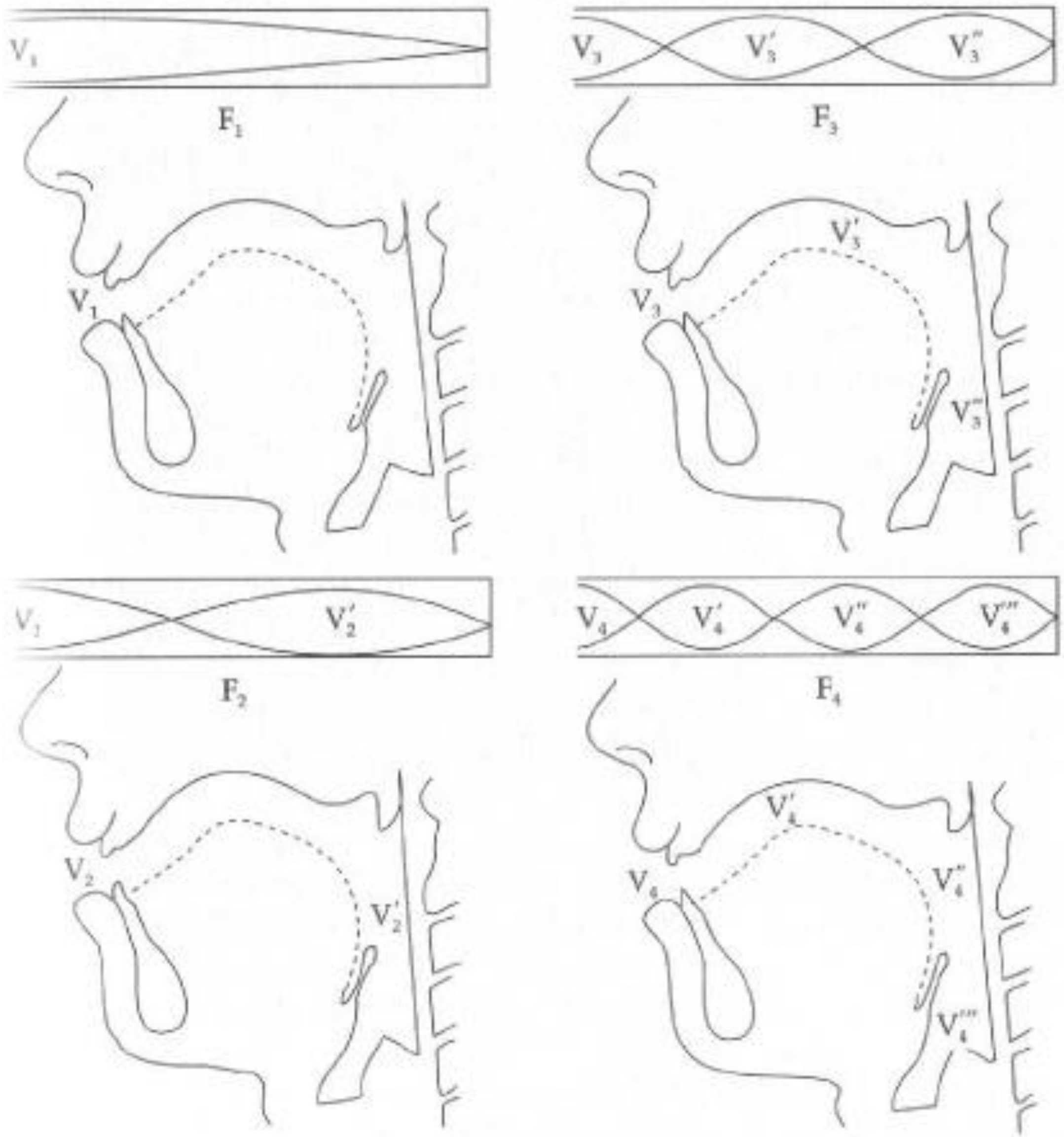
[Clark et al., 2007a, p.246]

UNIVERSITÄT DES SAARLANDES

T. Arai's cylinder-type models

[http://www.splab.net/Vocal_Tract_Model/index-e.htm]

Resonances: standing waves



parameter: v [Johnson, 1997, p.99]

Standing waves: interpretation

- interpretation of the graphical representation of standing waves in idealized vocal tract (neutral configuration, see previous figure):
- first 4 formants displayed (F1 – F4)
- in tube model and in vocal tract
- places of maximum sound velocity (sound velocity nodes, V_i)
- places of maximum sound pressure (wave maxima, "antinodes")
- localization of V_i in vocal tract

Dynamic area changes

- resonances of vocal tract with variable area cannot be straightforwardly visualized as in the neutral tube model
  - local area changes affect frequencies of resonances, depending on energy distribution of standing wave in tube along longitudinal axis ("z-axis")
  - e.g., constriction at lip end of tube has same effect as constriction at glottis end: lower resonance frequency
  - acoustic vowel system can be interpreted as representing geometrical changes with respect to neutral tube geometry and resulting changes of resonance frequencies away from neutral values

    $\rightarrow$ acoustic theory of vowel articulation [Ungeheuer (1962)]

Acoustic theory of vowel articulation

2.3.1 Ausgangspunkt Webster'sche Horngleichung (nach Ungeheuer, 1962)

Wir gehen nun von der Wellengleichung des Schnellenpotentials Φ für die Wellenausbreitung in einem Rohr veränderlichen Querschnittes, der sog. Webster'schen Horngleichung aus

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{1}{A}\frac{\partial \Phi}{\partial x}\frac{dA}{dx} = \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} \tag{45}$$

mit den bekannten Randbedingungen:

$$v(t) = 0 \quad \Rightarrow \quad \frac{\partial \Phi}{\partial x} = 0 \quad [\text{Glottis}, \; x = 0] \tag{46}$$

$$p(t) = 0 \quad \Rightarrow \quad \Phi = 0 \quad [\text{Mundöffnung}, \; x = l] \tag{47}$$

Mit Hilfe der Trennung der Variablen

$$\Phi(x,t) = \varphi(x) \cdot \psi(t) \tag{48}$$

können wir (45) schreiben

$$\frac{1}{\varphi}\left[\frac{d^2\varphi}{dx^2} + \frac{1}{A}\frac{d\varphi}{dx}\frac{dA}{dx}\right] = \frac{1}{c^2\psi}\,\frac{d^2\psi}{dt^2} \tag{49}$$

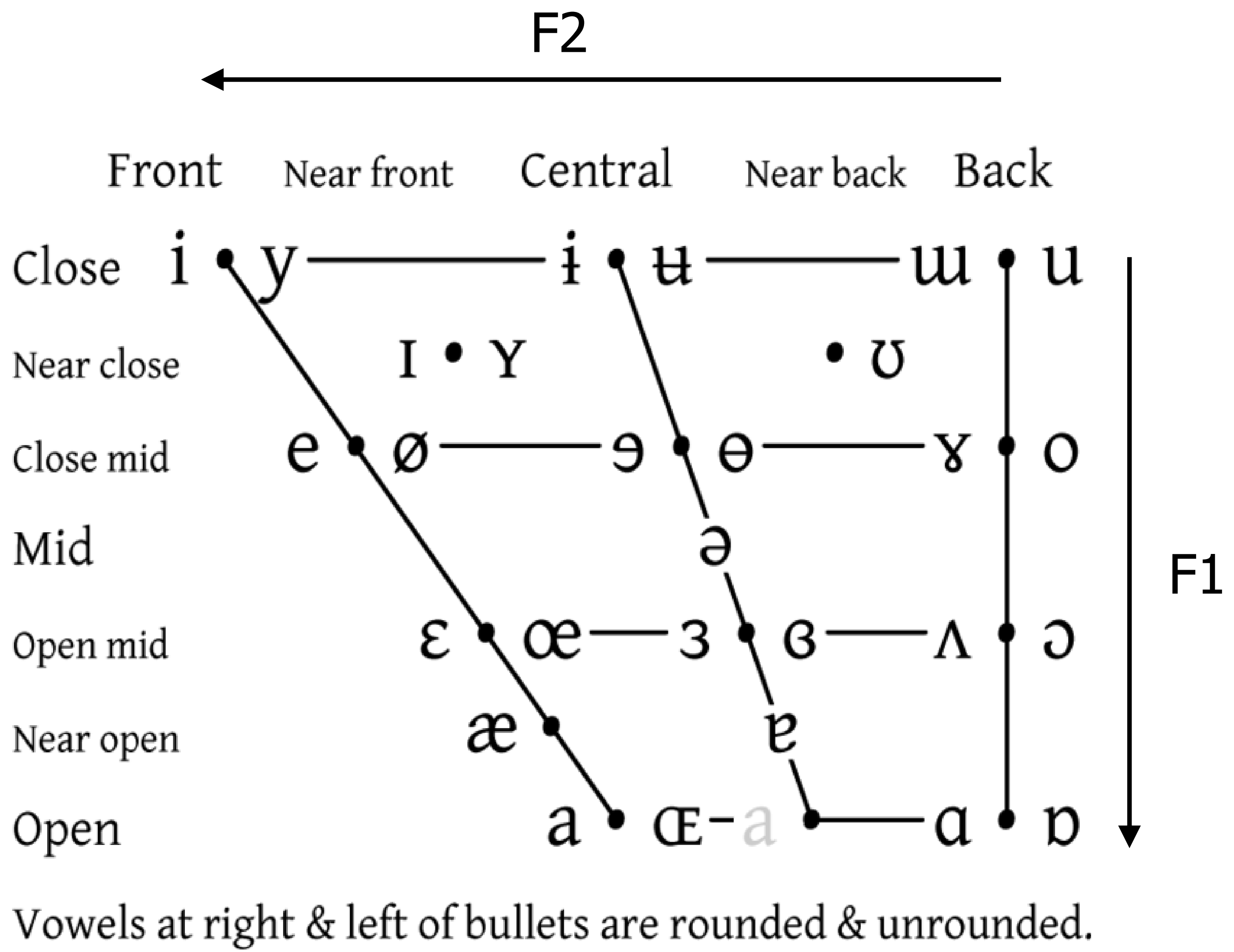
Die linke Hälfte hängt nur von x ab, die rechte nur von t. Damit können beide als gleich einer Konstante gesehen werden, die mit $-\Lambda$ bezeichnet sei:

$$\frac{1}{\varphi}\left[\frac{d^2\varphi}{dx^2} + \frac{1}{A}\frac{d\varphi}{dx}\frac{dA}{dx}\right] = -\Lambda = \frac{1}{c^2\psi}\,\frac{d^2\psi}{dt^2} \tag{50}$$

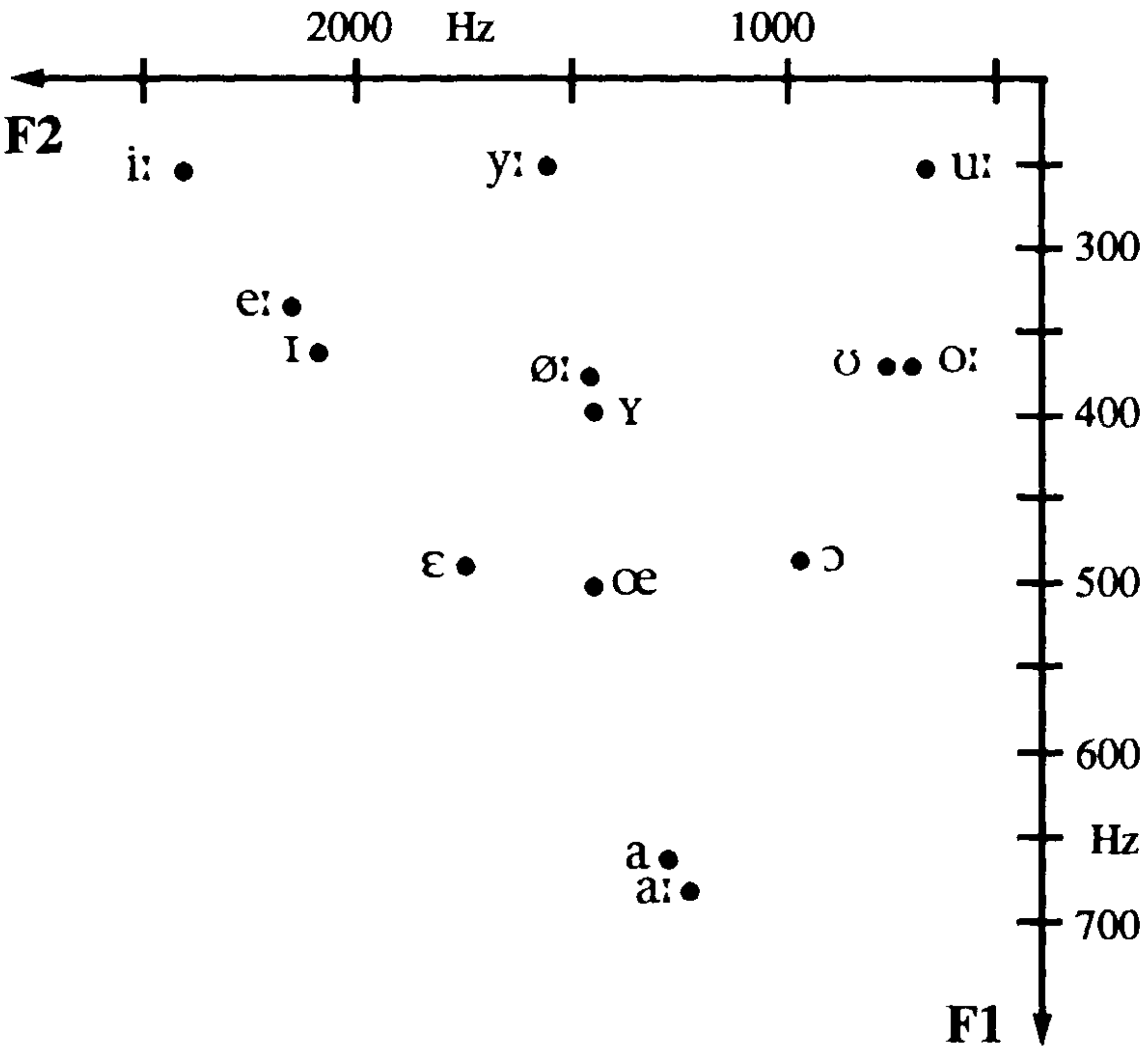
NOT REQUIRED

UNIVERSITÄT DES SAARLANDES

Vowels (IPA)

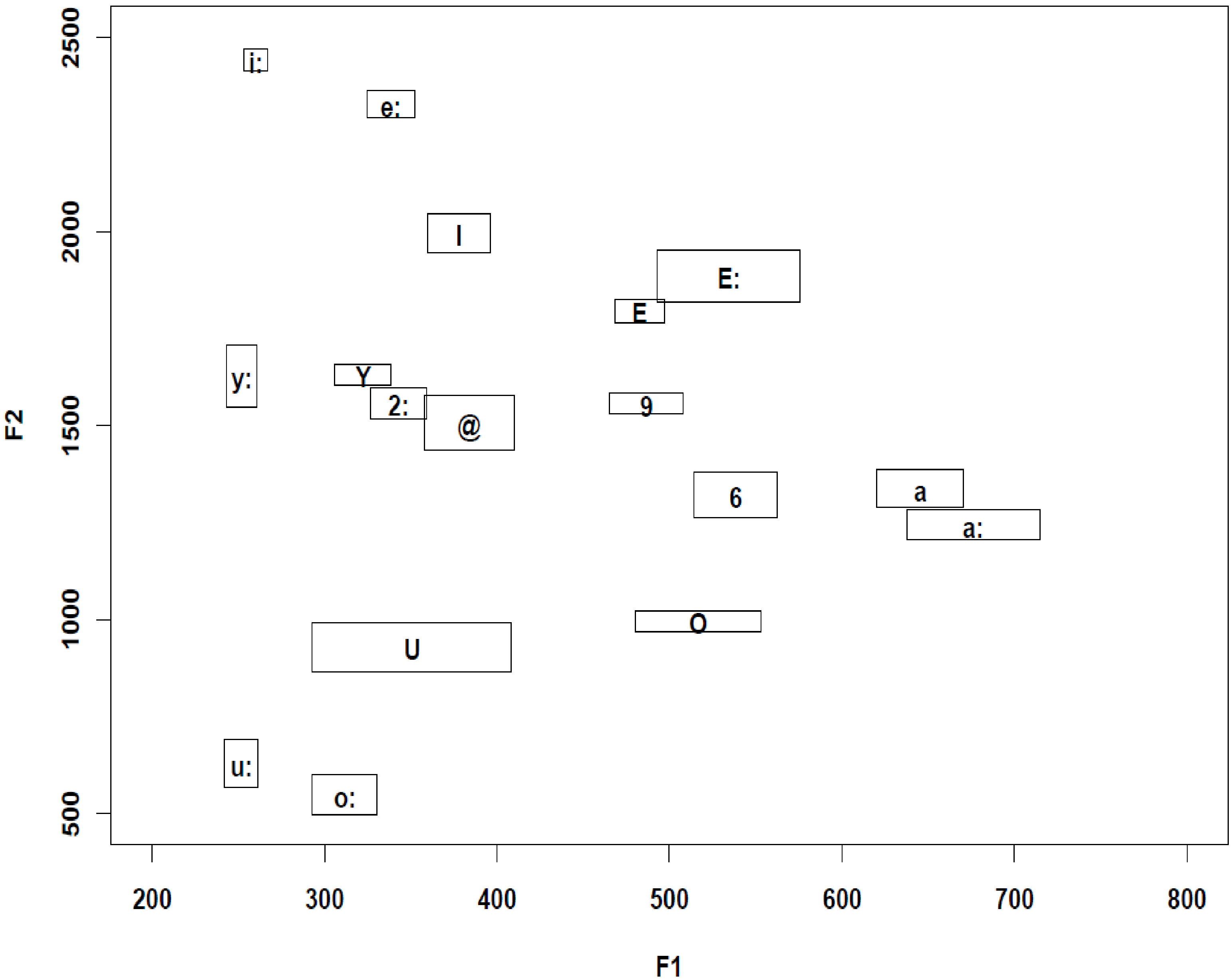

F2
Front
Near front
Central
Near back
Back
Close
Near close
Close mid
Mid
Open mid
Near open
Open
i • y
ɨ • ʉ
ɯ • u
ɪ • ʏ
• ʊ
e • ø
ɘ • ɵ
ɤ • o
ə
ɛ • œ
ɜ • ɞ
ʌ • ɔ
æ
ɐ
a • ɶ
ä
ɑ • ɒ
F1
Vowels at right & left of bullets are rounded & unrounded.

Vowels (German [Pompino-Marschall, 1995])



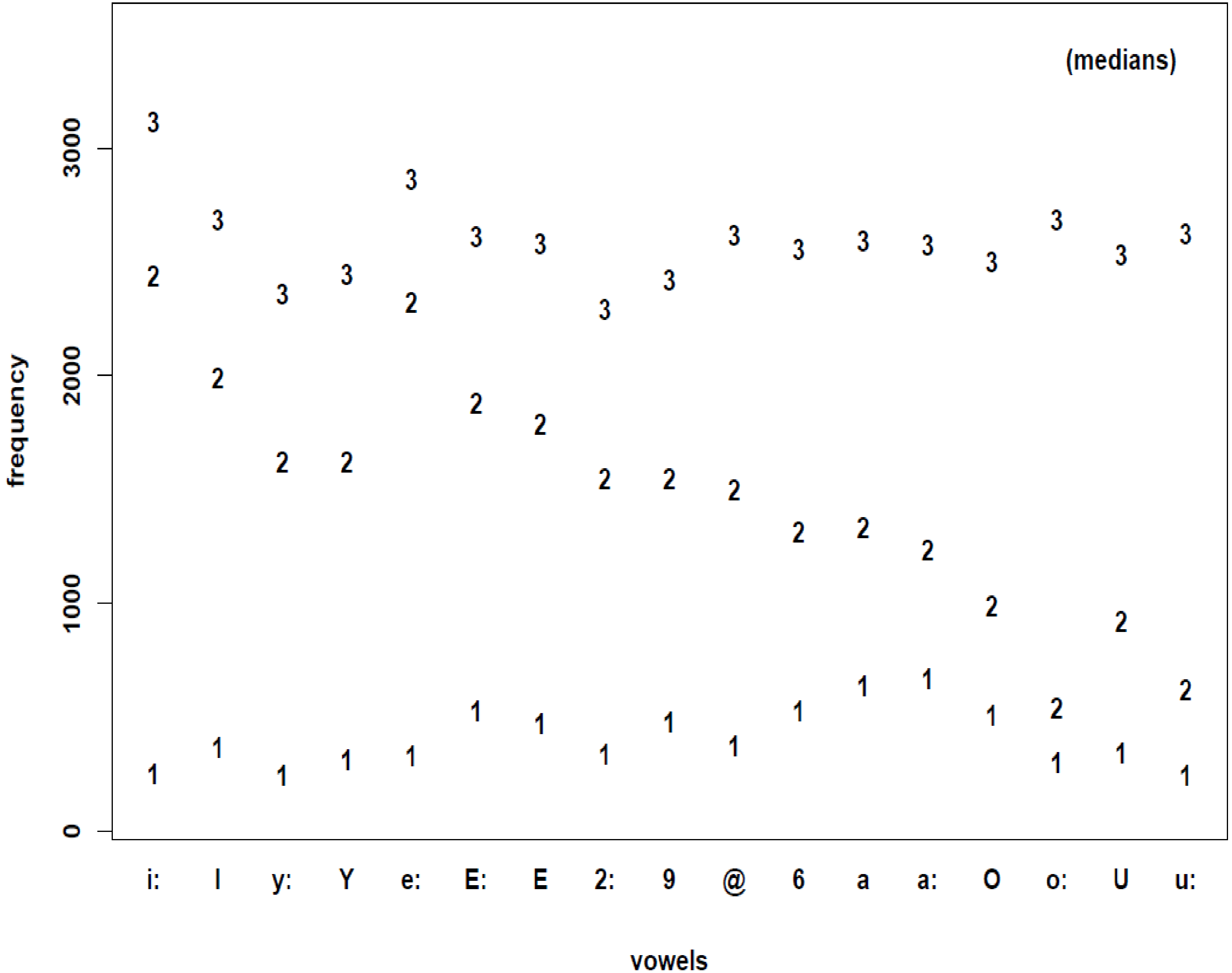
UNIVERSITÄT DES SAARLANDES

Vowels (German [Möbius, 2001])

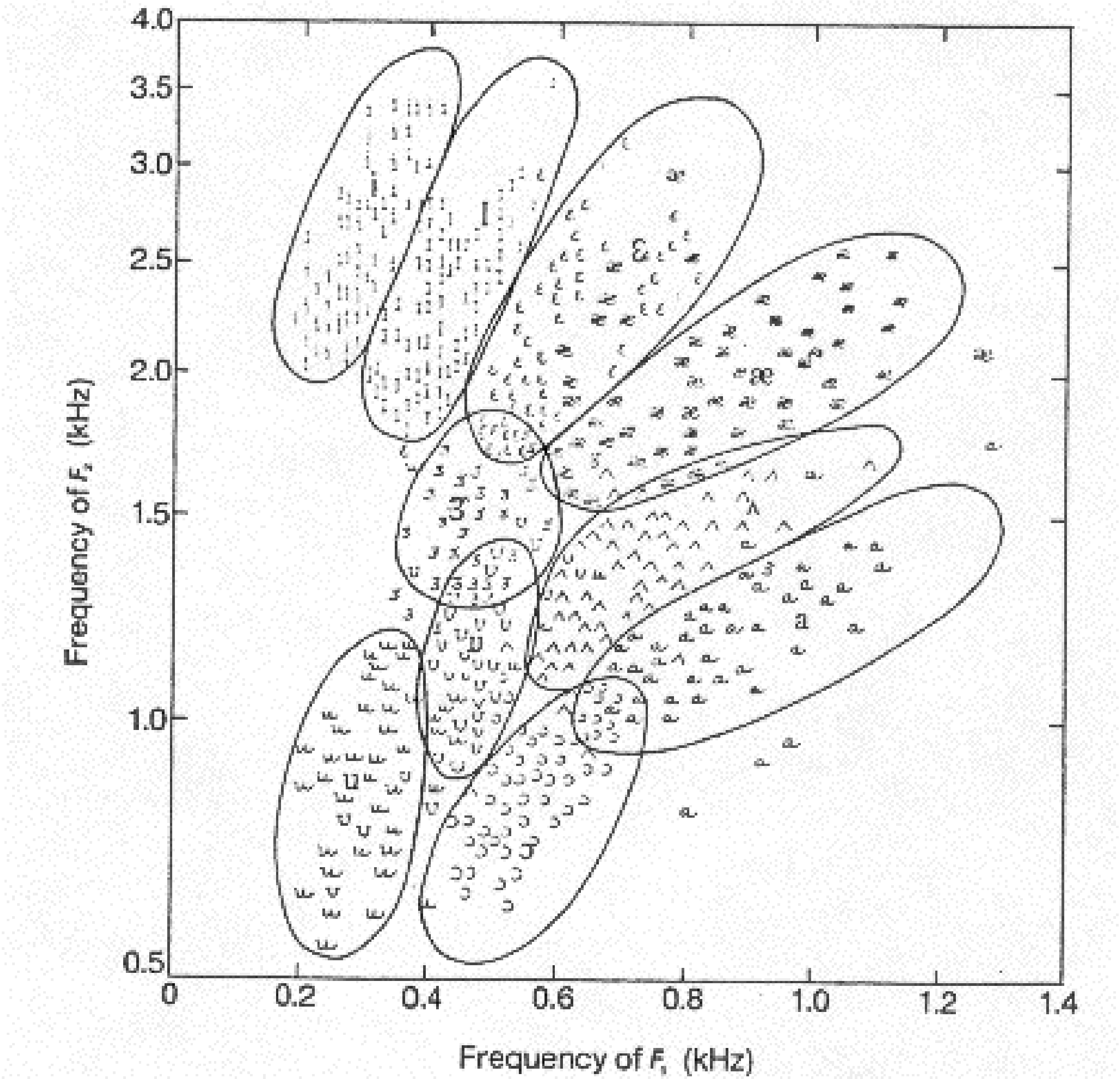


UNIVERSITÄT DES SAARLANDES

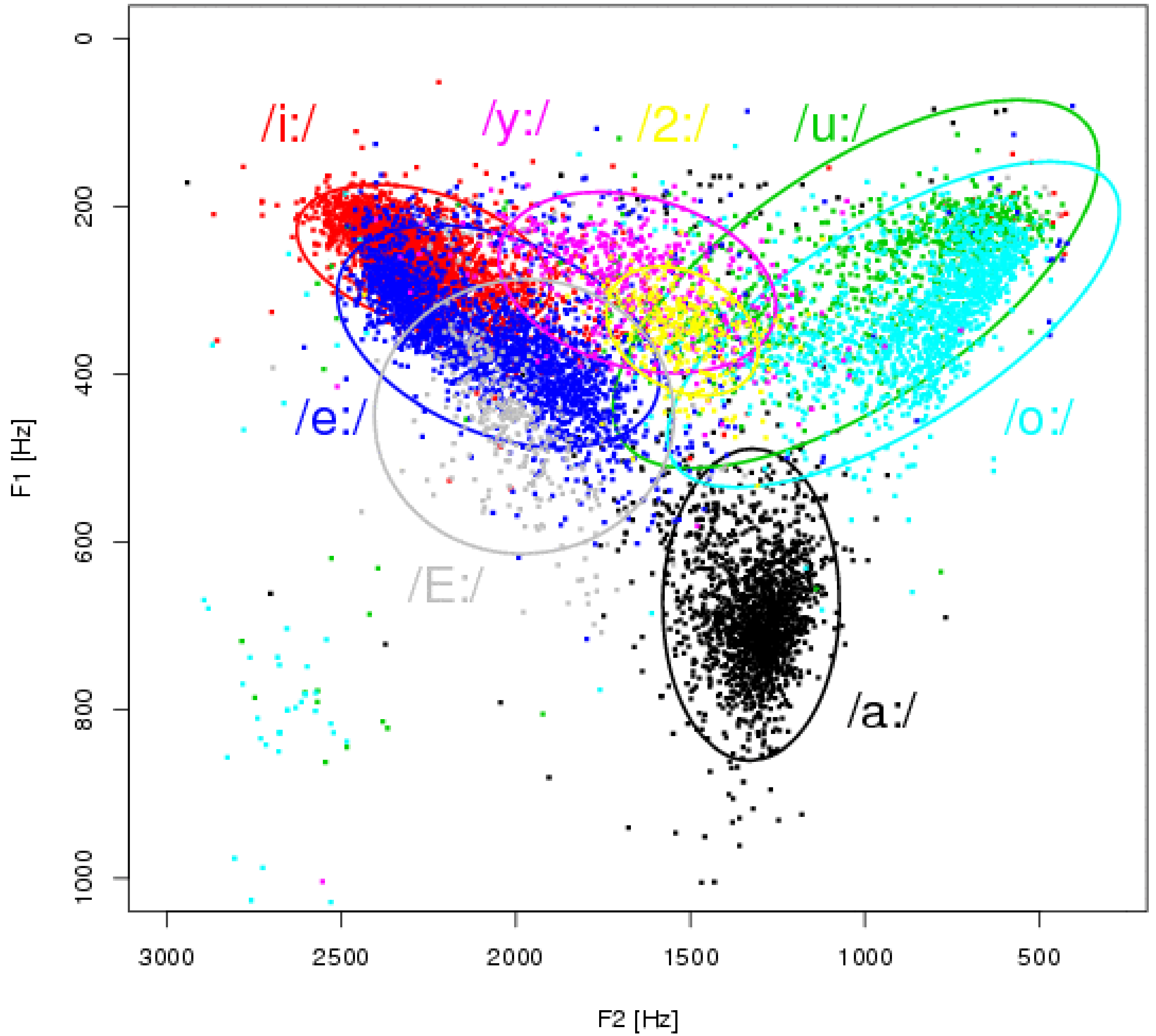
Vowels (German, F1/F2/F3 [Möbius, 2001])



Vowels (Am. English [Peterson and Barney, 1952])

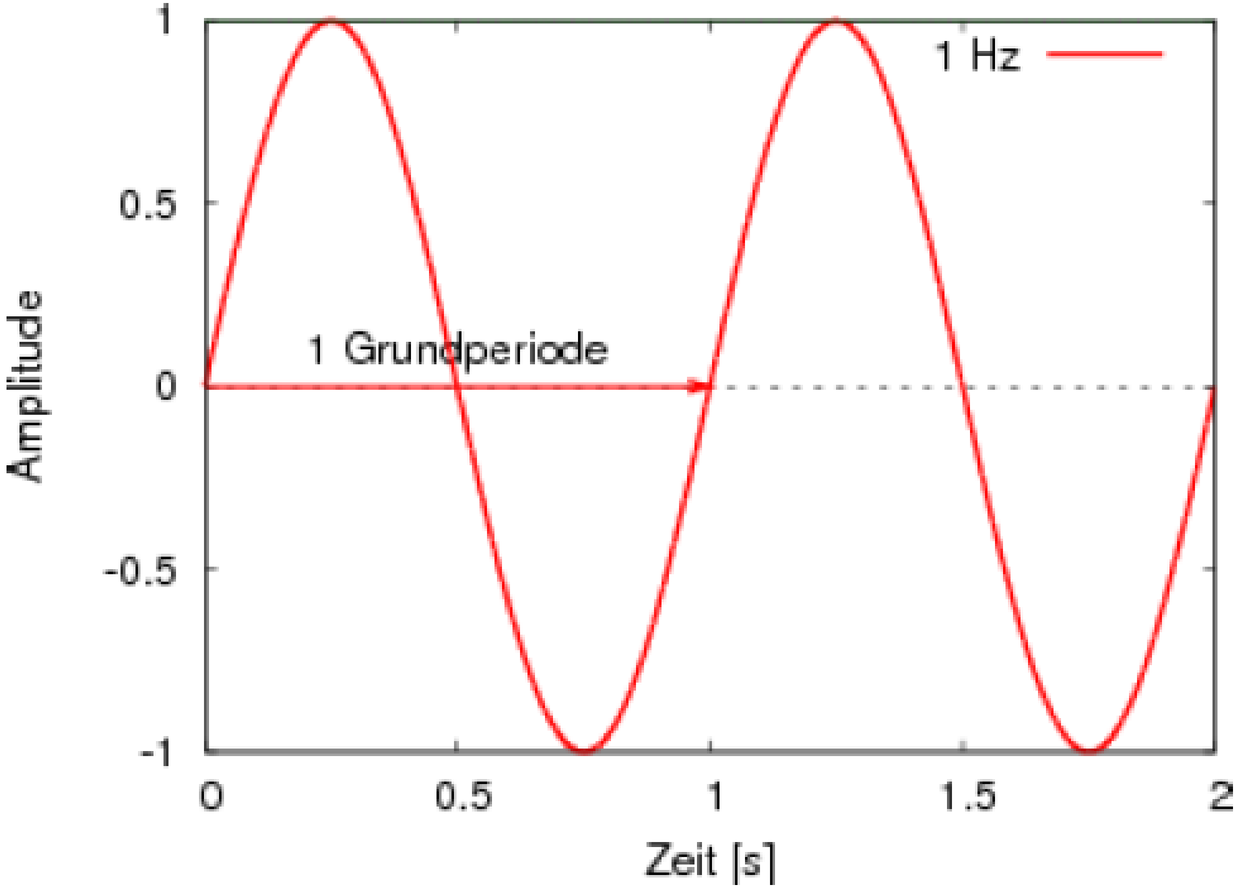


Vowels (German [Möbius])



UNIVERSITÄT DES SAARLANDES

Simple waveforms



UNIVERSITÄT DES SAARLANDES

Simple waveforms

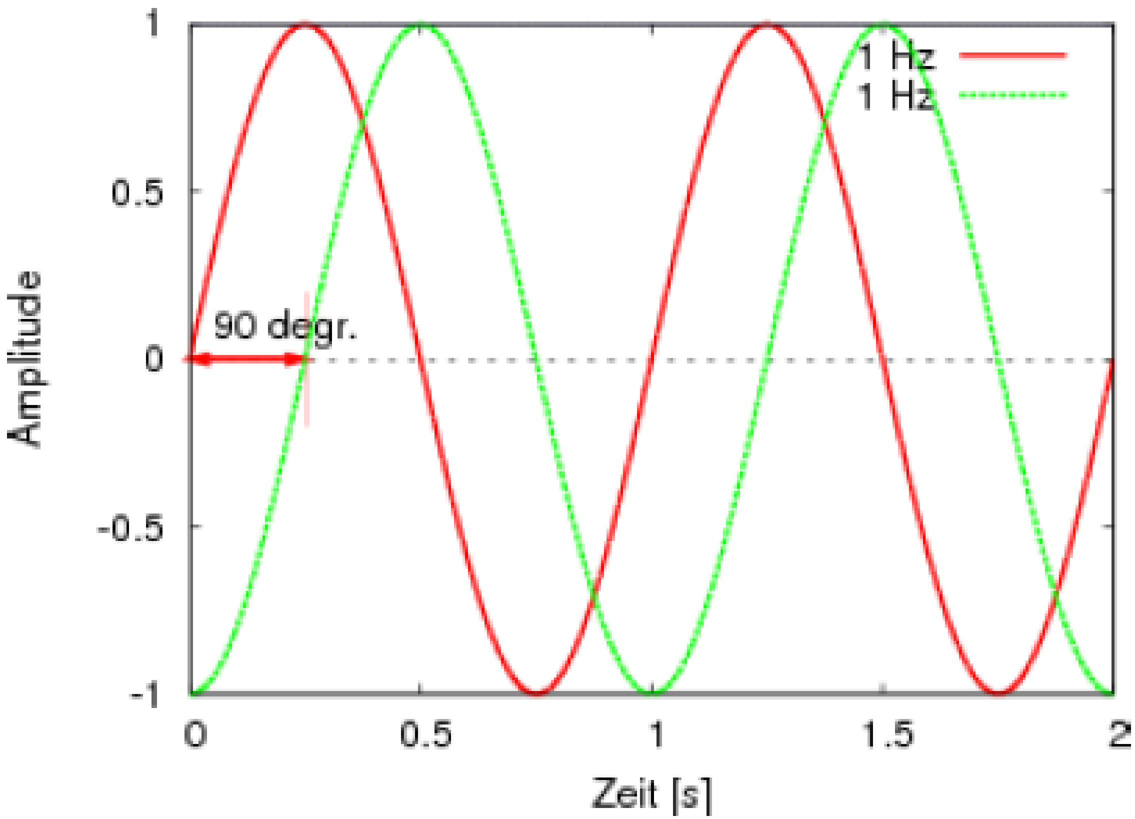
- Simple periodic oscillation: pure sine wave
  - cyclically recurring, simple oscillation pattern, determined by
    - fundamental period T_0
    - amplitude A
    - phase Φ
- Fundamental frequency [Hz]: 1 / fundamental period [s]

  $F_0 = 1 / T_0$

Simple waveforms

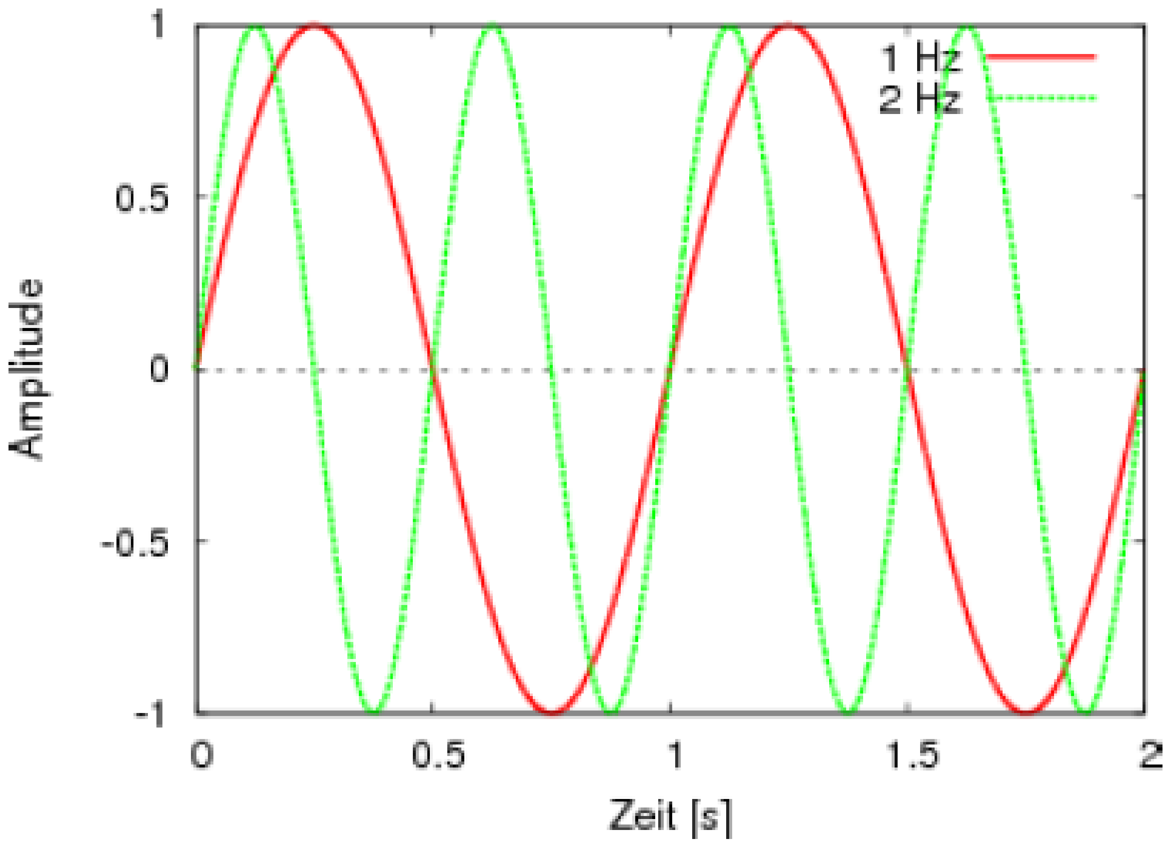
- Phase relation
  - two sine waves of same frequency and amplitude, but temporally displaced maxima, minima, and zero crossings

  → phase shift (here: angle 90º)



Simple waveforms

- Frequency differences
  - two sine waves of same amplitude and phase, but different frequency (here: 1 vs. 2 Hz)

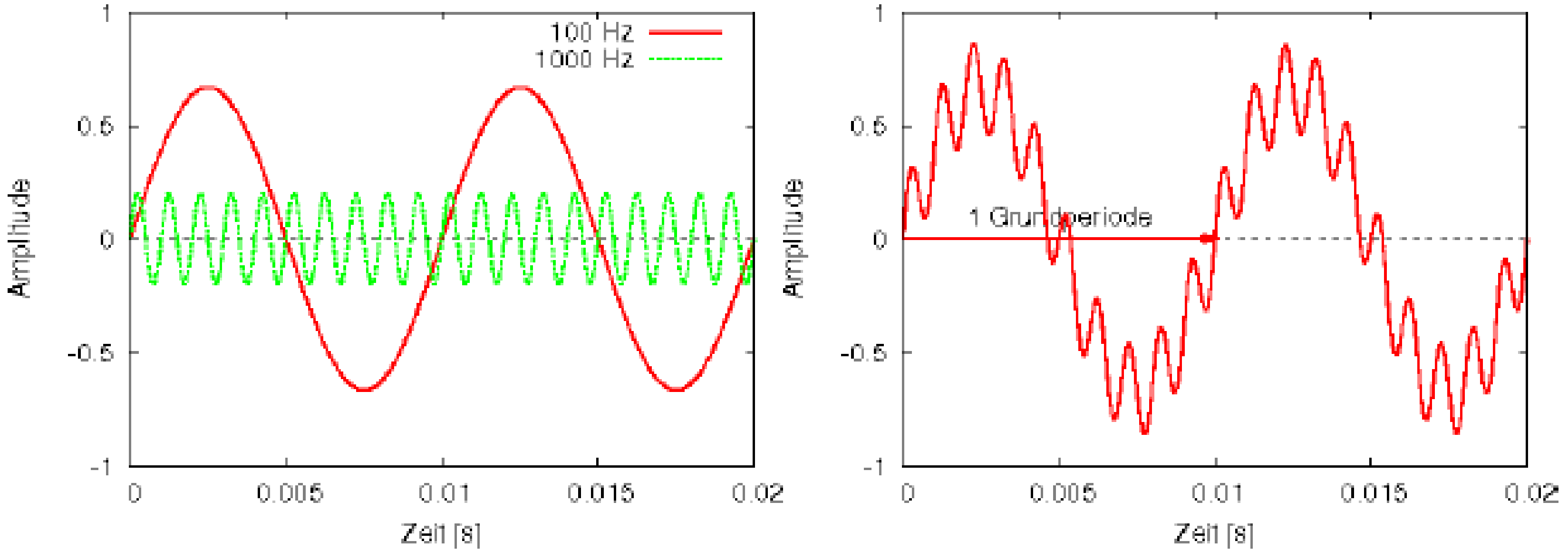


Complex waveforms

- Complex periodic signals
  - cyclically recurring oscillation patterns
  - composed of at least two sine waves
  - fundamental frequency = 1 / complex fundamental period
- Form of resulting complex wave depends on frequency, amplitude and phase relations between component waves

Complex waveforms

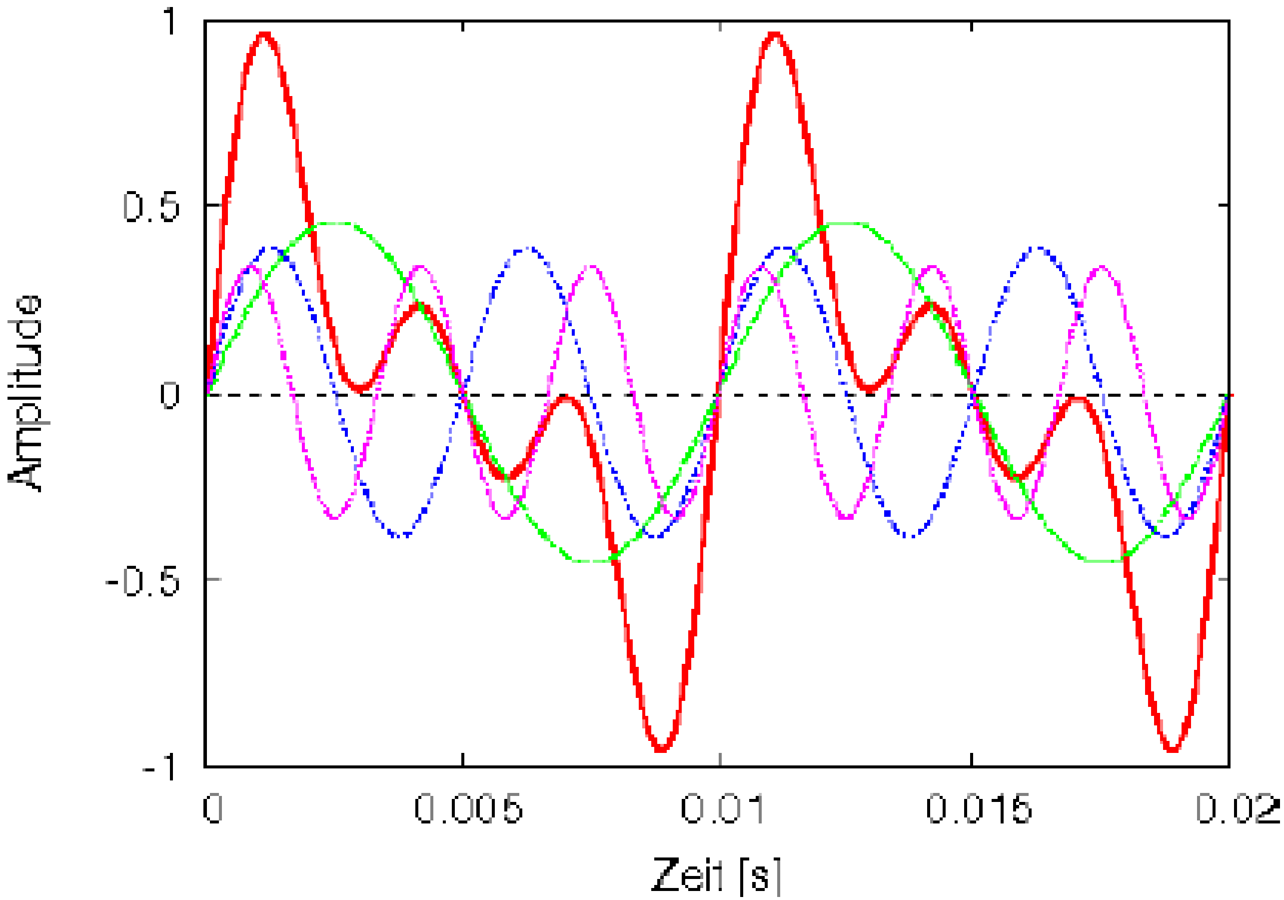
- Complex waveform: 2 components
  - two sine waves (100 Hz, 1000 Hz) with same phase and different amplitude (left)
  - complex wave (right) resulting from addition of the two components
  - F_0 = 100 Hz



UNIVERSITÄT DES SAARLANDES

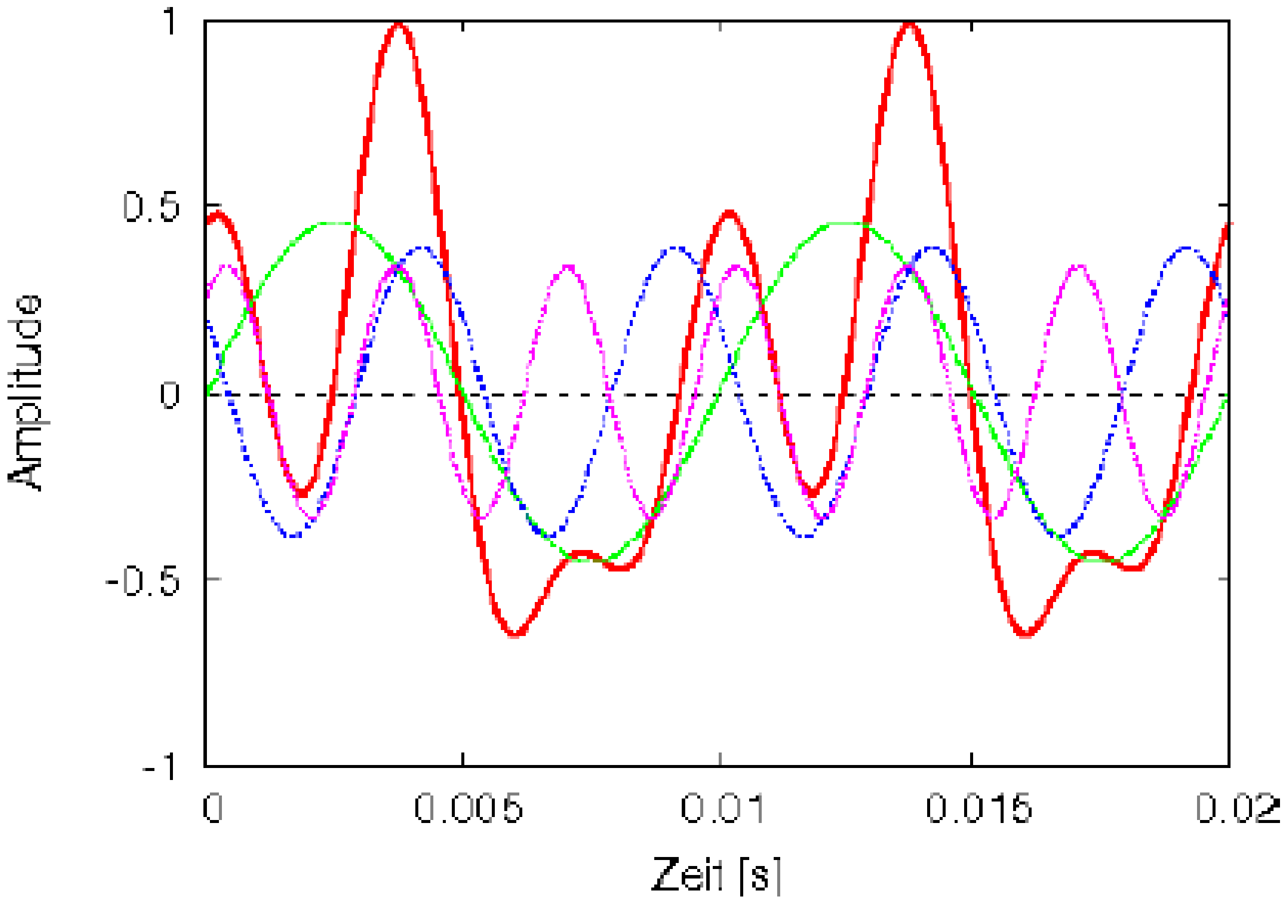
Complex waveforms

- Complex waveform (red): 5 components
  - five sine waves (100, 200, 300, 400, 500 Hz) with same phase
  - only 3 lowest frequency components displayed



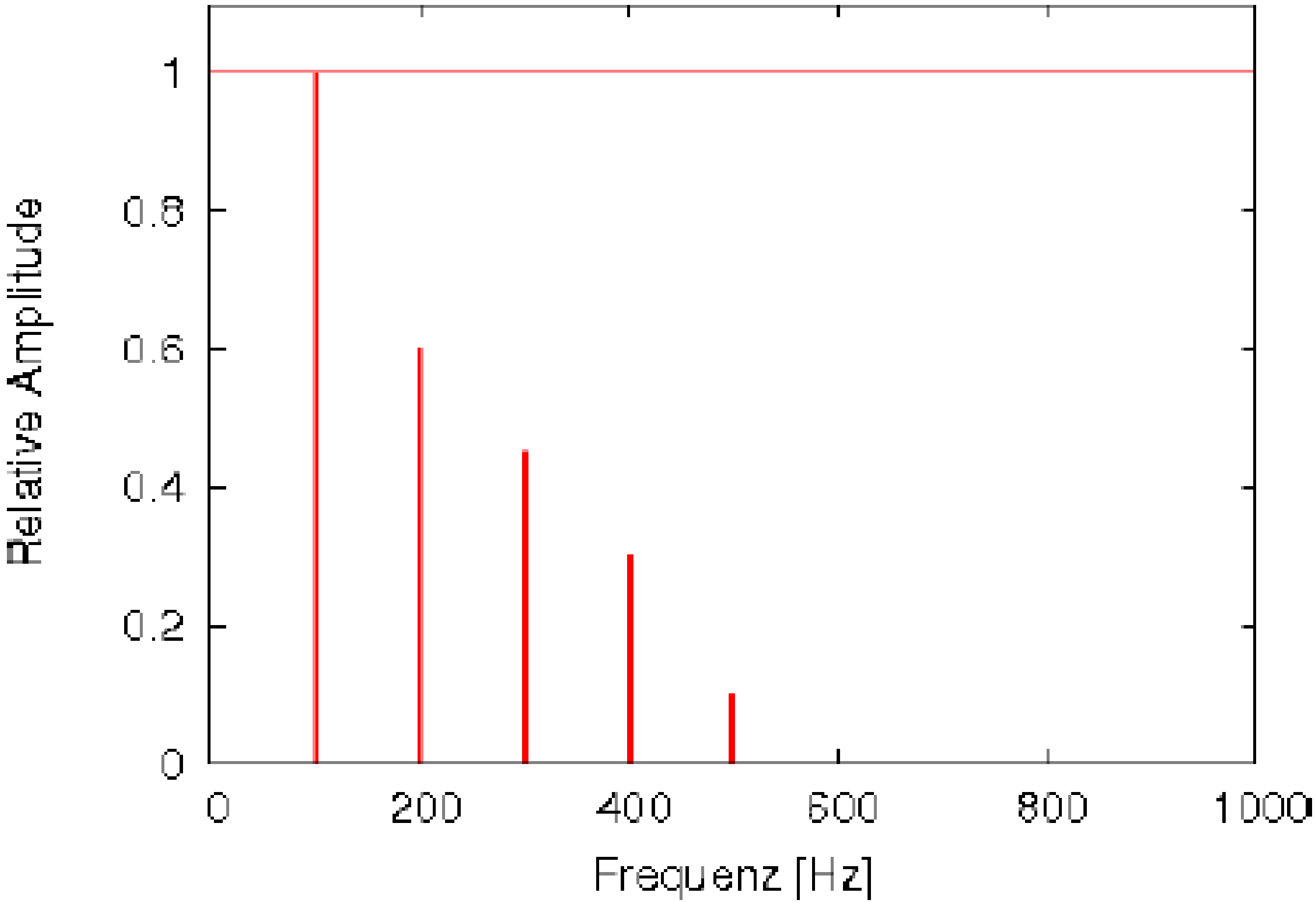
Complex waveforms

- Complex waveform (red): 5 components
  - five sine waves (100, 200, 300, 400, 500 Hz) with phase shifts
  - only 3 lowest frequency components displayed



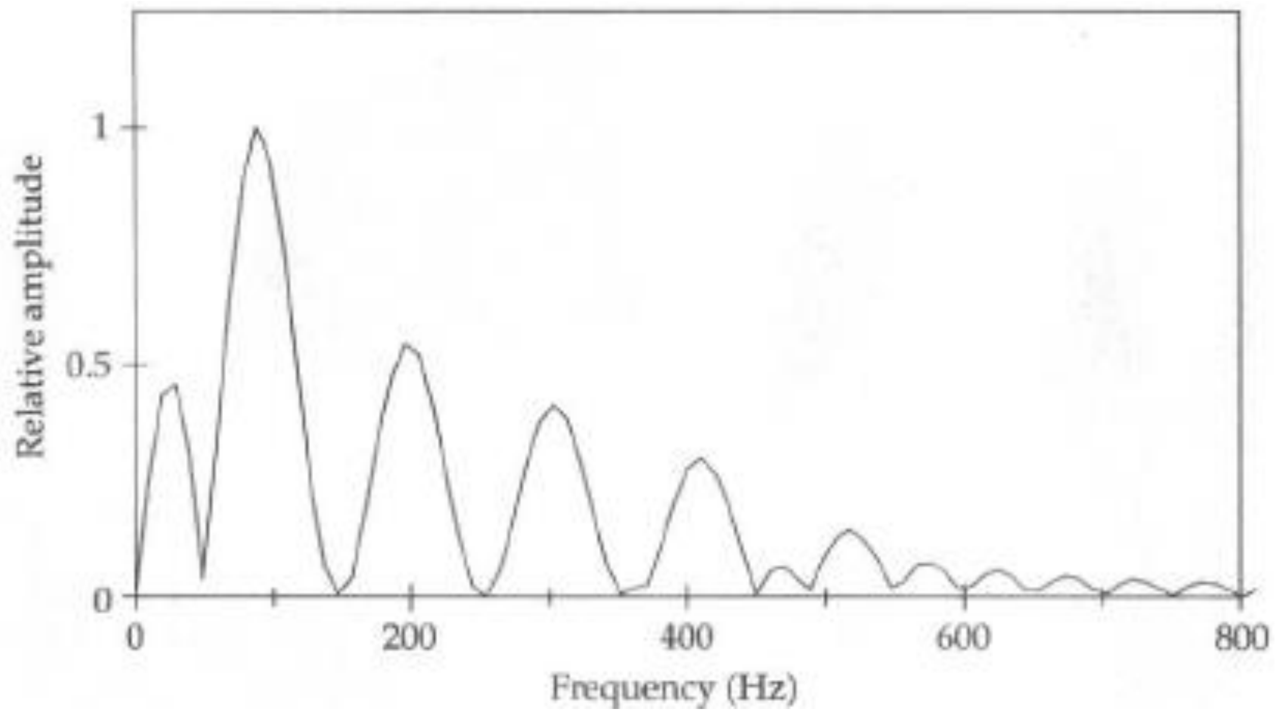
Power/line spectrum

- Line spectrum (amplitude over frequencies) of the complex waveform composed of five components (see above)



Fourier analysis

- Fourier's theorem: every complex wave can be analytically decomposed into a set of sine waves, each with specific values of frequency, amplitude and phase.



Fourier analysis: power spectrum of 5 component wave

Fourier analysis and power spectrum

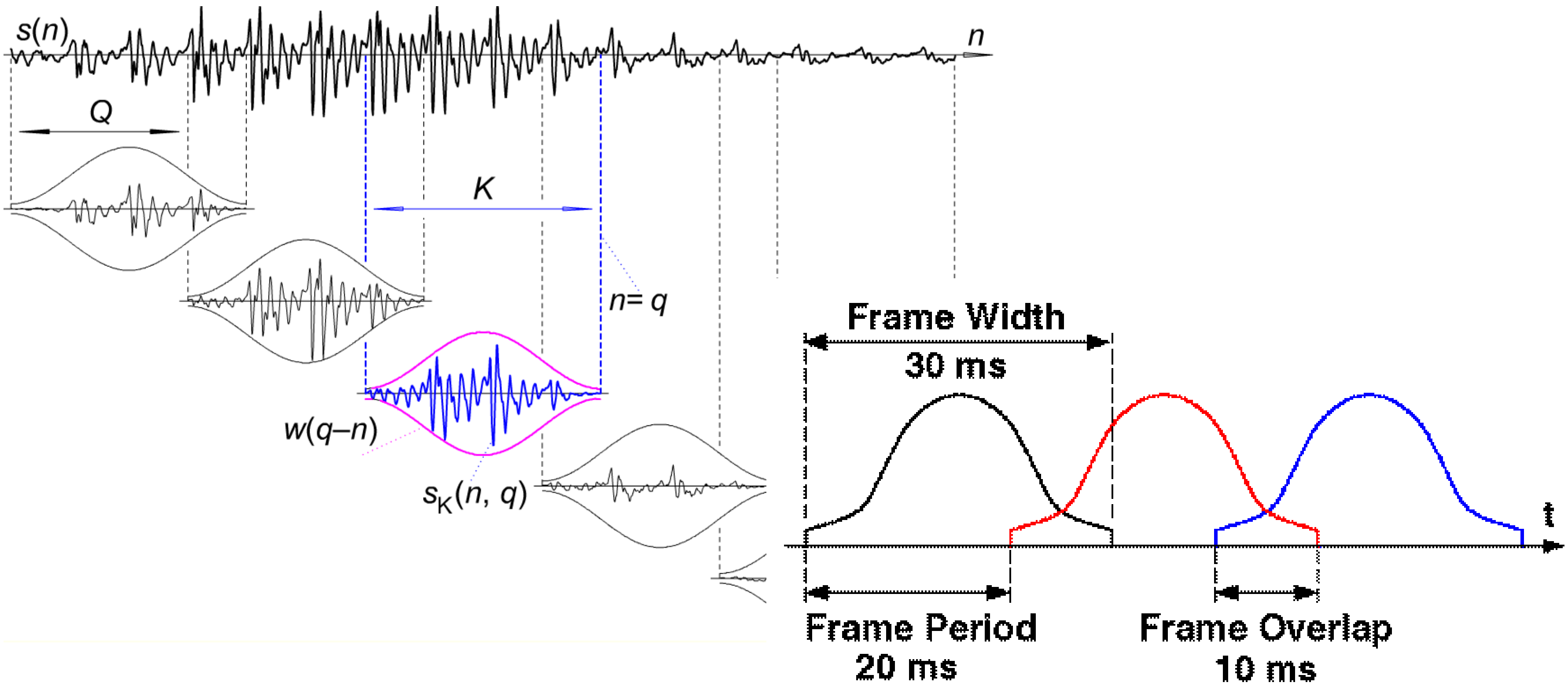
- Differences between result of Fourier analysis (Fast Fourier Transform, FFT) and idealized line spectrum:
  - broader peaks rather than lines
  - additional peaks (number of components is a parameter!)
- Reasons for these differences:
  - Fourier analysis assumes infinitely long signal, whereas analysis is performed over a few fundamental periods
  - Fourier analysis assumes periodicity, whereas speech signals are quasi-periodic, changing slowly from one fundamental period to the next, or even stochastic
  - digital (discrete) rather than analog (continuous) signal

Discrete Fourier Transform

- Discrete Fourier analysis (Discrete Fourier Transform, DFT)
  - digital Fourier analysis of complex signals, yielding a spectrum of sine wave components
  - transformation of data from time domain into frequency data
  - resolution parameters
    - sampling rate, e.g. 16000 Hz
    - window size (or frame length), e.g. 512 samples ~ 32 ms (512/16000)

Analysis window

- Windowing: splitting the input signal into temporal segments
  - window functions, e.g. Hamming window, cosine window, ...



UNIVERSITÄT DES SAARLANDES

Speech signal processing

- Typical parameter values in speech signal processing applications:
  - window length: 25 or 40 ms, 512 or 1024 samples (FFT)
  - window step size: 10 ms
    - resulting in a series of n-dimensional feature vectors, one vector every 10 ms

- Granularity of computed spectrum ca. 31 Hz (16000/512=31.25)
- Trading relation (uncertainty principle)
  - good frequency resolution $\leftrightarrow$ poor time resolution
  - good time resolution $\leftrightarrow$ poor frequency resolution

From spectrum to spectrogram

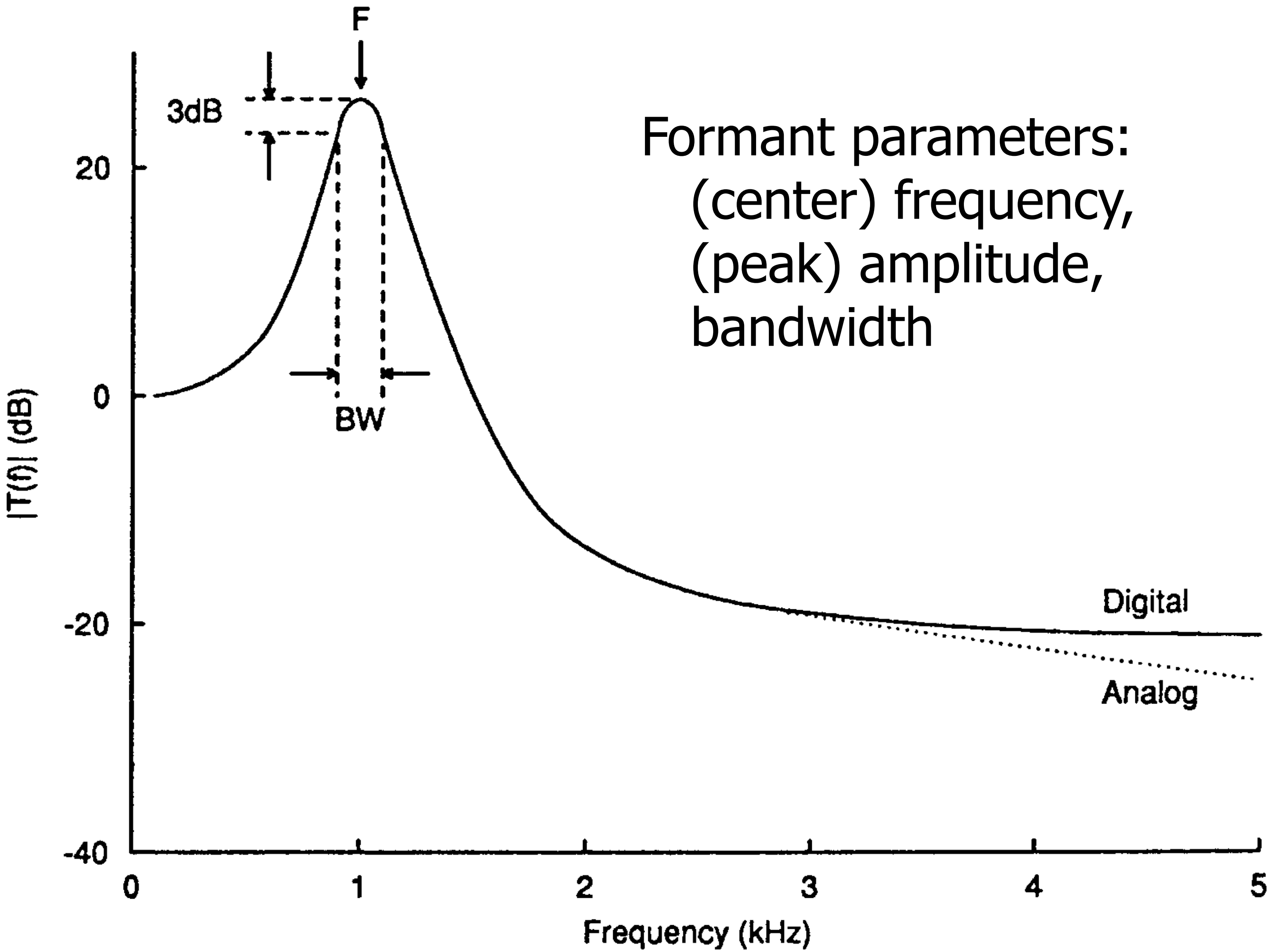
- Power spectrum:
  - snapshot taken at a specific instant of time in the speech signal

- Spectrogram:
  - narrow band spectrogram (e.g. 31 Hz): good frequency resolution
  - wide band spectrogram (e.g. 300 Hz): good temporal resolution
  - analysis window size/length:
    - short temporal window: good time resolution
    - long temporal window: good frequency resolution

Vocal tract vs. lossless tube

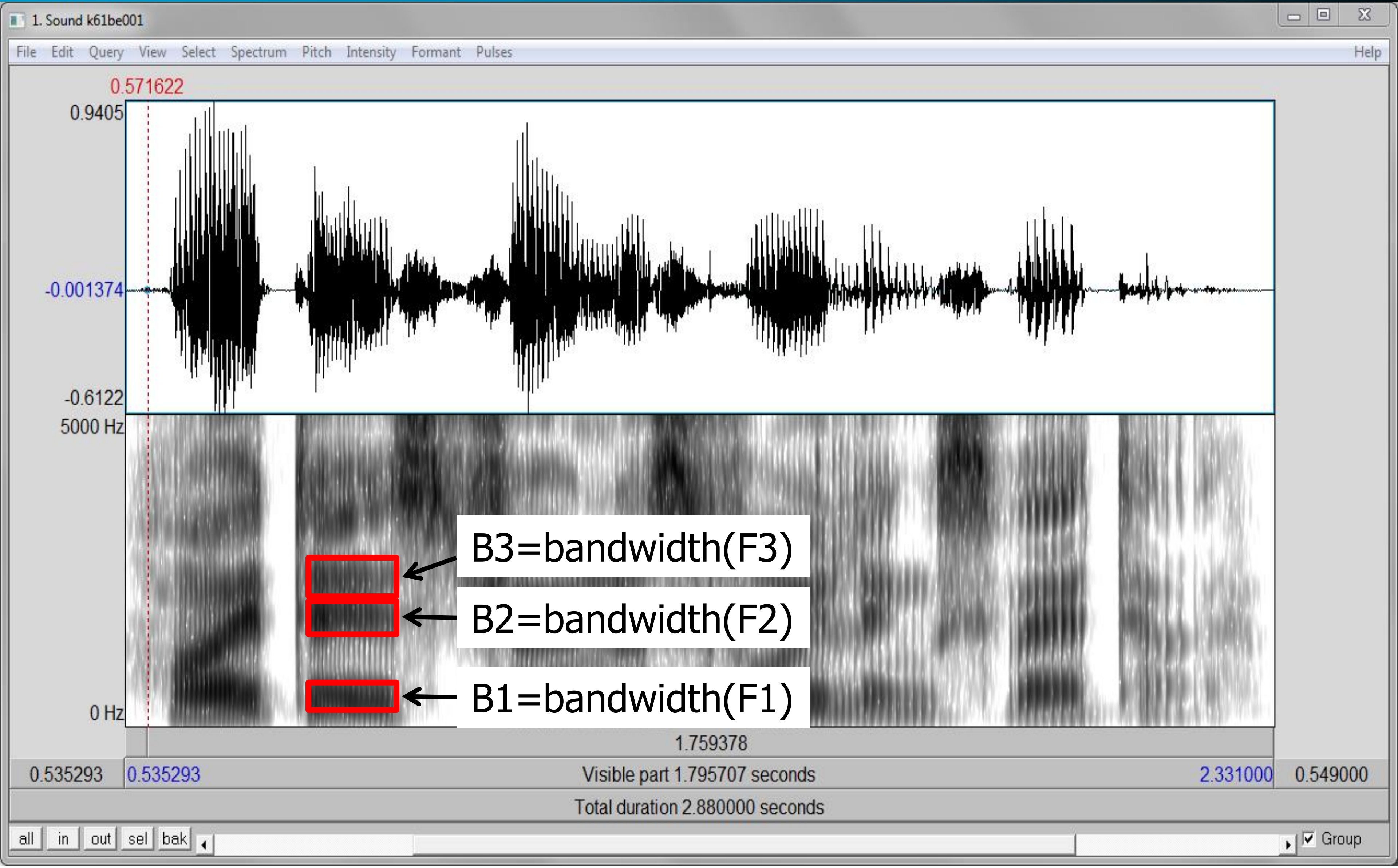
- losses in the vocal tract caused by
  - friction between air particles
  - vibration of vocal tract walls
  - viscosity of vocal tract tissue
  - radiation of sound energy into free acoustic field
- lossy vibrations are damped exponentially
- spectral equivalent of damping: **bandwidth**
  - defined as frequency range comprising 50% of power
  - corresponding to decrease of amplitude by 3 dB (or 0.707*A)
    - sound energy expressed in [dB]
    - sound energy is proportional to square of amplitude
    - 50% of power = energy maximum minus 3 dB
    - 0.5 * power = $\sqrt{0.5}$ * amplitude = 0.707 * amplitude

Resonance response



Formant parameters:
- (center) frequency,
- (peak) amplitude,
- bandwidth

Speech waveforms and spectrograms

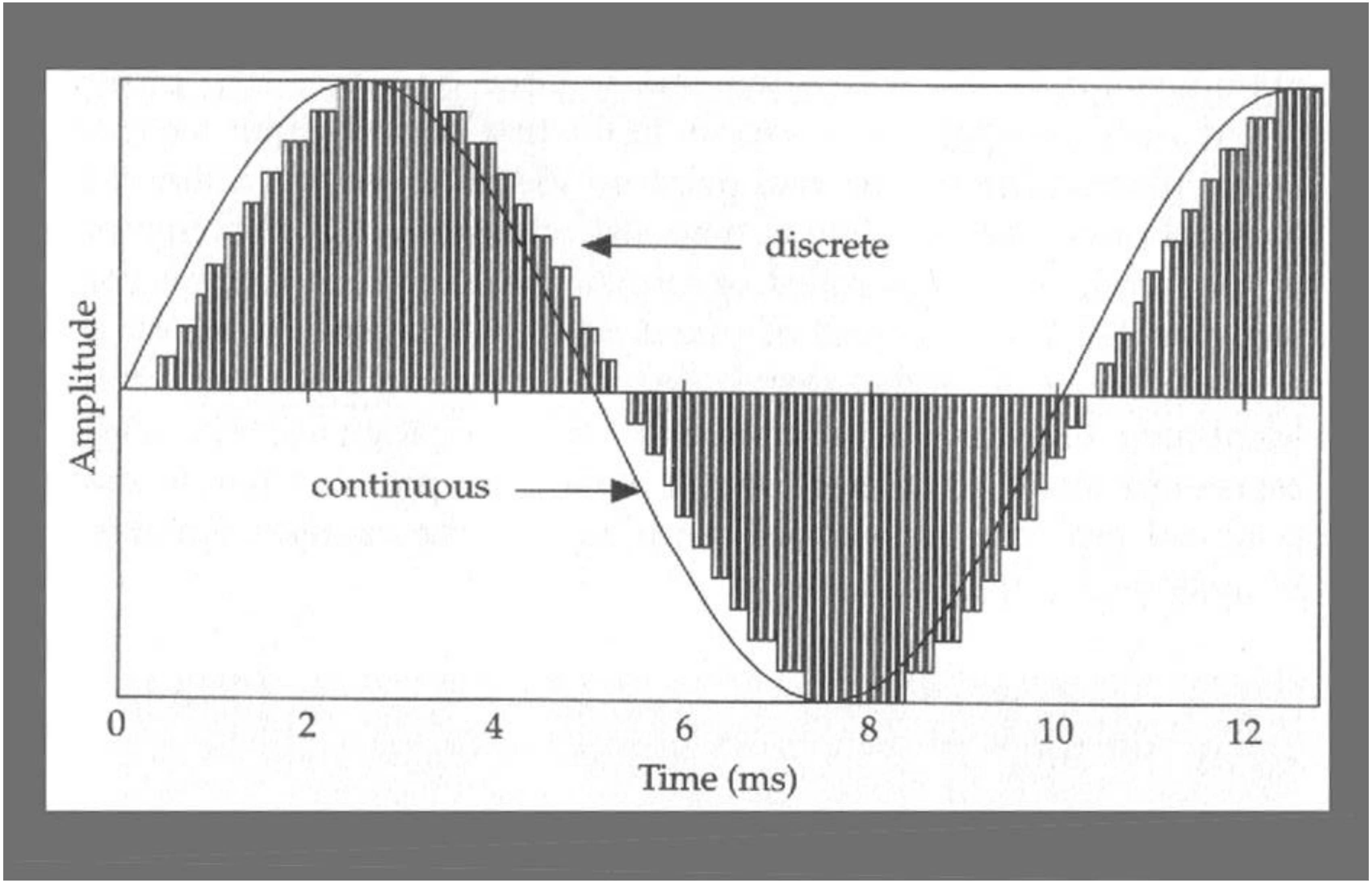


UNIVERSITÄT DES SAARLANDES

Continuous and discrete signals

- continuous (analog) signal
  - represented graphically as a continuous curve
  - amplitude values at all points in time
  - theoretically infinite number of time and amplitude values (arbitrary number of decimal places, e.g. "amplitude of 3.211178... volt at 1.034678 sec.")
- discrete (digital) signal:
  - represented graphically by individual, discrete bars
  - sequence of separate amplitude values
  - limited number of different time and amplitude values

Continuous and discrete signals

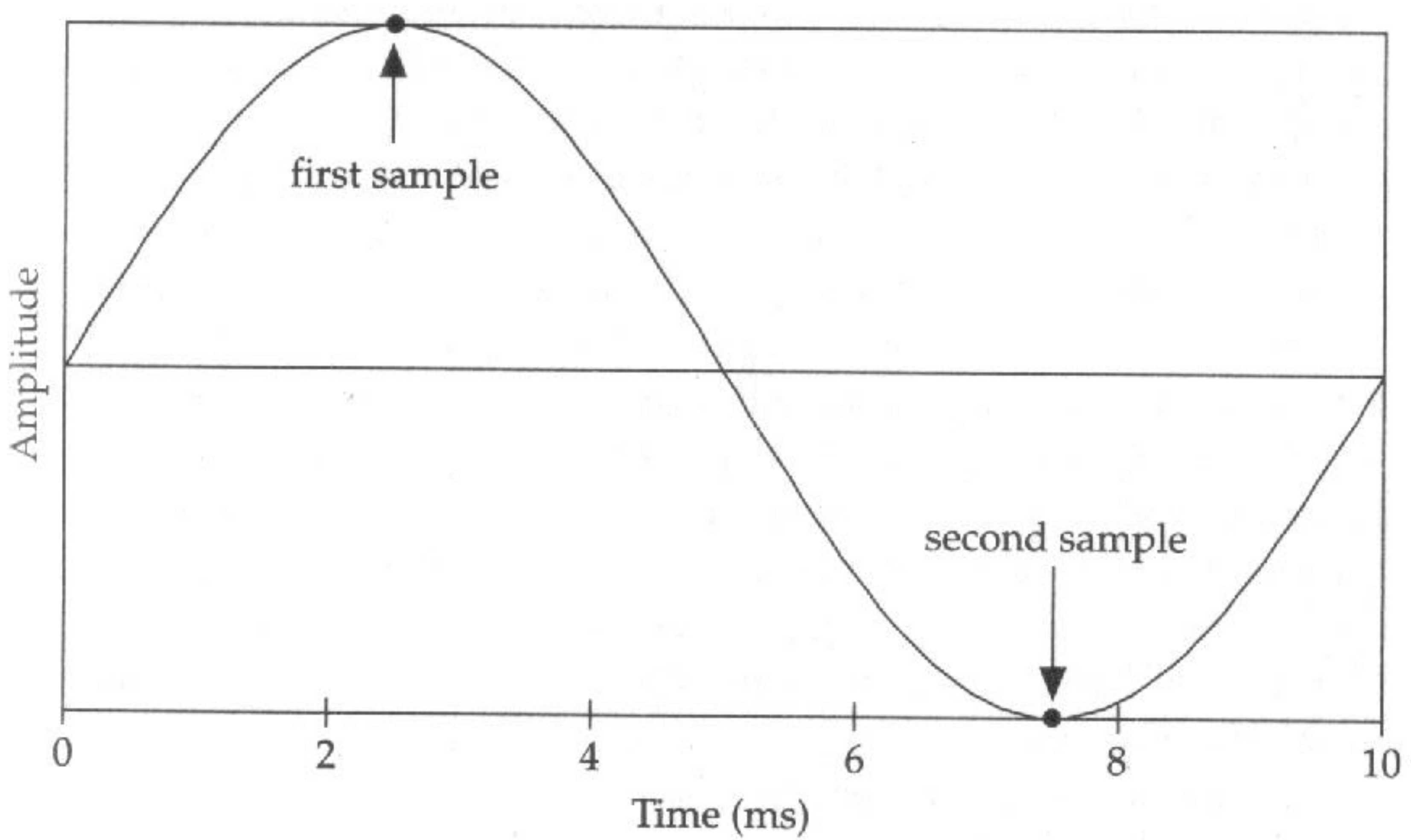


continuous vs. discrete sine wave [Johnson, 1997, p.23]

Analog-to-digital conversion

- A/D conversion – step 1: **sampling**
  - limitation of decimal places along time axis (x-axis)
  - slice-by-slice decomposition of continuous time signal
    - discrete points in time: **samples**
    - density of samples per time unit (sec.): **sampling rate** or **sampling frequency** [Hz]
- A/D conversion – step 2: **quantization**
  - limitation of decimal places along amplitude axis (y-axis)
  - slice-by-slice decomposition of continuous amplitudes
    - discrete amplitude values: **amplitude steps**
    - density of amplitude values: **quantization accuracy** [bit]

Sampling

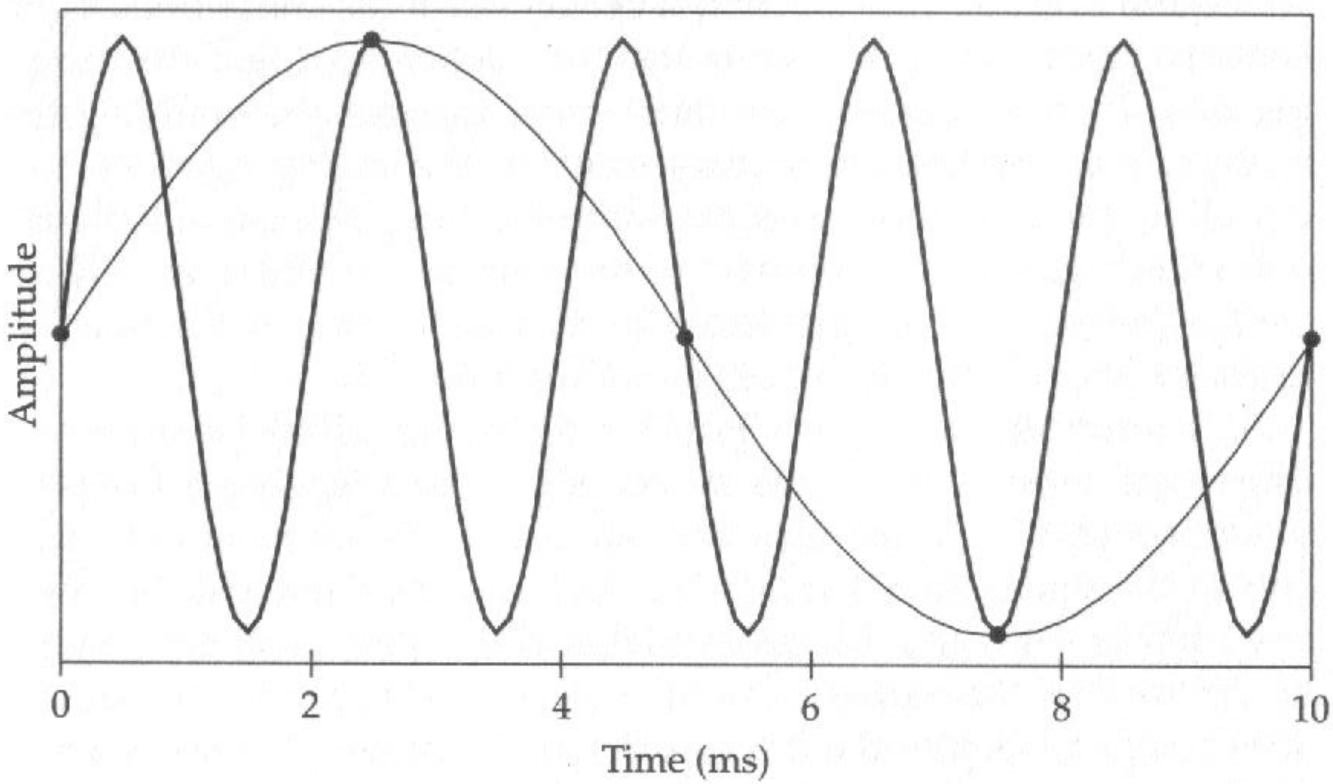


periodicity of sine wave can be represented by (minimally) 2 samples [Johnson, 1997, p.25]

Sampling theorem

- Which sampling rate is required for periodic (sine) waves?
  - 2 samples per fundamental period
    - sampling frequency $\geq$ 2 * fundamental frequency
    - e.g.: 100 Hz sine wave $\rightarrow$ 200 Hz sampling rate
    - known as **sampling theorem**
- What does this mean for complex (e.g. speech) signals?
  - useful information in speech signal of up to approx. 8 kHz
    - requires 16 kHz sampling rate
    - (cf. audio CDROM: 44.1 kHz)
  - **Nyquist frequency**: 0.5 * sampling frequency
    - highest frequency component of sampled signal

Aliasing effect by undersampling

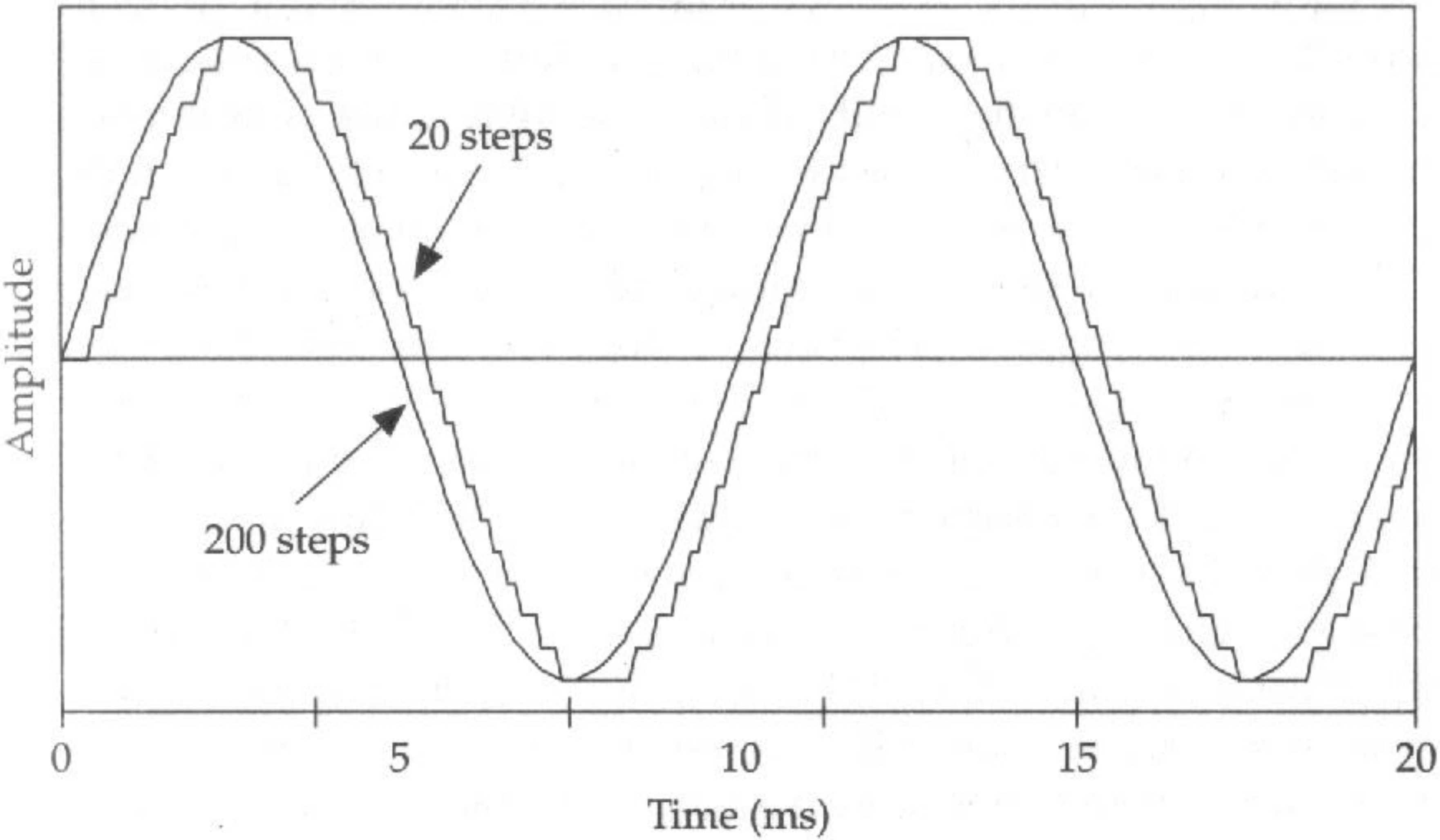


Undersampling of a sine wave [Johnson, 1997, p.27]

- digital signal has a low-frequency component and fails to represent correctly the high-frequency analog signal; audio demo:
  - in practice: use low-pass filter to remove all frequencies above Nyquist frequency (frequency band limitation)

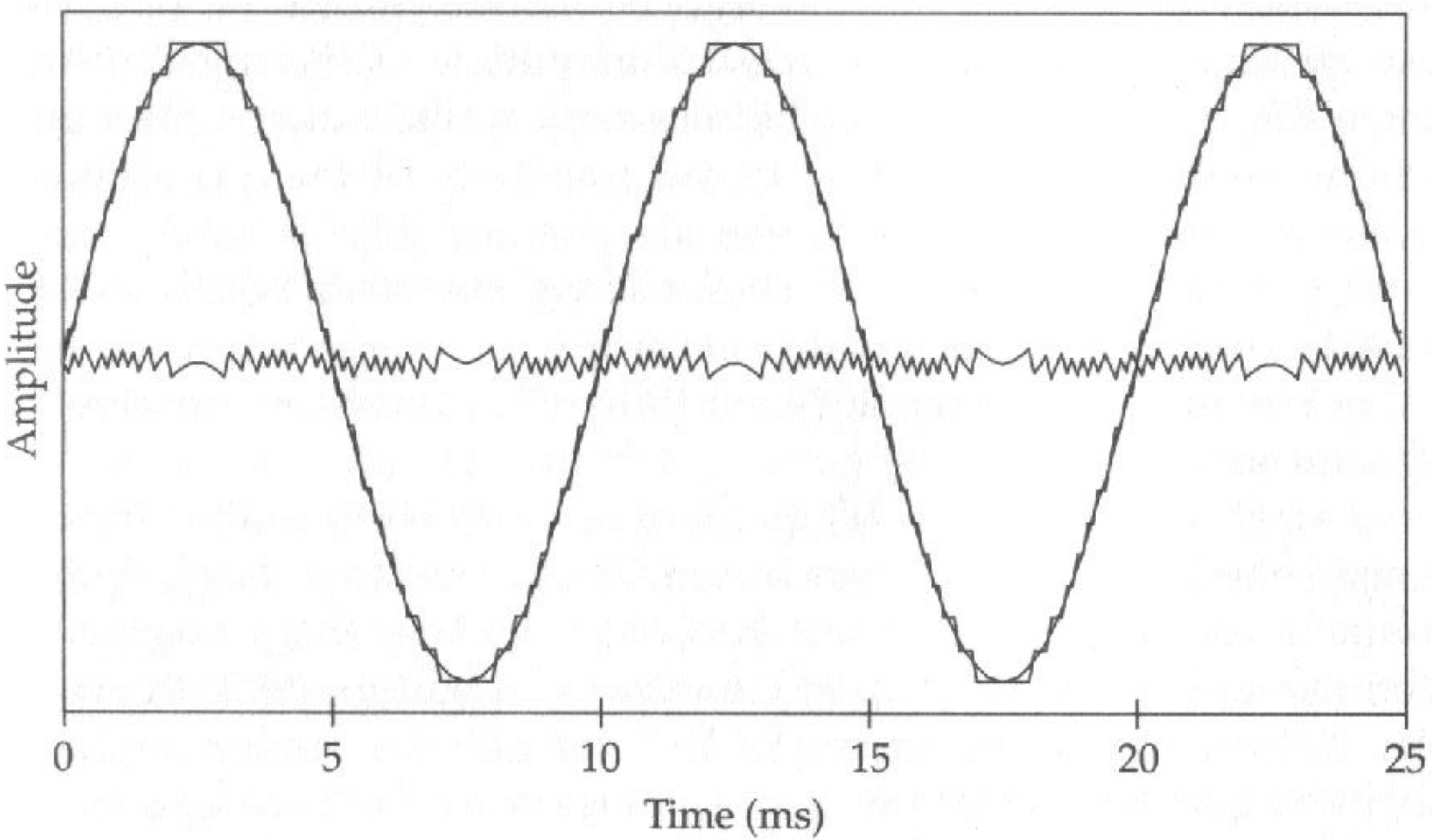
UNIVERSITÄT DES SAARLANDES

Quantization



- 2 different quantizations (20 vs. 200 steps) of sine wave amplitude [Johnson, 1997, p.29]

Quantization noise



- difference between continuous and quantized signal [Johnson, 1997, p.31]
- audio demo

UNIVERSITÄT DES SAARLANDES

Quantization steps and accuracy

- How accurately should waveform amplitudes be quantized? Or: How many quantization steps should be used?
  - digital representation by binary digits (bits): 0/1
    - 2 bit = 2^2 = 4 steps, or 3 bit = 2^3 = 8 steps, etc.
    - in practice: 16 bit = 2^{16} = 65536 steps, values: -32768 - 32768
  - quantization noise is negligible when using 16 bit
- How to report digitization:
  - "The speech signal was quantized with 16 bit accuracy (or bit depth) and a sampling rate (or sampling frequency) of 16 kHz."

Thanks!