Sprache und Spracherkennung

Signale, Systeme und Sensoren: Vorlesung 11

Prof. Dr. M. O. Franz

HTWG Konstanz, Fakultät für Informatik

Übersicht

Übersicht

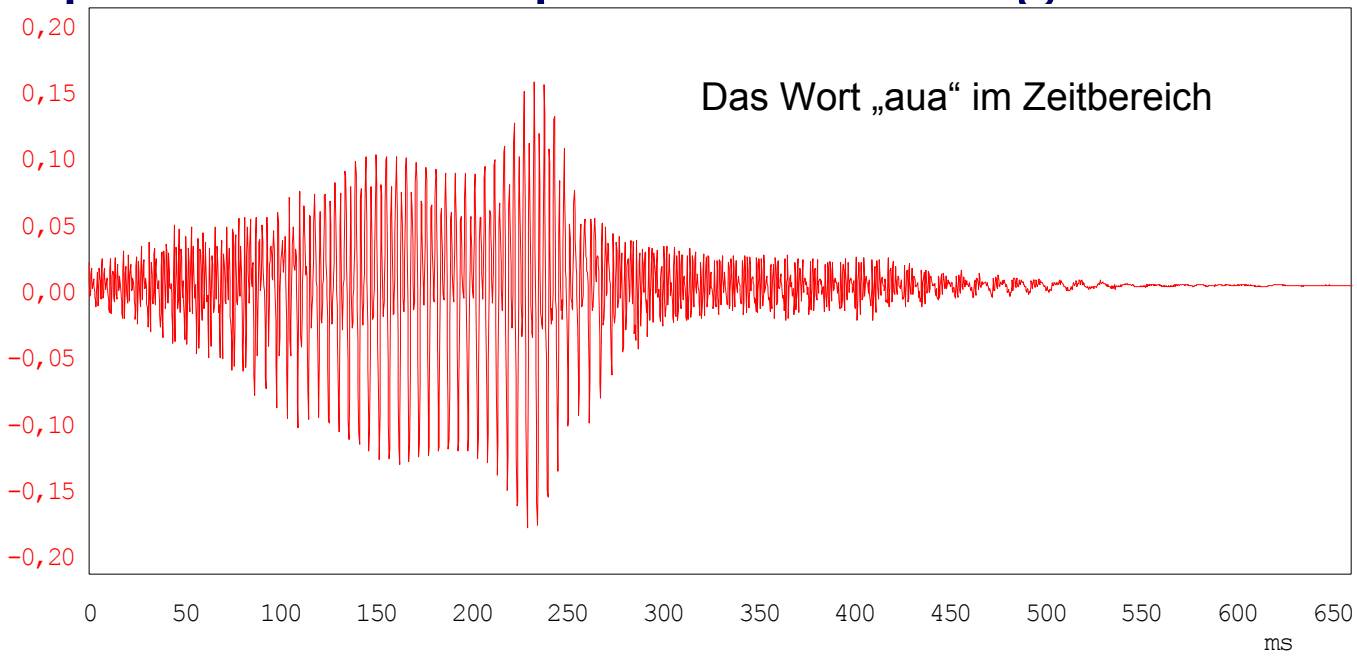
Wiederholung: Zeit-Frequenz-Komplementarität

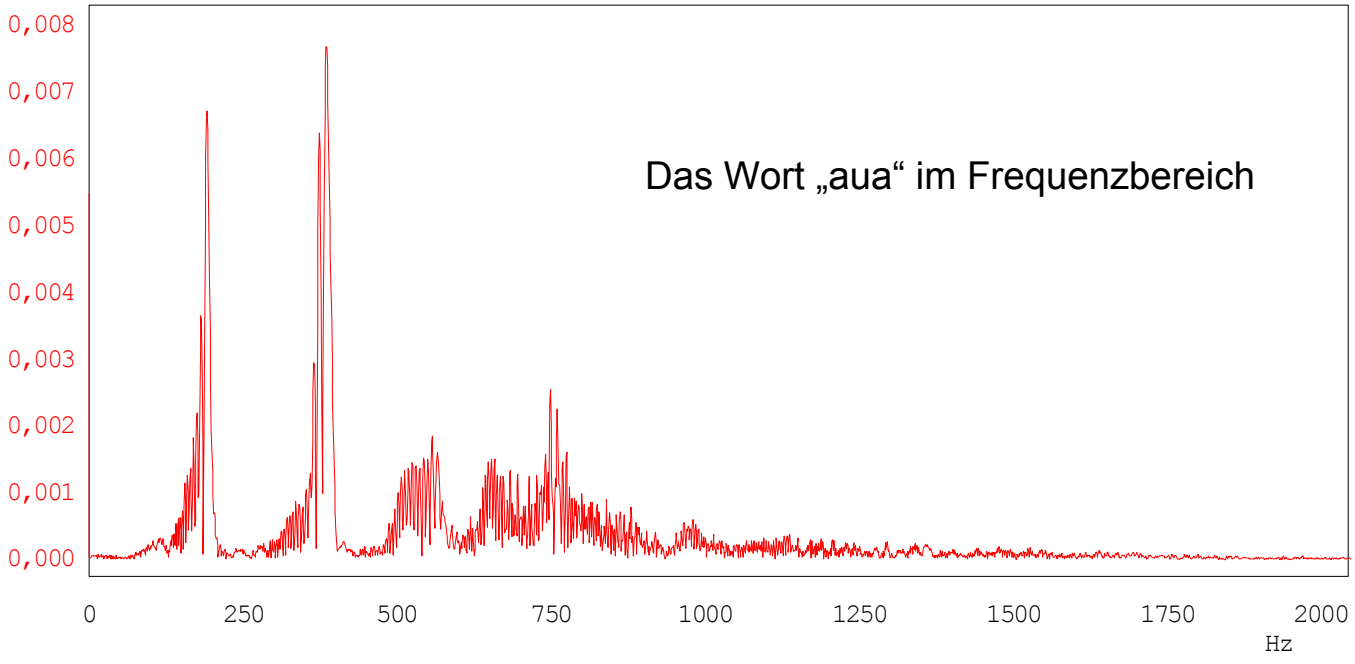
- Eine zeitliche Eingrenzung der Signaldauer Δt bedeutet eine Ausweitung des Frequenzbandes Δf. Umgekehrt gilt: Je eingeschränkter das Frequenzband eines Signals ist, desto größer muss zwangsläufig die Zeitdauer des Signals sein.
- Frequenzband und Zeitdauer eines Signals können nicht unabhängig voneinander gemessen werden. Eine genaue Messung des Frequenzbandes erfordert ein lange Zeitdauer, eine genaue Messung der Zeitdauer ein breites Spektrum.
- Es gilt die **Frequenz-Zeit-Unschärferelation**:

$$\sigma_t \cdot \sigma_\omega \geq 1 \quad \text{bzw.} \quad \Delta t \cdot \Delta f \geq 0.88.$$

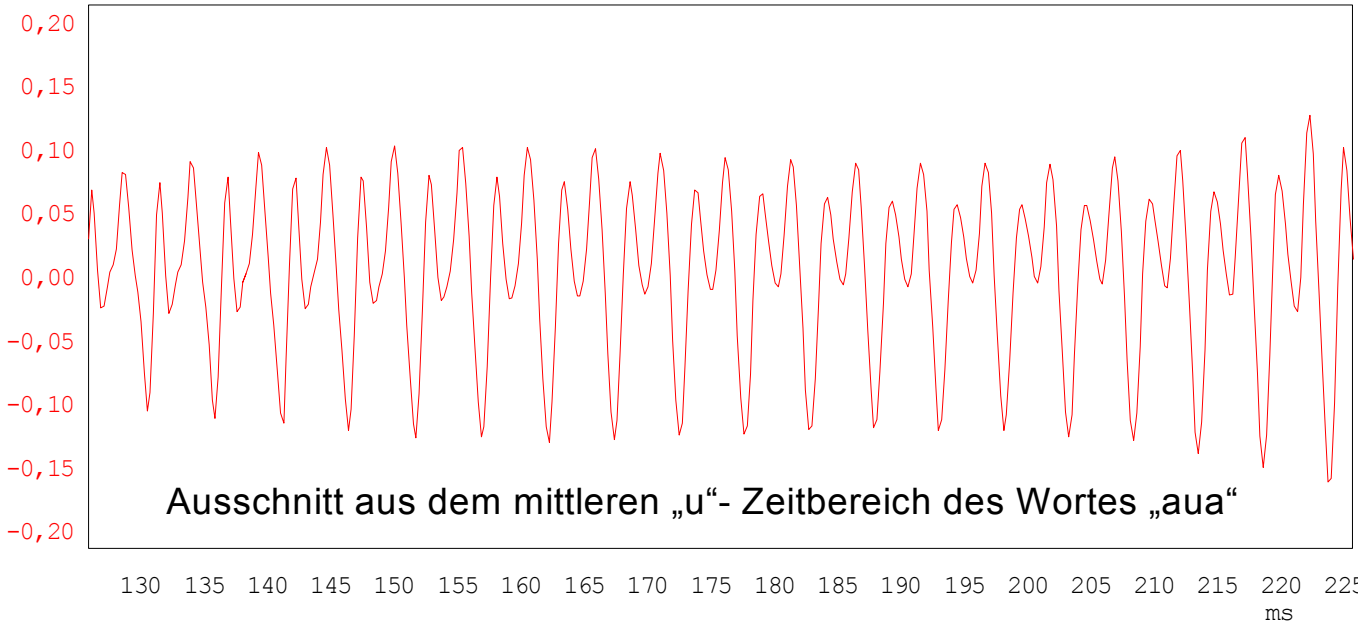
- **Fastperiodische Signale** wiederholen sich nur über einen begrenzten Zeitraum. Sie besitzen linienähnliche Spektren, die ausschließlich die ganzzahlig Vielfachen der Grundfrequenz umfassen.

Beispiel: Sprache als fastperiodisches Signal





Beispiel: fastperiodische Signalabschnitte

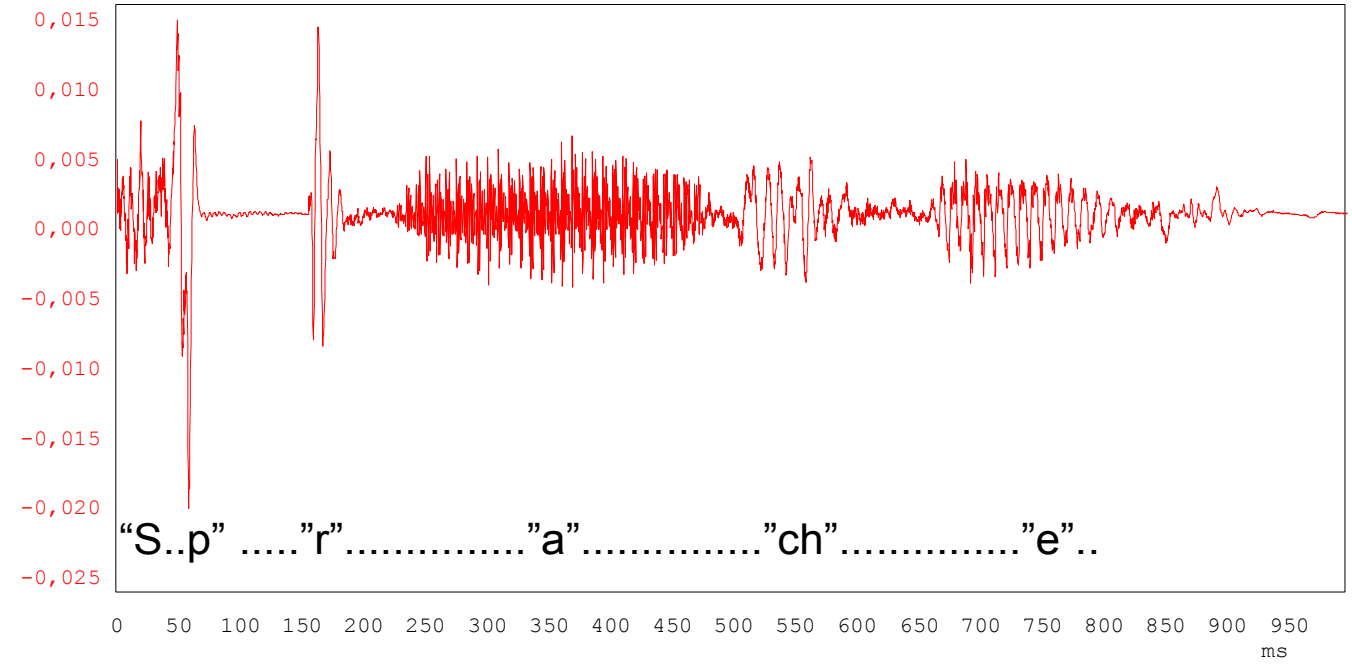


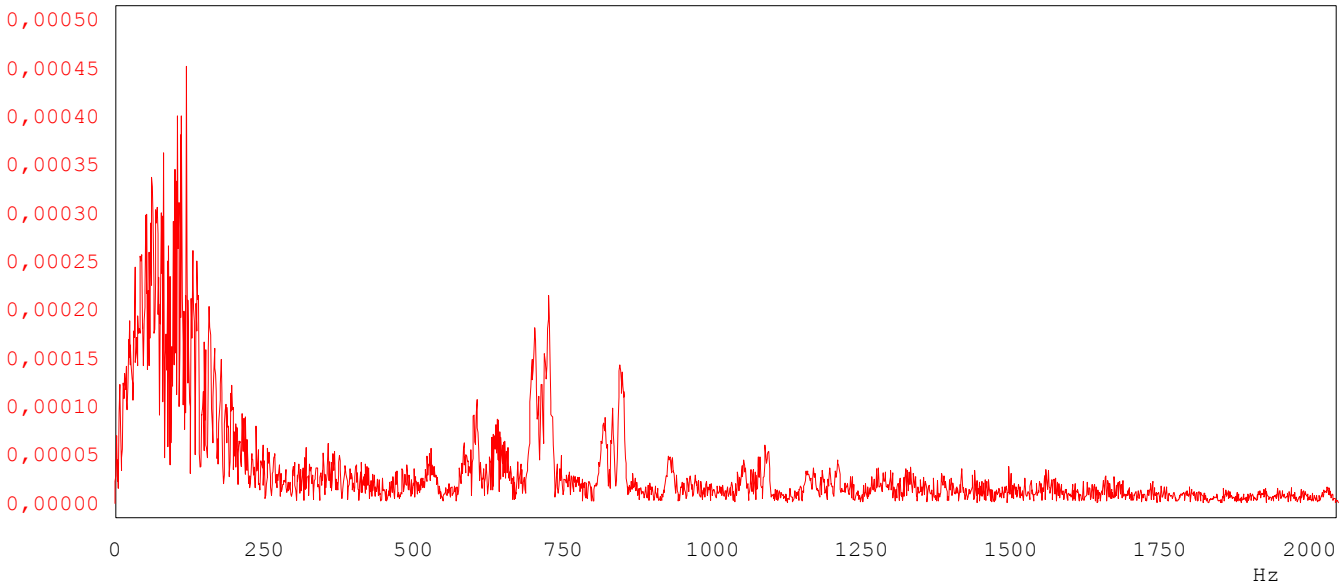
Quelle: Karrenberg, 2012

Vokale

- Vokale sind fastperiodische Signalabschnitte in Sprachsignalen.
- Je länger der Vokal gesprochen wird, desto klarer kann er wahrgenommen werden, desto präziser kann unser Gehör die Tonhöhe der Vokalfrequenzen bestimmen. Desto kürzer er ausfällt, desto unverständlicher wird er (s. Zeit-Frequenz-Komplementarität). Die Vokalerkennung entspricht also einer Frequenzmessung.
- Betrachtet man das Spektrum über das ganze Signal, so werden die Spektren der einzelnen fastperiodischen Abschnitte einfach gemischt und können so in ihrer zeitlichen Abfolge nicht mehr unterschieden werden. Um eine Folge von Vokalen erkennen zu können, benötigt man eine lokale Form der Fourieranalyse innerhalb eines gleitenden Zeitfensters.
- Eine ähnliche Analyse wäre auch sinnvoll für Konsonanten, obwohl diese nicht fastperiodisch sind, sondern aus Rausch-, Explosiv- oder Reibelauten entstehen.

Beispiel: “Sprache”





Quelle: Karrenberg, 2012

Erkennung von Sprachlauten

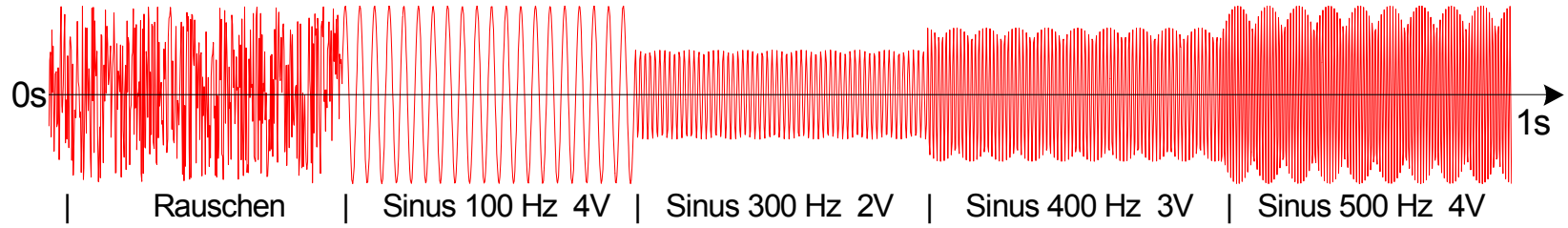
- **Phon** (auch: Laut, Sprachlaut): die kleinste unterscheidbare „Lauteinheit im Lautkontinuum“ - ein minimales Schallsegment, das noch als selbständig wahrgenommen wird.
- **Phoneme**: die Menge aller Phone, die in einer gesprochenen Sprache die gleiche bedeutungsunterscheidende Funktion haben (z.B. gerolltes “r” und Rachen-“r”).
- Jedes Phon besitzt einige charakteristische Frequenzen, die - ähnlich wie ein Fingerabdruck - nahezu unverwechselbare, denkbar einfache Muster sind.
- Akustische Mustererkennung geschieht in der Natur wie in der Technik überwiegend im Frequenzbereich.
- Die Frequenzmuster von fastperiodischen und quasiperiodischen Signalen - z.B. Vokalen - sind besonders einfach, da sie lediglich aus mehreren „verschmierten“ Linien (“peaks“) verschiedener Höhe bestehen.

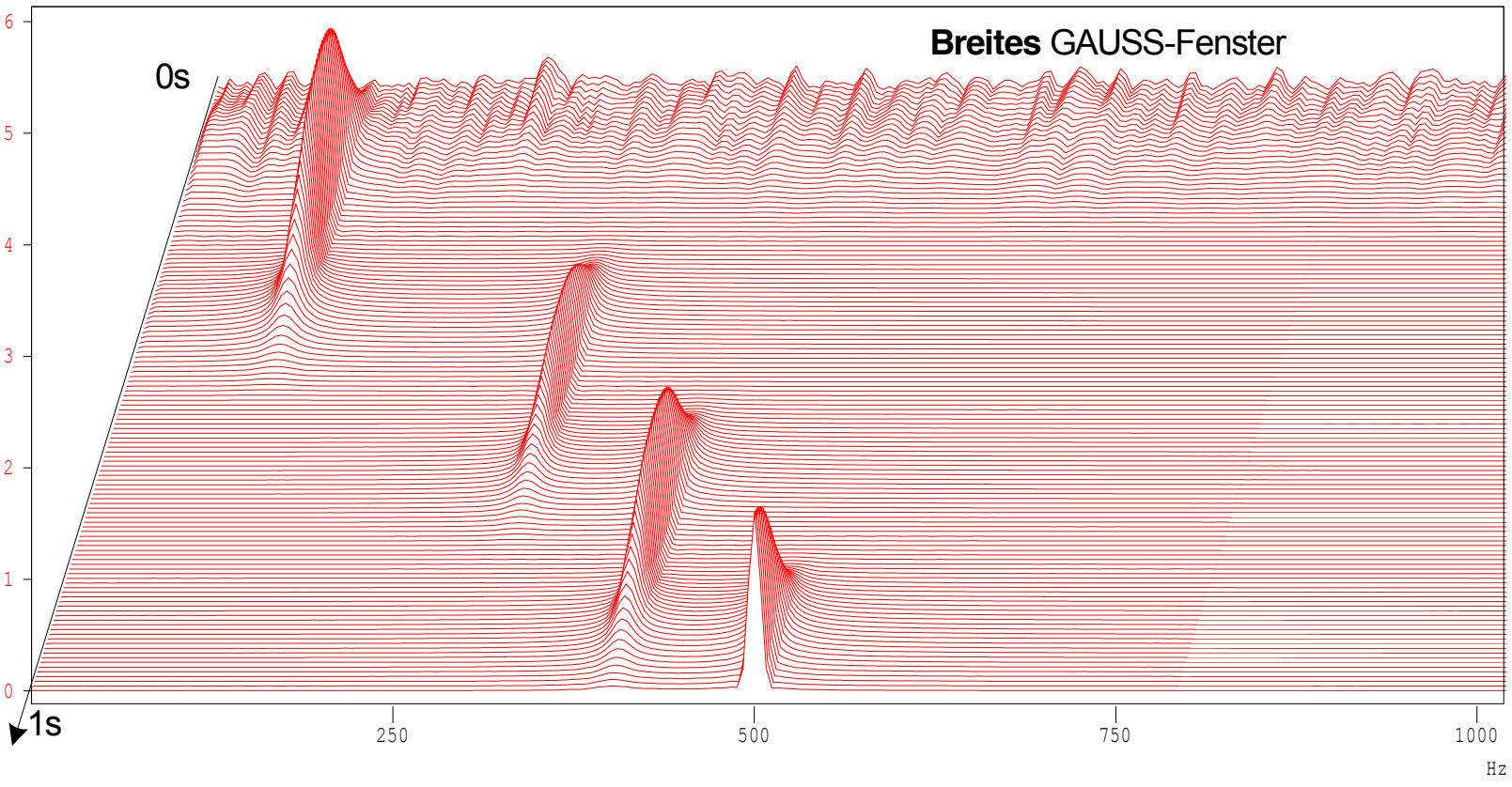
Kurzzeit-Fouriertransformation

Kurzzeit-Fourieranalyse (s. Vorl. 7):

- Signal wird in eine Folge überlappender Fenster zerlegt. Natürlich müssen die Fenster dicht genug aneinander liegen, um alle zeitlichen Veränderungen des Spektrums mitzubekommen.
- Jedes Fenster wird mit einer geeigneten Fensterfunktion multipliziert.
- Fourieranalyse innerhalb des Fensters.
- Wird ein kurzes Zeitfenster gewählt, lässt sich relativ genau zeitlich lokalisieren, wann ein relativ breites Band benachbarter Frequenzen wahrnehmbar war.
- Wird ein längeres Zeitfenster gewählt, lässt sich relativ ungenau zeitlich lokalisieren, wann ein relativ schmales Band benachbarter Frequenzen wahrnehmbar war.

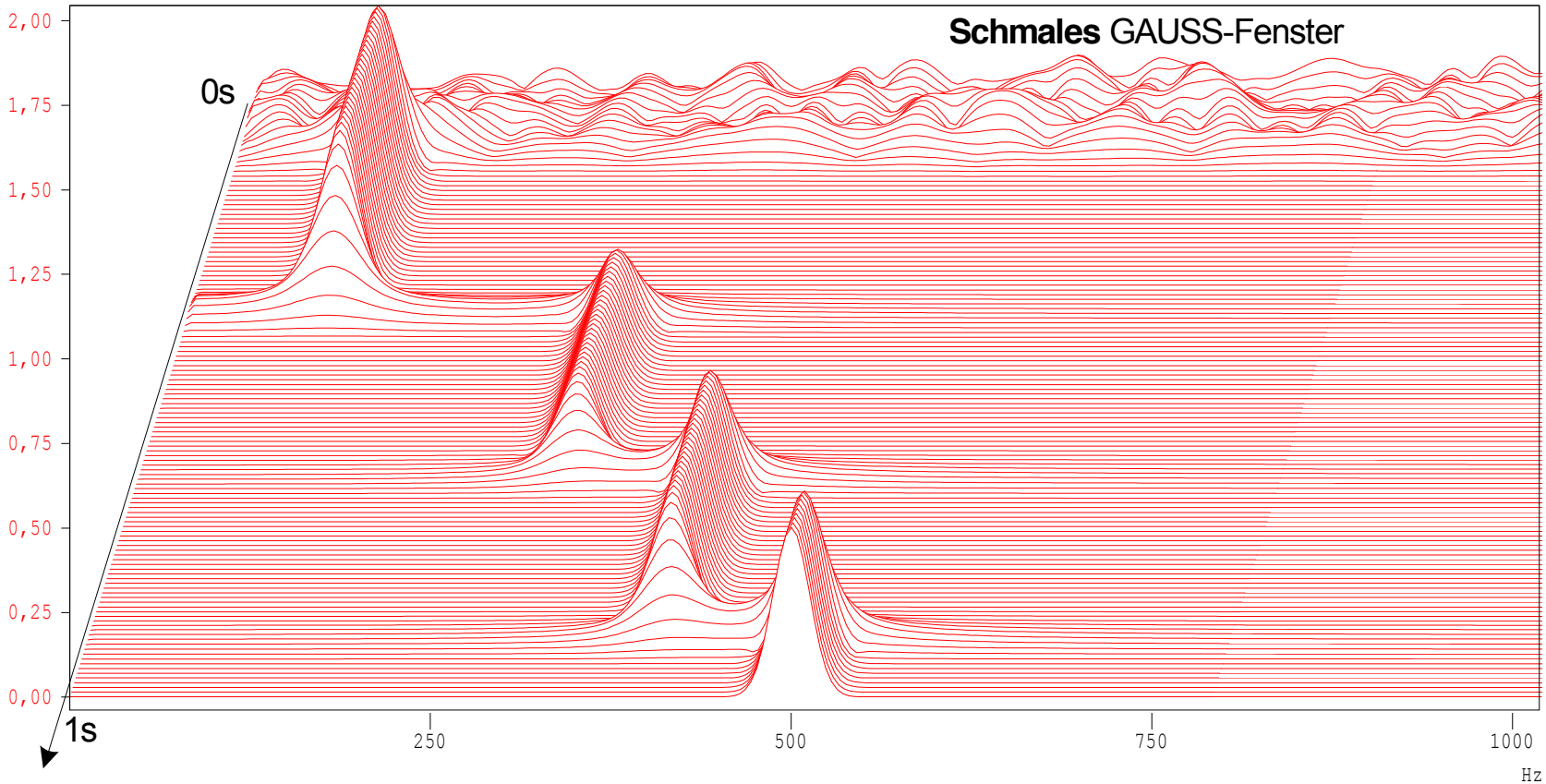
Zeit-Frequenz-Landschaft mit breitem Gauß-Fenster





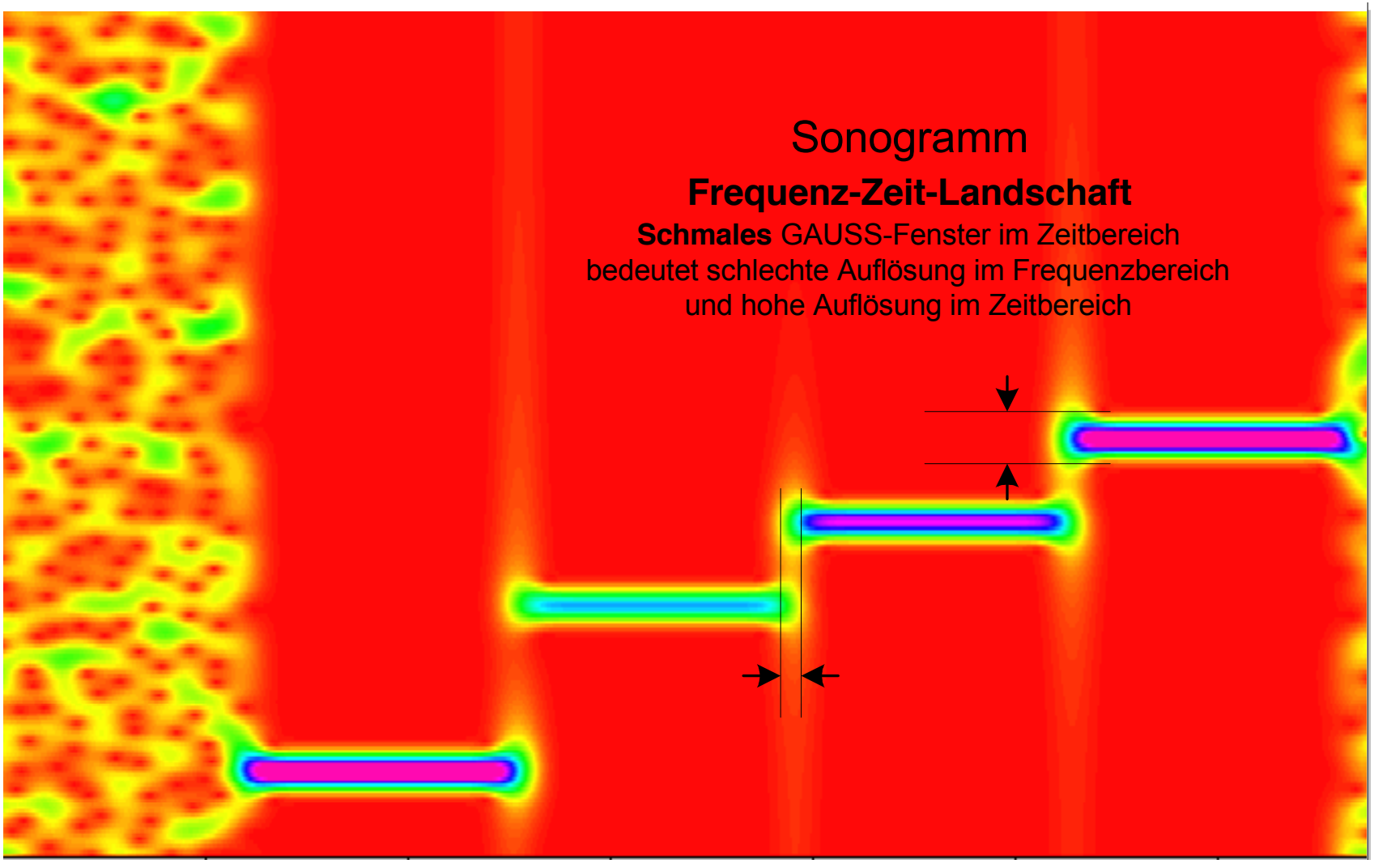
Quelle: Karrenberg, 2012

Zeit-Frequenz-Landschaft mit schmalem Gauß-Fenster



Quelle: Karrenberg, 2012

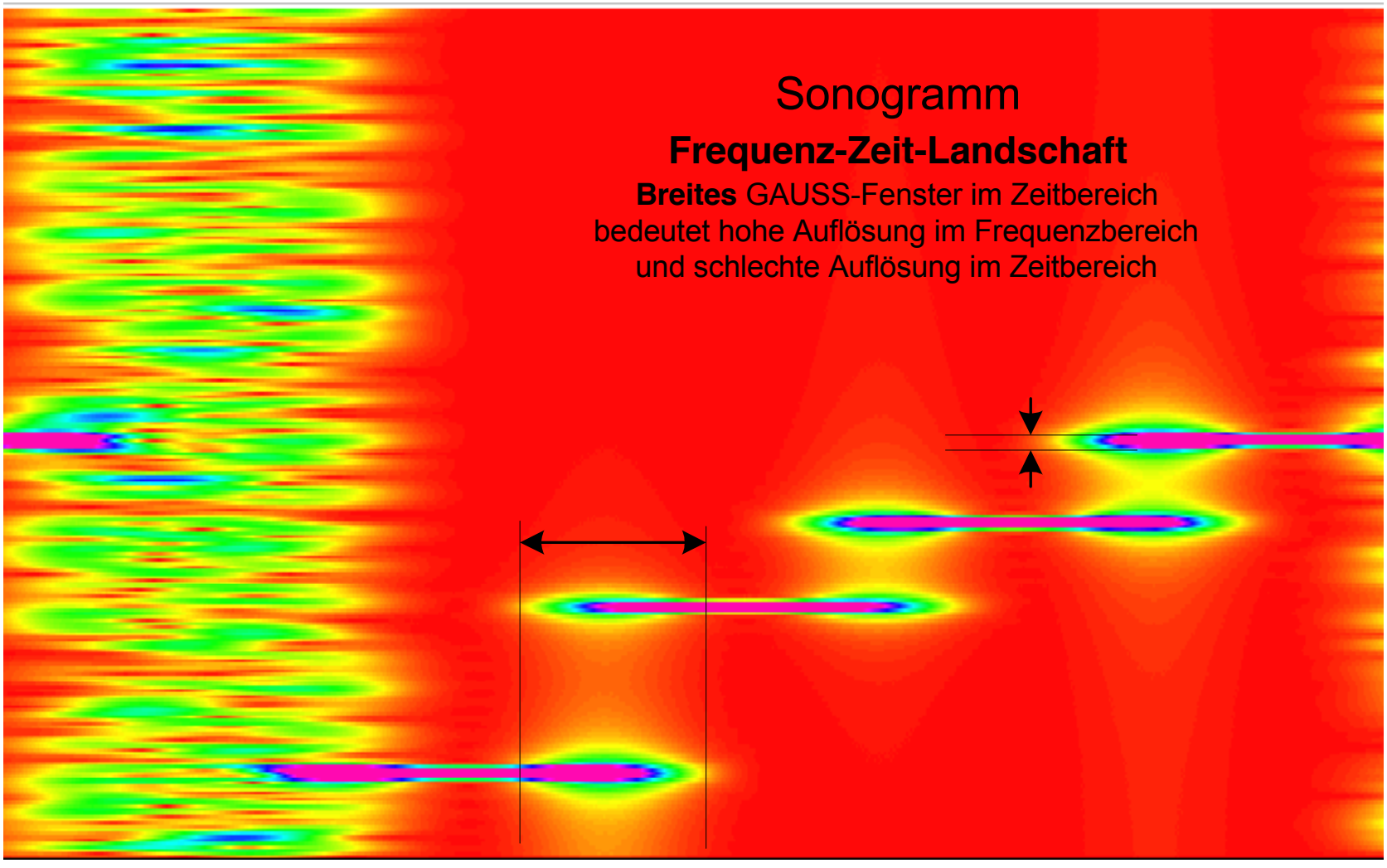
Sonogramm mit schmalem Gauß-Fenster



Sonogramm oder **Spektrogramm**: x-Achse: Zeit, y-Achse: Frequenz

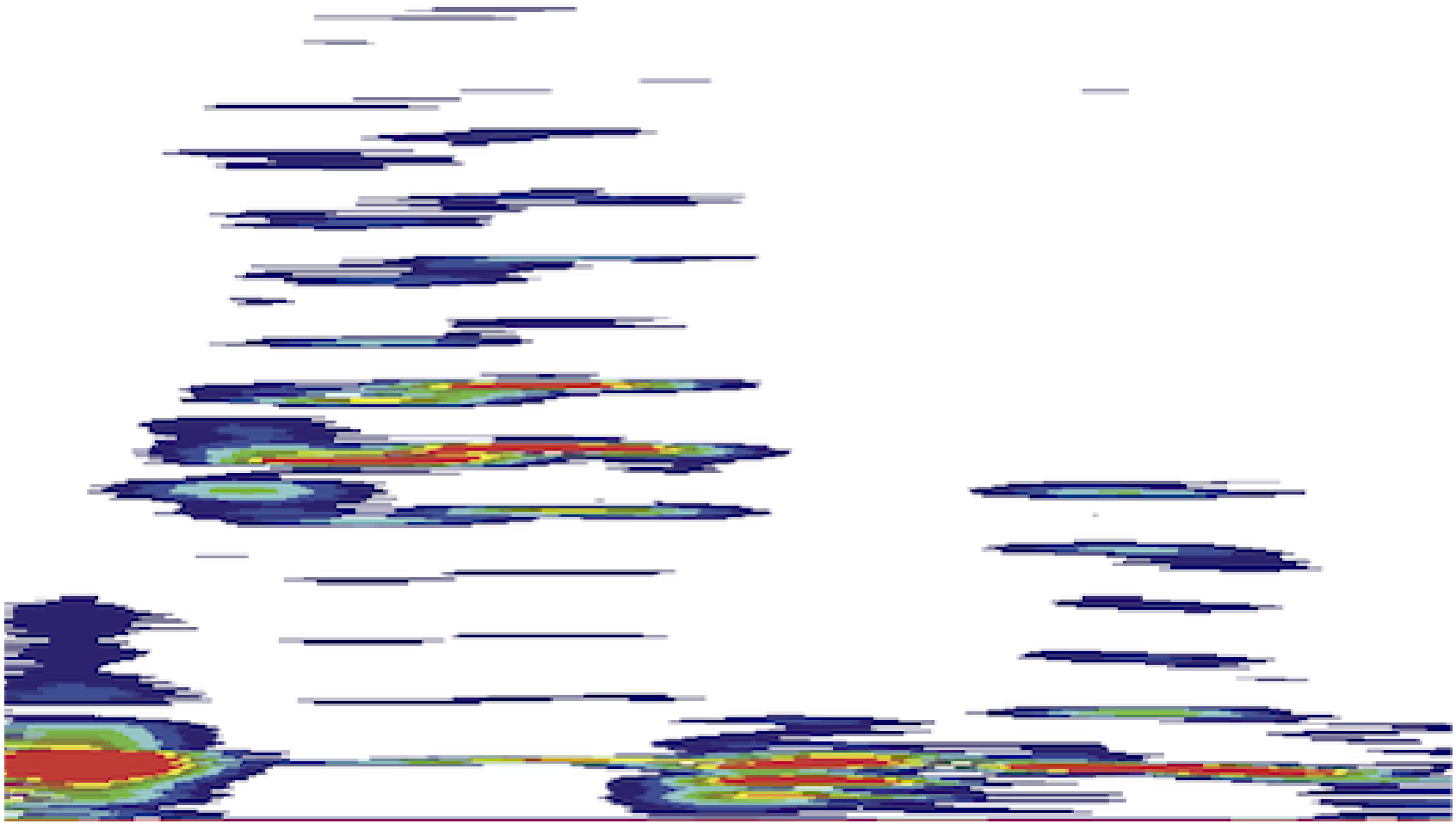
Quelle: Karrenberg, 2012

Sonogramm mit breitem Gauß-Fenster



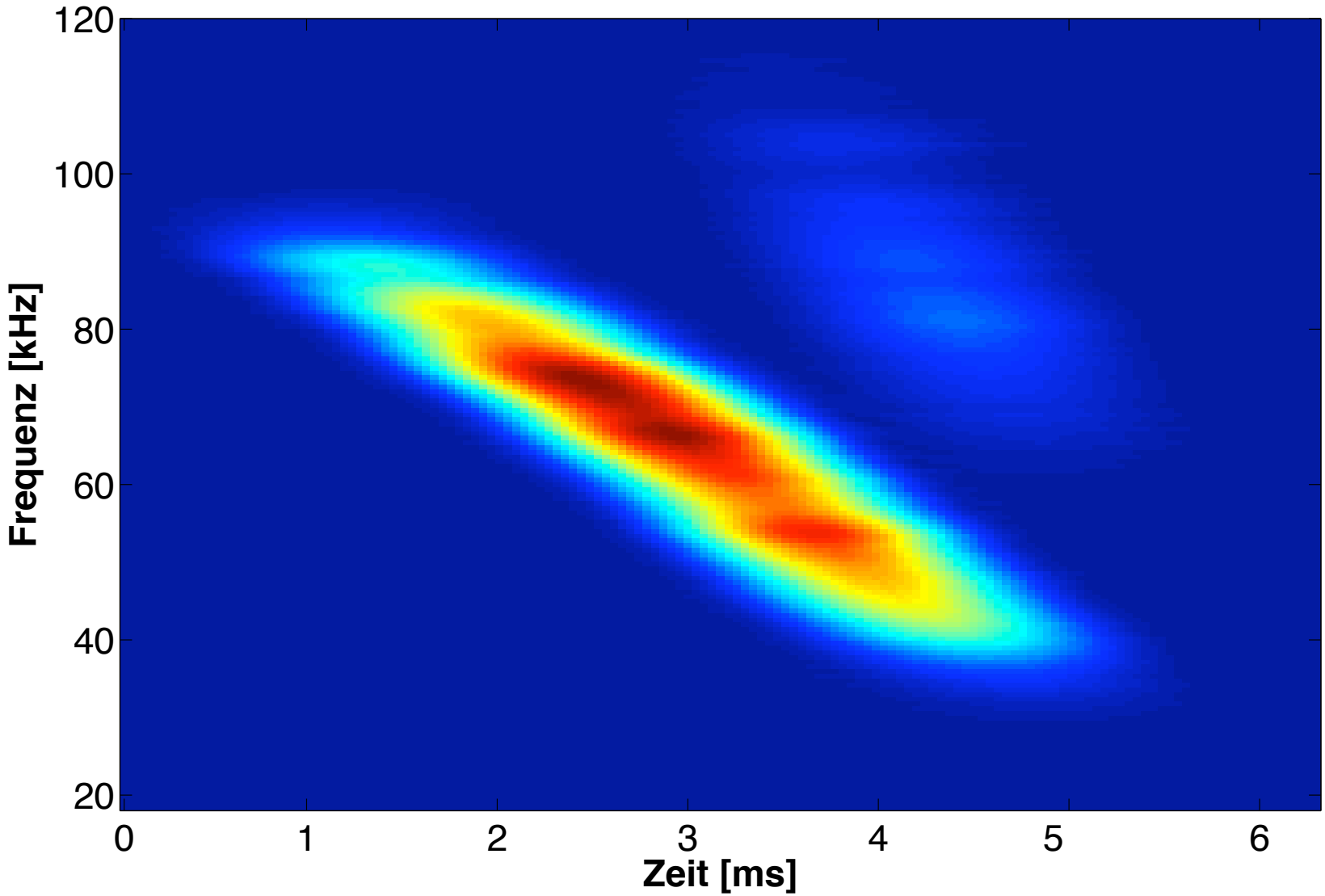
Quelle: Karrenberg, 2012

Sonogramm für “Sprache”



Quelle: Karrenberg, 2012

Sonogramm des Echoortungslautes einer Fledermaus



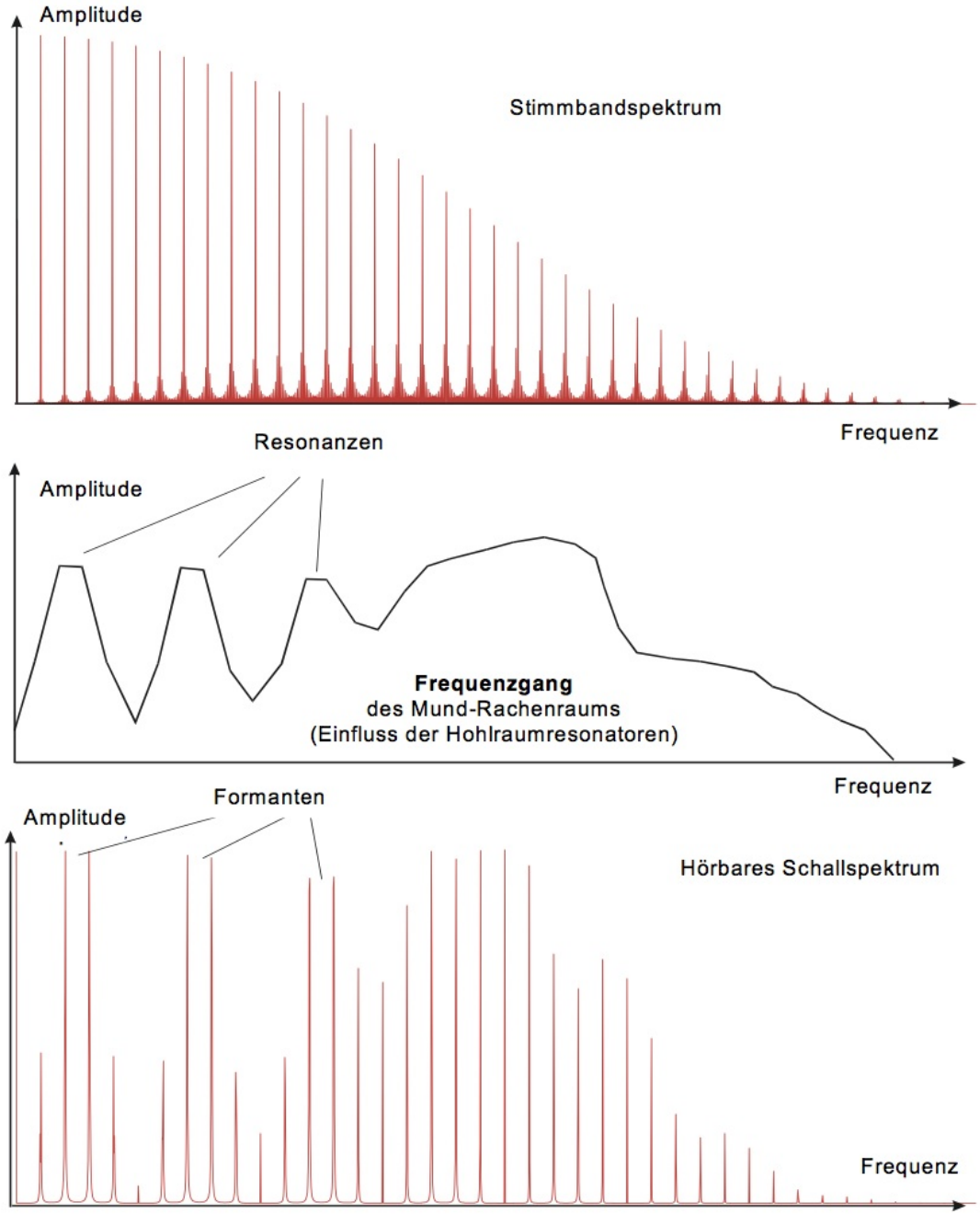
Quelle: Franz et al., 2012

Übersicht

Formanten

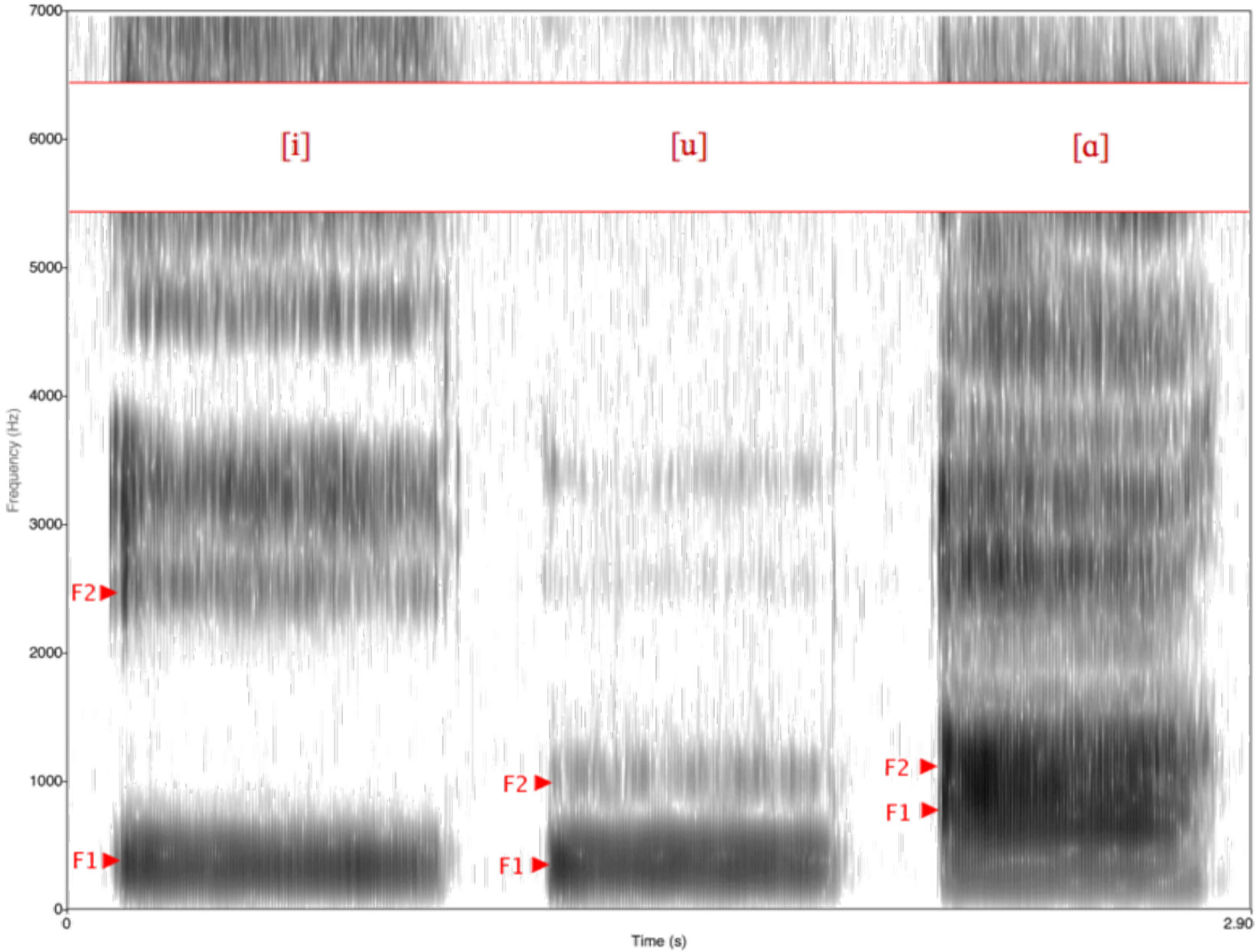
- Baut sich an den Stimmbändern durch den Luftstrom aus den Lungen ein Überdruck auf, so öffnen sich diese, der Druck baut sich ruckartig ab, sie schließen sich wieder usw. Bei Vokalen geschieht dies fastperiodisch, bei Konsonanten nichtperiodisch. Sprachgrundfrequenz: ca. 240 Hz (Frau), 130 Hz (Mann).
- Im komplexen Hohlraumsystem des Mund- und Rachenraums entstehen bei bestimmten Anregungsfrequenzen *stehende Wellen* (ähnlich wie in Orgelpfeifen oder Flöten).
- Dieser Hohlraumresonator verstärkt diejenigen Frequenzen, bei denen sich in seinem Inneren stehende Wellen bilden können und schwächt diejenigen, für die das nicht gilt, er wirkt also als frequenzselektiver Verstärker und Filter.
- Diejenigen Frequenzbereiche, bei denen die relative Verstärkung am höchsten ist, bezeichnet man als **Formanten**. Die ersten beiden Formanten f_1 und f_2 charakterisieren die Vokale, der dritte und vierte Formant f_3 und f_4 sind für das Sprachverständnis nicht mehr wesentlich.

Erzeugung von Formanten



Quelle: Karrenberg, 2012

Sonogramm von “i”, “u” und “a”

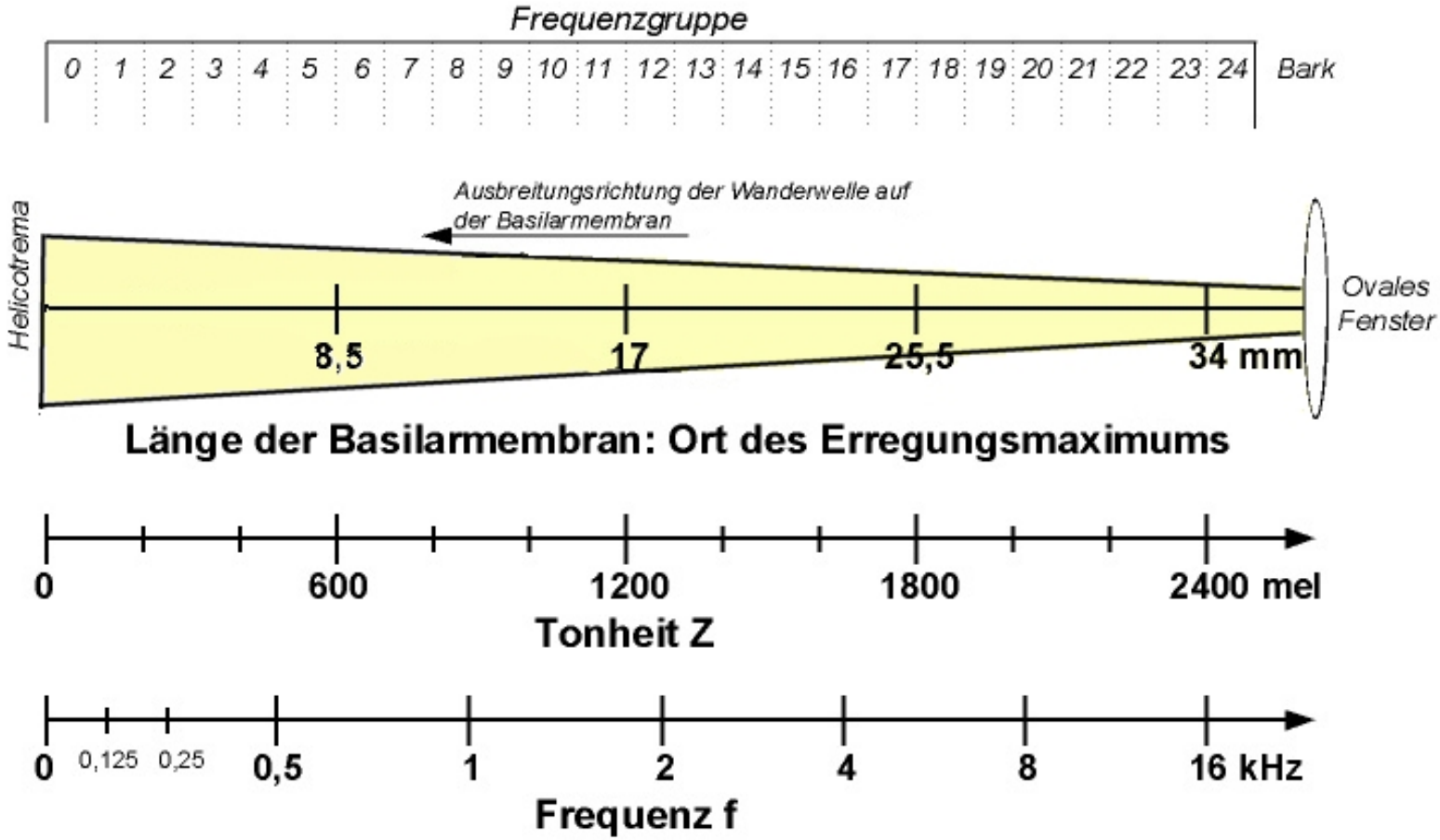


Quelle: Wikipedia

Kurzzeit-Fourieranalyse im Innenohr

- Beim Hören werden die von außen aufgenommenen Schallwellen durch die Bewegungen des Steigbügels über das ovale Fenster in der Hörschnecke (Cochlea) übertragen.
- Durch die Ausbreitung der Wanderwelle in der Innenohrflüssigkeit kommt es zur Auslenkung der Basilarmembran, welche die Cochlea in zwei Kammern teilt.
- Die Basilarmembran schwingt nicht gleichmäßig über ihre gesamte Länge, sondern zeigt ein ausgeprägtes, frequenzabhängiges Maximum. Der Ort dieses Maximums wird durch die variierende Dicke und Breite der Membran festgelegt, so dass an jeder Stelle eine unterschiedliche Frequenz bevorzugt wird.
- Die Schwingungen werden über die inneren Haarzellen an das Nervensystem weitergeleitet. Dem am ausgeprägtesten feuernden Bereich der Basilarmembran wird so eine Frequenz zugeordnet. Je näher der Ort der maximalen Auslenkung dem ovalen Fenster ist, desto höher der Ton [video].

Zuordnung von Erregungsmaximum zu Frequenz auf der Basilarmembran



Quelle: Wikipedia

Übersicht

Ein einfaches Spracherkennungssystem...

- Für einfache Aufgaben müssen oft nur wenige Wörter erkannt werden, z.B. “Hoch”, “Runter” etc.
- Für jedes zu erkennende Wort wird ein Referenzspektrum (**Prototyp**) gespeichert. Der momentane Sprachinput wird mit den Referenzspektren verglichen. Das ähnlichste Referenzspektrum wird als die wahrscheinlichste Wortbedeutung interpretiert.
- Referenzspektren können z.B. durch Mittelung über mehrere Aufnahmen desselben Wortes gewonnen werden.
- Der Vergleich kann z.B. über ein gleitendes Fenster passender Größe geschehen, oder durch den Beginn des Wortes getriggert werden.
- Ein solches Mustererkennungssystem wird als **Nächster-Nachbar-Klassifikator** oder **Prototypen-Klassifikator** bezeichnet.

Das einfachste Ähnlichkeitsmaß: Korrelation

Korrelation zweier Signale $f(t), g(t)$ oder Spektren $F(\omega), G(\omega)$:

$$C_{fg} = \int_{-\infty}^{\infty} f(t) \cdot g(t)\, dt \quad \text{bzw.} \quad C_{FG} = \int_{-\infty}^{\infty} F(\omega) \cdot G(\omega)\, d\omega.$$

- Wenn $f(t)$ und $g(t)$ ähnlich zueinander sind, dann sind sie oft gleichzeitig positiv oder gleichzeitig negativ (sie **kovariieren**). Somit ist ihr Produkt oft groß und damit ihre Korrelation.
- Sind beide zueinander unähnlich, dann bleibt das Produkt klein, weil $f(t)$ oft nahe an 0 ist, wenn $g(t)$ groß ist, und umgekehrt, oder negative und positive Produkte löschen sich im Integral gegenseitig aus.
- Im Rechner stehen nur abgetastete Werte f_k und g_k zur Verfügung. Die diskrete Korrelation berechnet sich als

$$C_{fg} = \frac{1}{n} \sum_{k=1}^{n} f_k \cdot g_k.$$

Ähnlichkeit unabhängig vom Signalmittelwert: Kovarianz

Allgemein gilt: die Korrelation ist umso höher, je größer die Mittelwerte der Signale sind, unabhängig davon, ob sie zusätzlich kovariieren oder nicht, d.h. Signale mit hohem Mittelwert sind immer "ähnlicher" bzw. stärker korreliert.

Dieser Nachteil wird vermieden, wenn bei beiden Signalen vorher der Mittelwert abgezogen wird (**Kovarianz**):

$$\sigma_{fg} = \int_{-\infty}^{\infty} (f(t) - \mu_f) \cdot (g(t) - \mu_g)\, dt$$

mit

$$\mu_f = \int_{-\infty}^{\infty} f(t)\, dt \quad \text{und} \quad \mu_g = \int_{-\infty}^{\infty} g(t)\, dt.$$

Für diskrete Signale ist die Kovarianz mit $\mu_f = (1/n) \sum_k f_k$ und $\mu_g = (1/n) \sum_k g_k$ definiert als:

$$\sigma_{fg} = \frac{1}{n} \sum_{k=1}^{n} (f_k - \mu_f) \cdot (g_k - \mu_g).$$

Korrelationskoeffizient nach Bravais-Pearson

Je höher die Standardabweichung der beiden Signale, desto höher ist ihre Kovarianz. Man teilt daher die Kovarianz durch die Standardabweichungen σ_f und σ_g und erhält so den **Korrelationskoeffizienten** nach Bravais-Pearson:

$$r_{fg} = \frac{\sigma_{fg}}{\sigma_f \cdot \sigma_g} \quad \text{mit}$$

$$\sigma_f^2 = \int_{-\infty}^{\infty} (f(t) - \mu_f)^2 \, dt \quad \text{bzw.} \quad \sigma_f^2 = \frac{1}{n-1} \sum_{k=1}^{n} (f_k - \mu_f)^2$$

- r_{fg} nahe an 1 bedeutet eine hohe Ähnlichkeit (**positive Korrelation**).
- r_{fg} nahe an 0 bedeutet keine Ähnlichkeit (**keine Korrelation**).
- r_{fg} nahe an -1 bedeutet eine "Anti-Ähnlichkeit" (**negative Korrelation**).

Architektur des Spracherkenners