Speech Signal Processing

Basics: Physikalische Größen

Ein Signal hat:

A · Amplitude — max. Ausschlag einer harmonischen Schwingung vom arithmetischen Mittel

f · Frequenz [Hz] · Anzahl der Schwingungen pro Sekunde

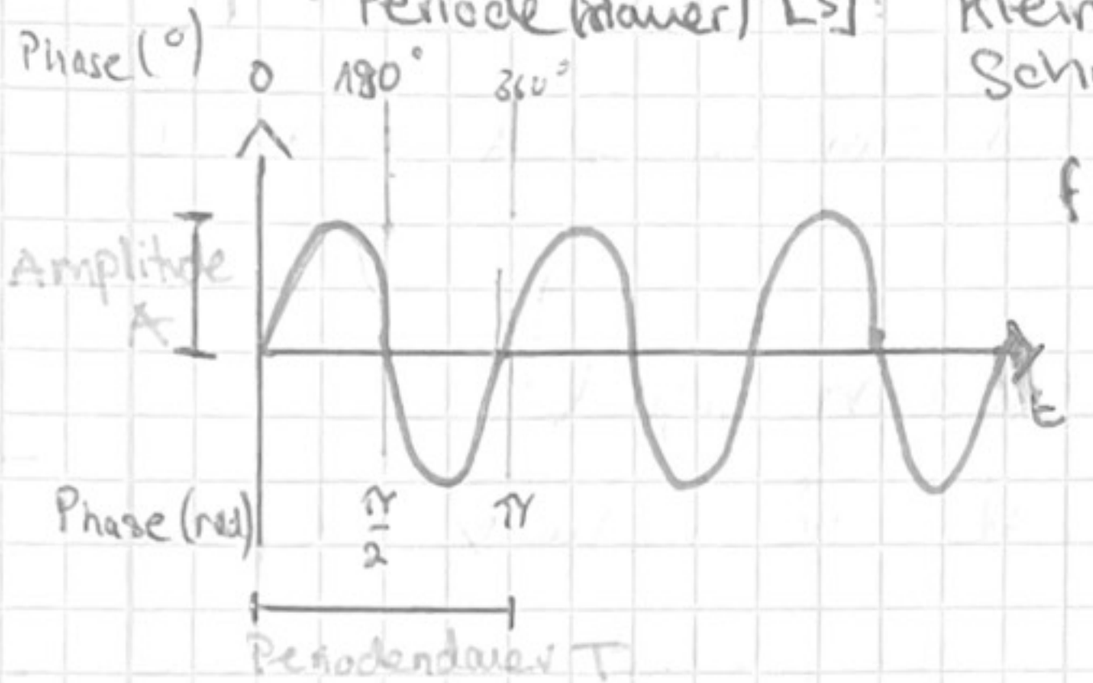
t · Time-Domain [s] · Zeit, i.d.R. auf der y-Achse

Phi ϕ · Phase — Position des Signals innerhalb den Zyklus; Angabe in Grad (°) oder Radian (rad)

Omega ω · Kreisfrequenz — Phasenwinkel pro Zeit, abh. von Frequenz
$\omega = 2\pi f = \frac{2\pi}{T}$

T · Periode (dauer) [s] — kleinstes Intervall nach dem sich eine Schwingung wiederholt.

$f(x) = \sin(x)$



Q · Pegel / (Laut-) Stärke [dB]: Logarithmus der Druckunterschieden der Schallwellen und dem Normalwert, der "Schalldruck".

- Warum logarithmisch? → der von Menschen wahrnehmbare Lautstärke-Bereich ist sehr groß und nicht gut auf einer linearen Skala darstellbar.
- Was bedeutet das?
- Ein Unterschied von ± 10 dB entspricht etwa einer Verdopplung / Halbierung der Lautstärke

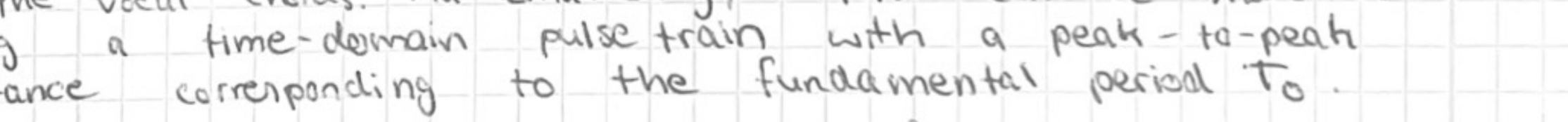
·	~ 30 dB	empfinden wir als ruhig (Schlafzimmer bei Nacht)
:		
·	~ 60 dB	normales Sprechen
·	~ 70 dB	Staubsauger
·	~ 80 dB	Verkehrsstraße
·	~ 100 dB	Disco
·	~ 130 dB	Schmerzensgrenze
·	~ 140 dB	Düsenflugzeug

Lecture: Speech Production

- How do we produce speech?
  - Lungs produce airflow
  - In the larynx (Kehlkopf) the vocal cords start vibrating and produce ^voiced sound.
  - In the vocal tract the sound is formed to produce a speech sound
- The most important speech sounds are
  - Voiced sounds: vowels (a,e,i,o,u), sounds w/ mixed excitation (/v/)
  - unvoiced sounds
    - Fricative (/s/, /th/, /sh/)
    - plosive (/k/, /p/, /t/)

Lecture: Source-Filter-Model

The production of a speech signal can be ~~used~~ described using a source-filter-model:

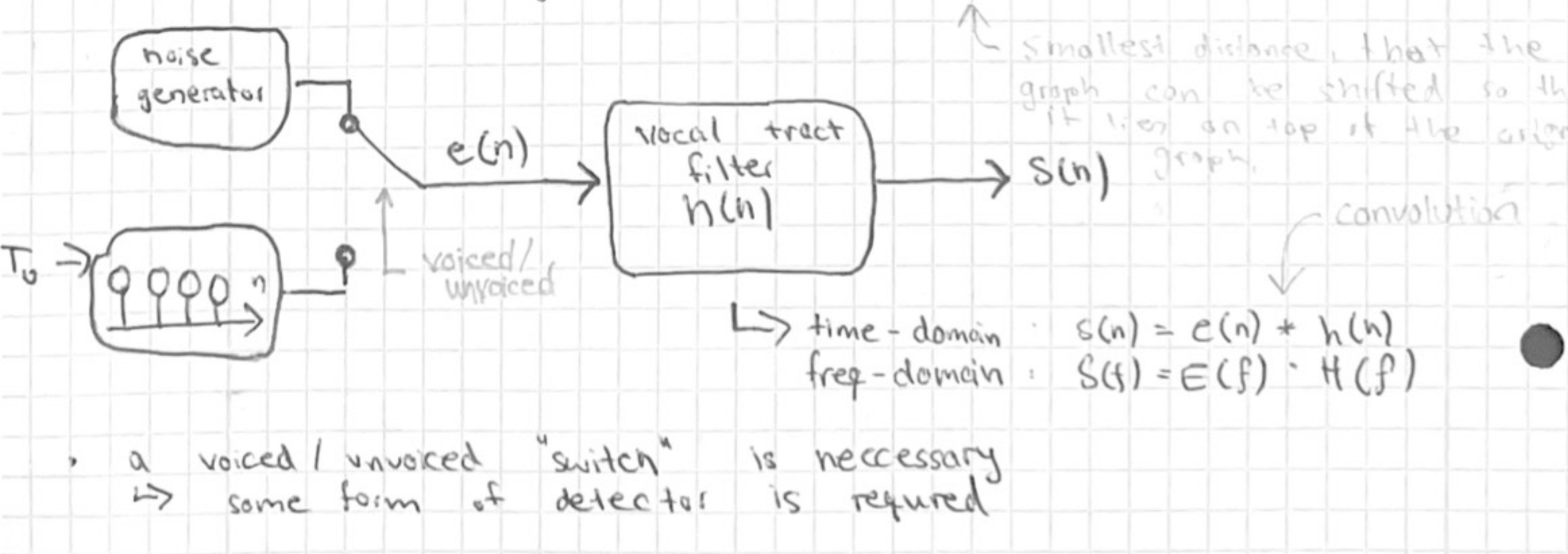


[Simplyfying assumption: source and filter are mutually independant]

→ in humans:
- Source: airflow, vibration of vocal cords
- ~~Filter~~: vocal tract, tongue, lips, palate etc.

— Excitation = Anregung —— (source part of the model)

- unvoiced excitation is noise-like, lacking in periodicity
  ↳ can be modelled using Gaussian white noise
- voiced excitation is produced through periodic opening and closing of the vocal chords. Mathematically, this can be modelled using a time-domain pulse train with a peak-to-peak distance corresponding to the fundamental period T_0.
  ↑ Smallest distance, that the graph can be shifted so that it lies on top of the original graph.



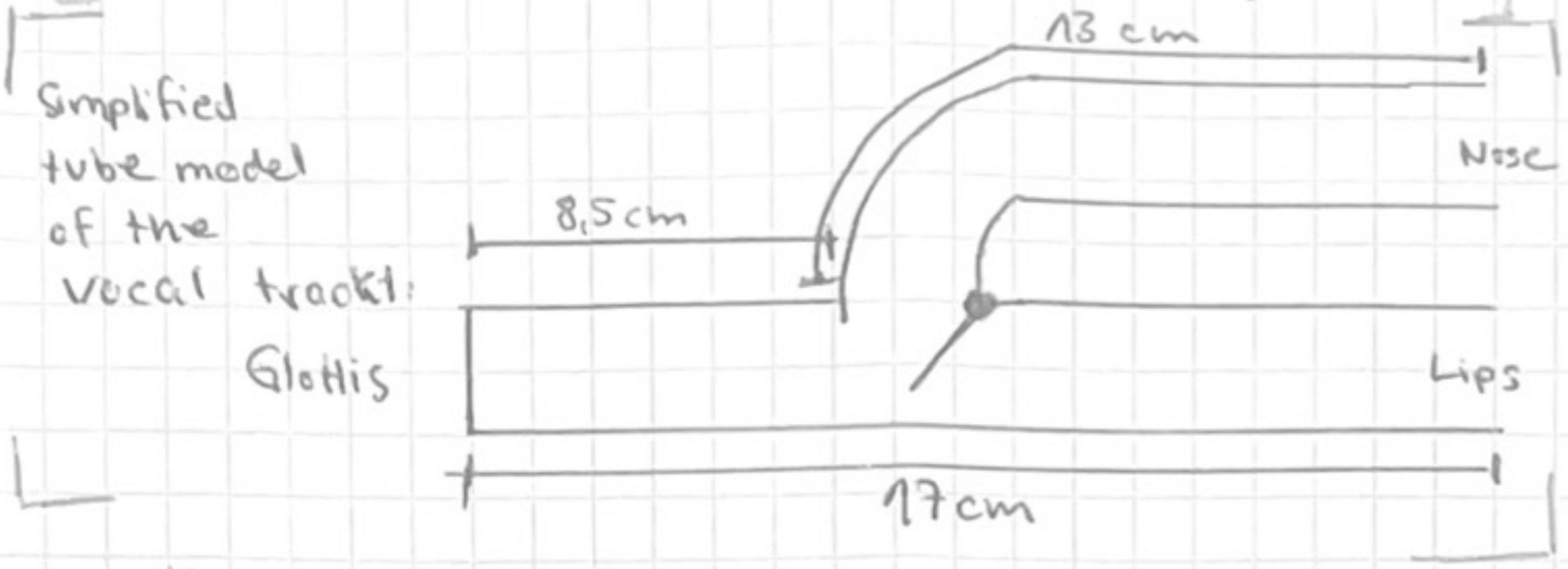
↳ time-domain: $s(n) = e(n) * h(n)$
freq-domain: $S(f) = E(f) \cdot H(f)$

- a voiced / unvoiced "switch" is neccessary
  ↳ some form of detector is requred

— vocal tract (filter model) —

we can model the vocal tract as a filter through which the excitation signal passes.

[this model is an immense simplification



- the length of the tuben correspond to certain frequency ranges called formants
- these formants (resonance frequencies) can be mathematically described using a vocal tract transfer function

→ the filtering is represented by a convolution of the excitation signal and the vocal tract impulse response in the time domain

$$s(n) = e(n) * h(n)$$

or a multiplication of the excitation spectrum and the vocal tract transfer function in the frequency domain

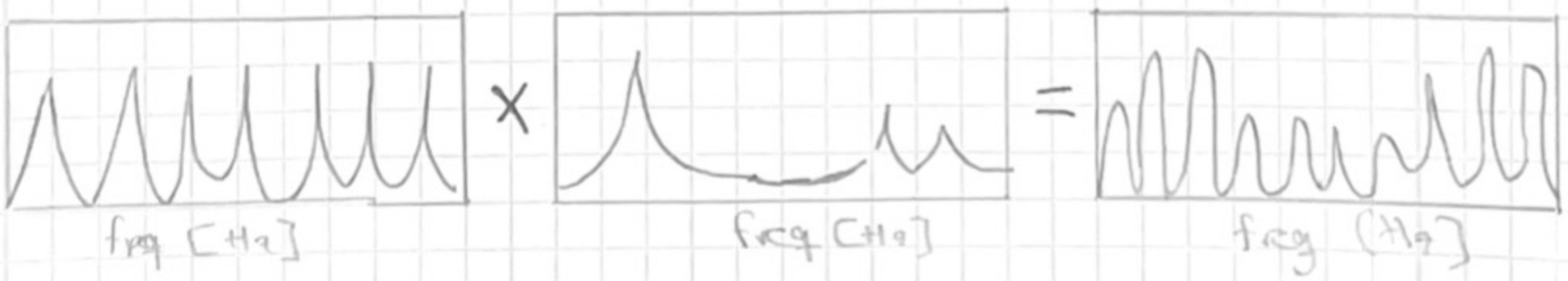
$$S(f) = E(f) \cdot H(f)$$

- the spectrum of the excitation signal consists of the fundamental frequency f_0 and a set of harmonics at integer multiples of f_0.
- formants appear as spectral peaks in the transfer function

[→ f_0 and it's harmonics are the only fequencies in the final signal, because the ~~transfer function~~ excitation signal samples the transfer function at discrete points.

→ fundamental frequency f_0: intonation and prosody (no meaning), excitation signal

└ formant frequencies: added thrugh the trasfer function, allowing the harmonics of the excitation signal to resonate within ther formant freq. regions]



excitation $e(n)$ × vocal tract transfer function $h(n)$ = spectrum $s(n)$