

Da bist Du sprachlos:

Trennung und Rekonstruktion von Sprachsignalen

Prof. Dr.-Ing. Thorsten Herfet
Lehrstuhl Nachrichtentechnik
<mailto:herfet@nt.uni-saarland.de>



Inhalt

- Problemstellung und Anwendungsfelder
- Eigenschaften von Sprachsignalen
- Die Fouriertransformation
 - Kontinuierlich vs. Short-Time (STFT)
- Orthogonalität
- Trennung von Sprachsignalen
 - Zeit-/Frequenzmasken
 - ITD / ILD
 - Algorithmen in der STFT-Ebene
- Zusammenfassung



Begriffserklärung

- Binaurales Hören
 - Hören mit zwei Ohren / Mikrofonen
- Sprachsignale
 - Von menschlichen Sprechern erzeugte Signale am Ausgang der Mikrophone
- Multimodaler, humanoider Roboter
 - Humanoid = menschähnlich
 - Multimodal = hören und sehen





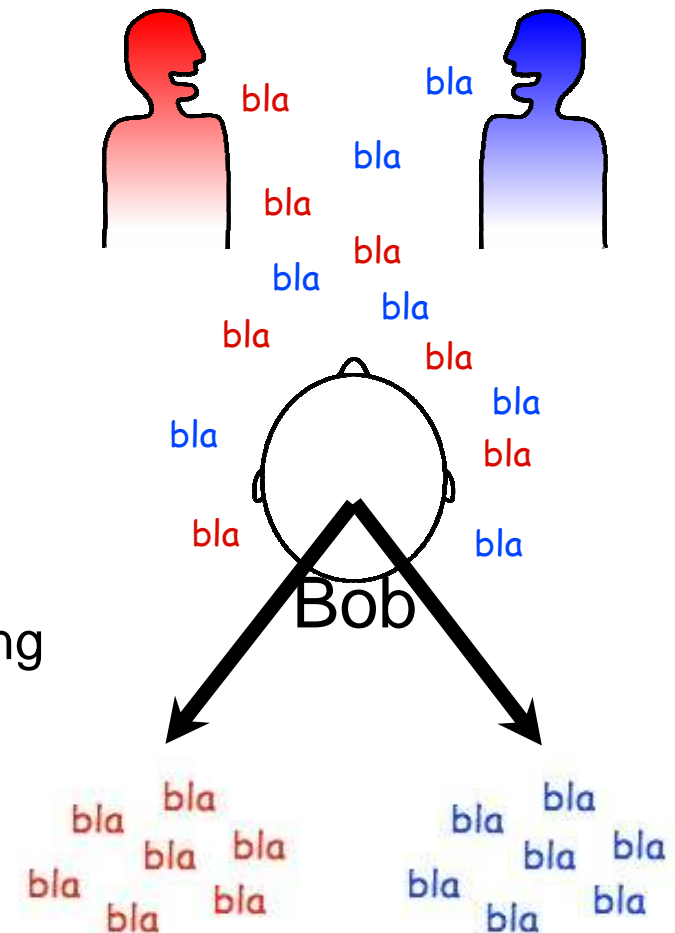
Anwendungsfelder

- **Grundlagenforschung: Kognitive Kanäle**
 - Untersuchung akustischer Quellentrennung bei vorhandenem Hintergrundwissen (z. B. menschliche Sprache)
 - Nachbildung der menschlichen Signalverarbeitung
- **Grundlagenforschung: Multimodale Quellenseparation**
 - Untersuchung optisch/akustischer Modenfusion
 - Mitarbeit im Cluster Multimodal Computing & Interaction
- **Anwendungsfelder:**
 - Humanoide Roboter / Avatare
 - Hörgeräte (Lokalisierungsunterstützung)
 - Audiokonferenzen



Problemstellung

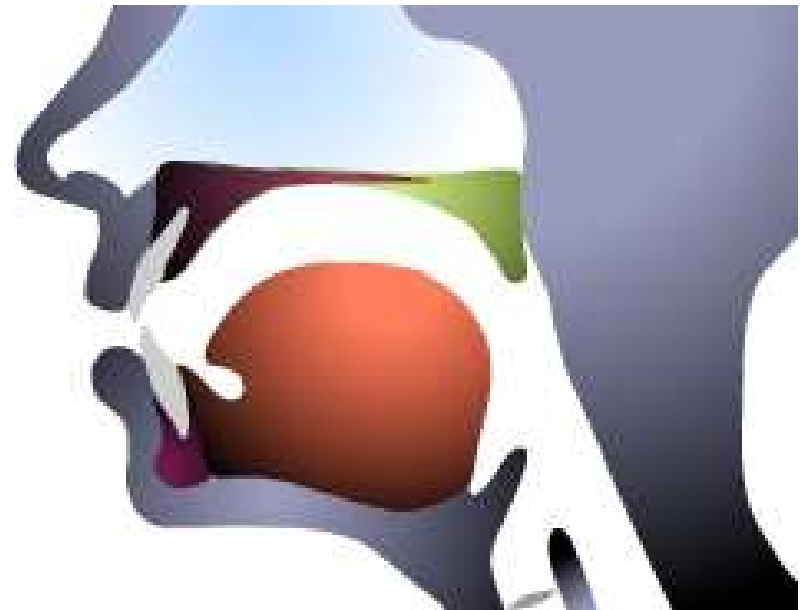
- Cocktail-Party Effekt
 - Mehrere Personen unterhalten sich
 - Verschiedene Richtungen, verschiedene Stimmen
 - Mehr Quellen als Sinken
 - Nicht durch Gleichungssystem lösbar
 - Wissen über Art der Quelle fließt ein
 - Kein „simples“ Beamforming
 - Auf Quelle angepasste Signalverarbeitung





Akustische Eigenschaften

- Obstruenten
 - Frikative: /s/, /f/, /v/, /z/, ...
 - Plosive: /g/, /k/, /d/, /p/, ...
- Resonanten
 - Vokale: /a/, /e/, /i/, ...
 - Nasale: /n/, /m/, ...





Die Fouriertransformation

- Zerlegung eines Signals in gewichtete Summe harmonischer Grundfunktionen

$$s(t) = \int_{-\infty}^{\infty} S(f) \cdot \exp(j2\pi ft) df, \quad S(f) = \int_{-\infty}^{\infty} s(t) \cdot \exp(-j2\pi ft) dt$$

- Analyse von Signalen im sog. „Spektrum“
 - Eindeutige Rücktransformation vorhanden
 - Problem: Integration über unendlich langen Zeitraum
 - Praktisch nicht implementierbar
 - Signale sind nicht stationär
 - Problem: Kontinuierlich
 - Nicht im Computer behandelbar



STFT

- Short Time Fourier Transformation
 - Endliches Fenster, erzwungen durch Fensterfunktion $w(n)$
 - Diskret (nur diskrete Frequenzen $f_q=q/(NT)$)

$$X(k, q) = \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} w(n) \cdot x(n+k) \cdot \exp\left(-j2\pi \frac{qn}{N}\right)$$

- Neue Probleme:
 - Spektraler Inhalt stationärer Signale wird verfälscht
 - Bsp. (an Tafel): $\cos(2\pi f_0 t)$
 - $0.5 \times \delta(f+f_0) + 0.5 \times \delta(f-f_0)$ im stationären Fall
 - $0.5 \times \text{sinc}(f+f_0) + 0.5 \times \text{sinc}(f-f_0)$ im STFT-Fall



OL = 0



L/2

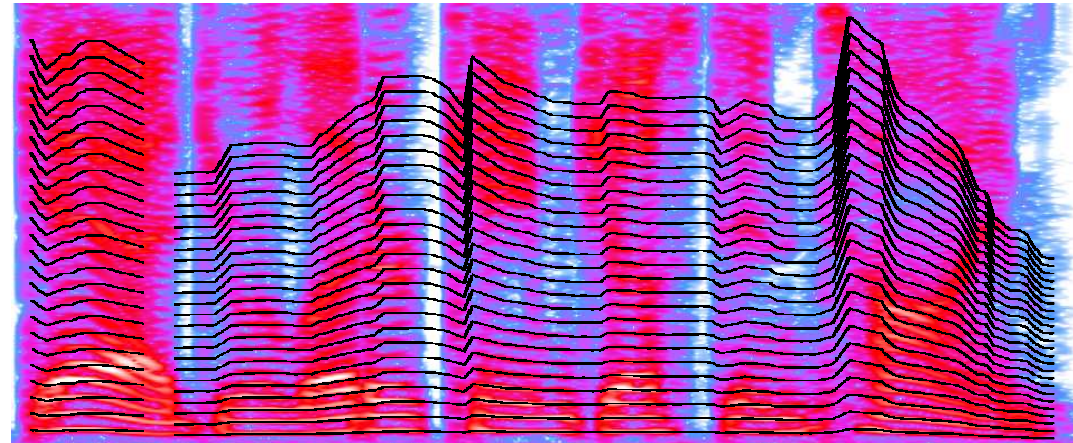
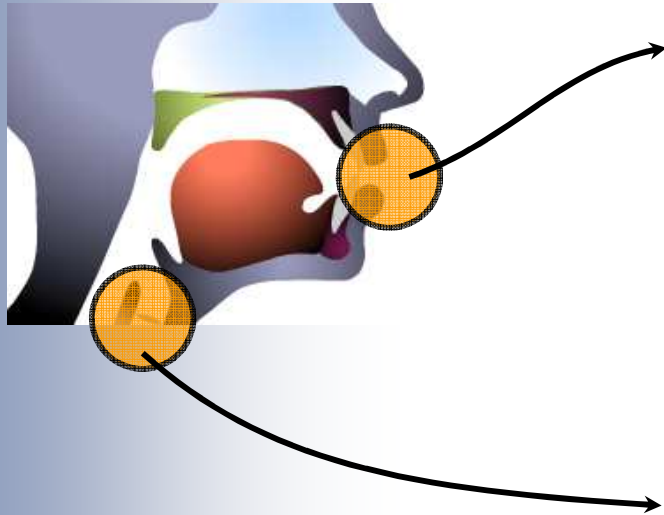


Anwendung

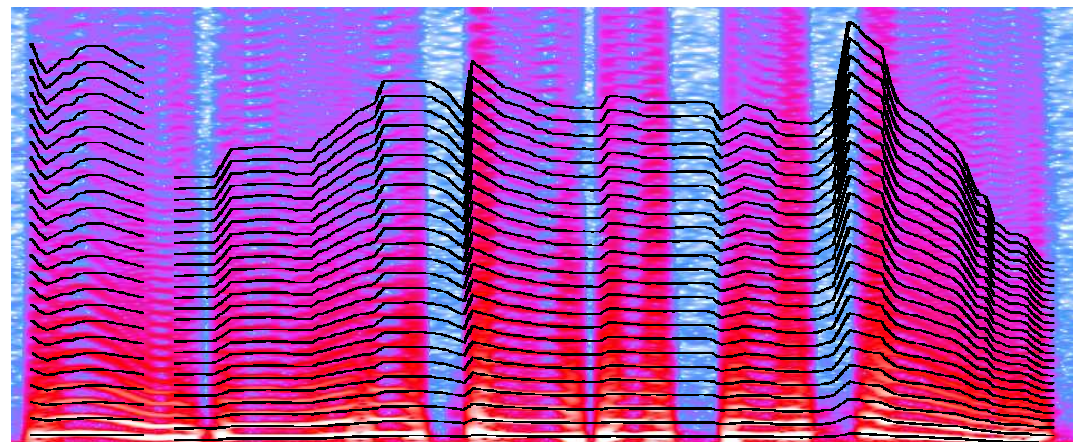
- Die STFT stellt ein probates Mittel zur Analyse dar, aber:
 - Die Fensterlänge muss sorgfältig gewählt werden
 - Typische Längen sind im Bereich 20–60 ms
 - Die **Schrittweite** muss sorgfältig gewählt werden
 - Typisch ist halbe Fensterlänge
 - Bedenke: Das rekonstruierte Signal ergibt sich durch ISTFT!
 - Signalabschnitte müssen „**nahtlos**“ **aneinander passen**
 - Die STFT behandelt das Spektrum linear
 - Alle Frequenzbänder sind gleich breit
 - Das menschliche Ohr „**hört logarithmisch**“
 - Frequenzbänder werden zu höheren Frequenzen hin breiter



Beispiel



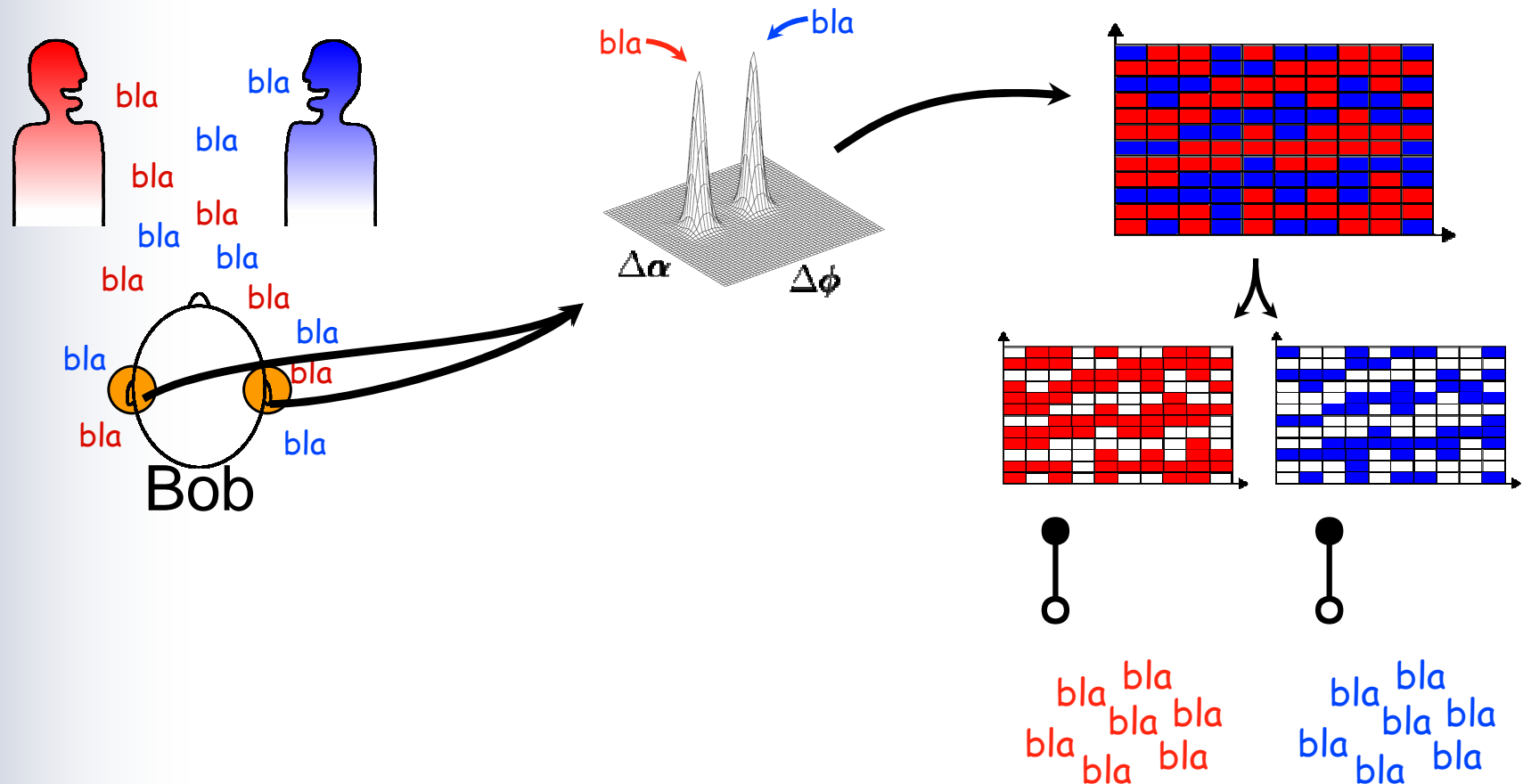
„Gad, your letter came just in time.“





DUET

- Degenerate Unmixing Estimation Technique





Orthogonalität

- Definition: Zwei Signale s und g sind orthogonal, wenn gilt:

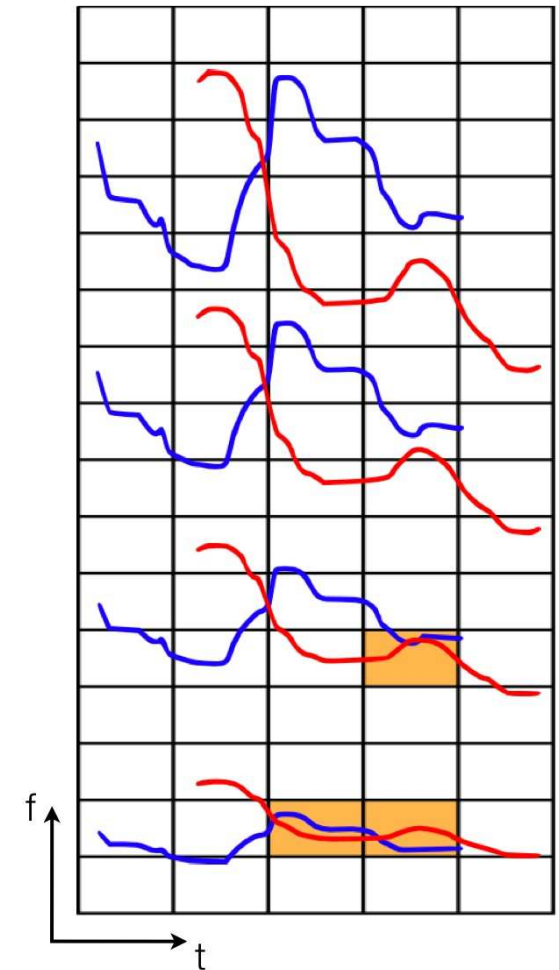
$$\int_{-\infty}^{\infty} s^*(t) \cdot g(t) dt = 0$$

- Mögliche Orthogonalitäten im:
 - Frequenzbereich (unterschiedliche Grundfrequenz)
 - Zeitbereich (kein gleichzeitiges Sprechen)
 - Raum (unterschiedliche Position der Sprecher im Raum)



Orthogonalität

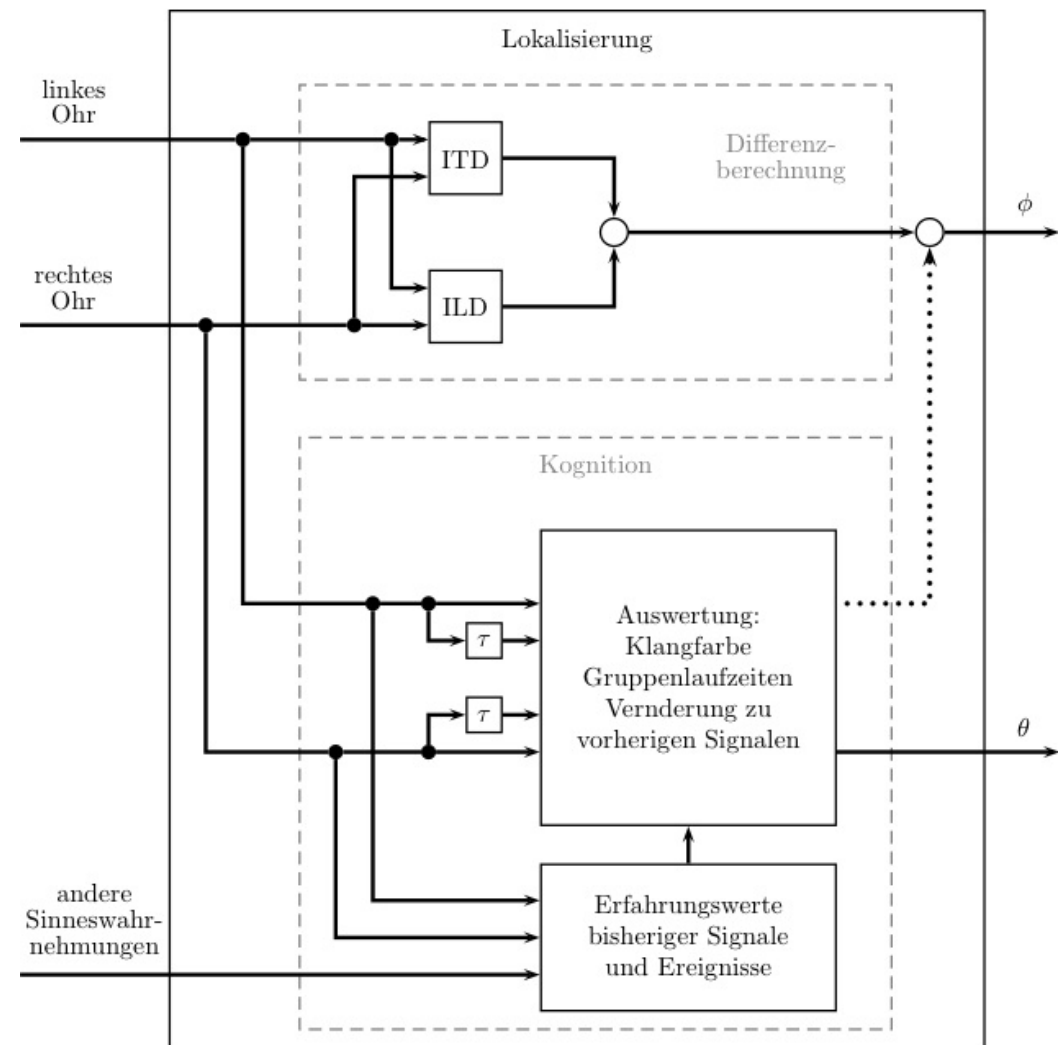
- Im Frequenzbereich
 - Eine Stimme besteht aus einer **Grundfrequenz** und dazugehörigen **Harmonischen** (*Partials*)
 - Im Allgemeinen sind die Partials zweier Personen verschieden
 - Ist der Abstand *ausreichend groß*, können zwei Stimmen separiert werden
- Probleme:
 - Welcher Abstand ist „ausreichend groß“?
 - Wie werden sich kreuzende Partials behandelt?





Lokalisierung

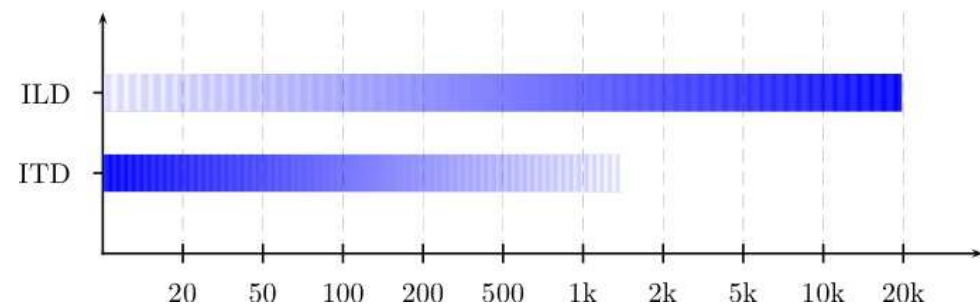
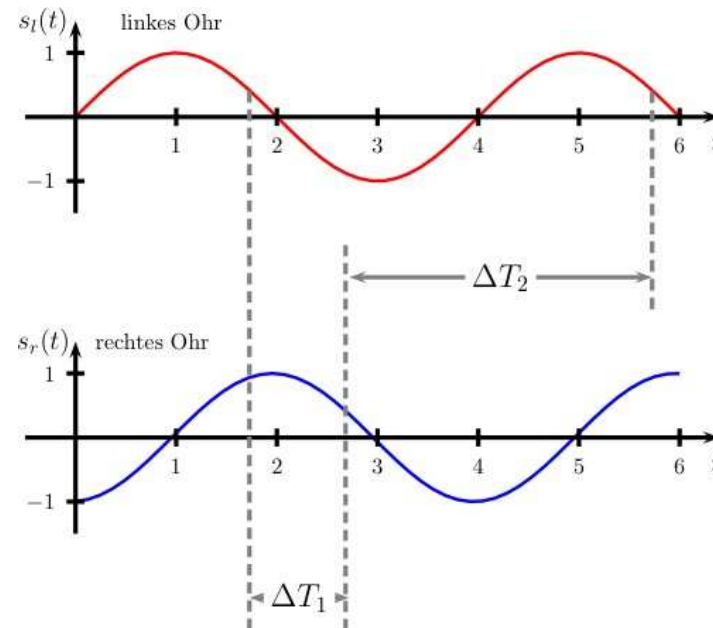
- Medianebene:
 - Laufzeitdifferenz (ITD)
 - Pegeldifferenzen (ILD)
- Horizontalebene:
 - Veränderungen in Klangfarbe
 - Verzerrungen im Bereich 1-10 kHz
- Verbesserung der Ergebnisse durch marginale Kopfbewegungen





Lokalisierung

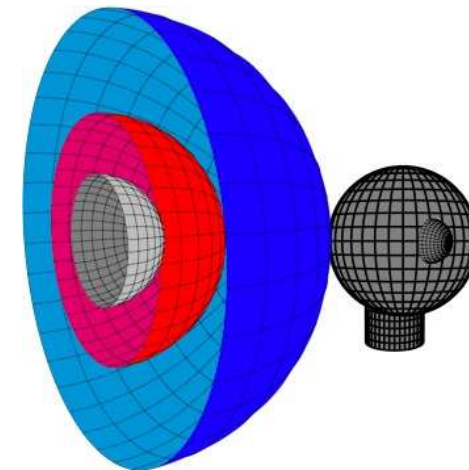
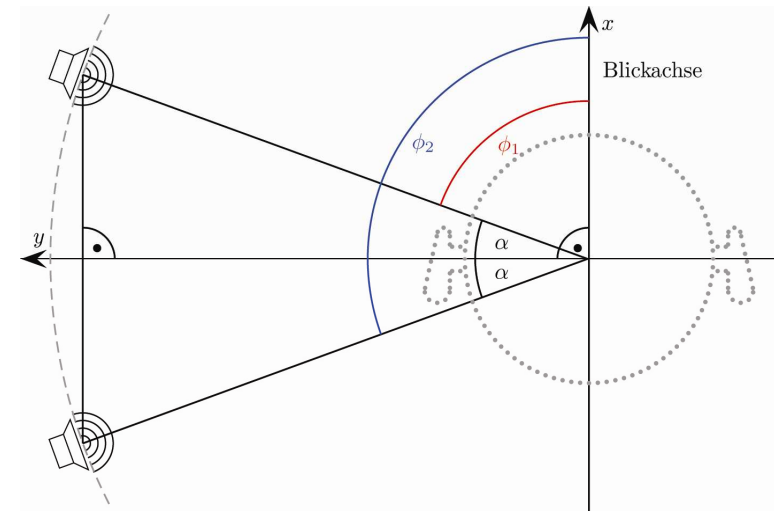
- ITD
 - Unterschiedliche Laufzeit zu den Ohren führt zur Phasenverschiebung zwischen den Signalen →
- ILD
 - Abschattung durch den Kopf führt zu unterschiedlichen Pegeln der Signale
- Duplex-Theorie:
 - ITD bei *tiefen* Frequenzen
 - ILD bei *hohen* Frequenzen





Lokalisierung

- Probleme:
 - Symmetrie zur Frontalebene führt zu einer Hinten-Vorne-Vertauschung
 - Lokalisierung in der Medianebene nicht durch Differenzmessung möglich
 - Beide Probleme führen zu einem *Kegel der Verwirrung*





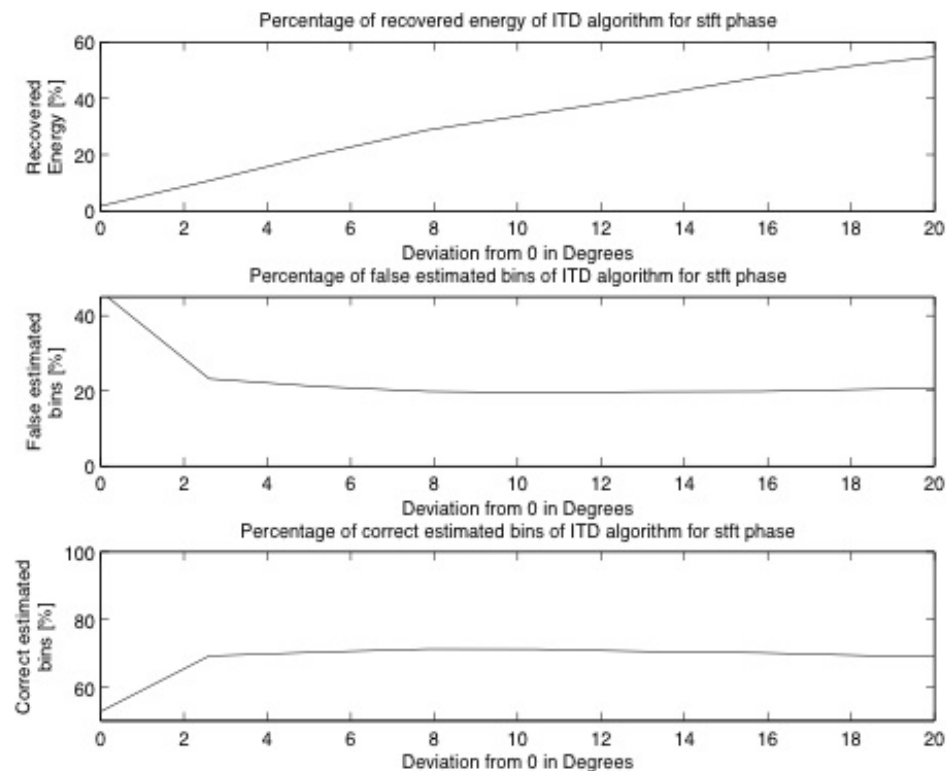
Anwendung ITD auf STFT

- Auswirkungen der ITD auf das STFT-Spektrum
 - ITD führt zu einer Phasenverschiebung
 - Signale von unterschiedlichen Positionen werden somit unterschiedlich in ihrer Phase verschoben
 - Gilt für die Fundamentalfrequenz (vgl. nächste Folie)
 - Harmonische werden entsprechend ihres Frequenzunterschieds verschoben
- Separation der Signale
 - Anteile im Spektrum (Bins) können aufgrund Ihrer Phase einer Quelle zugeordnet werden



Anwendung ITD auf STFT

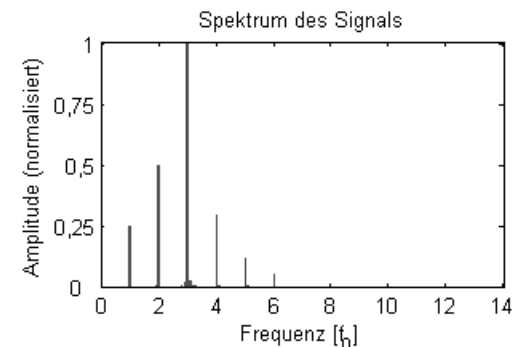
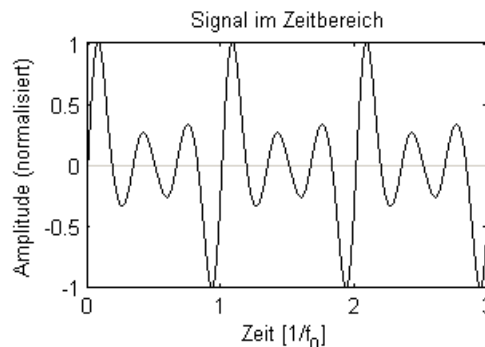
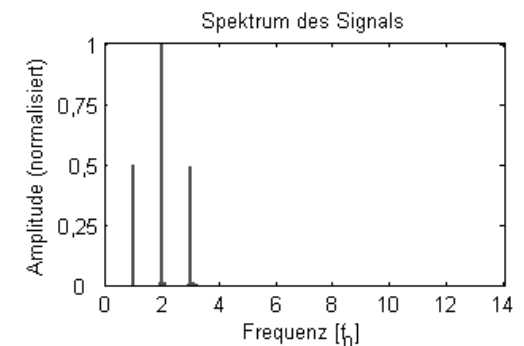
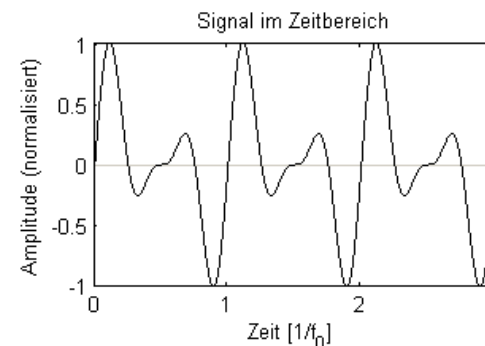
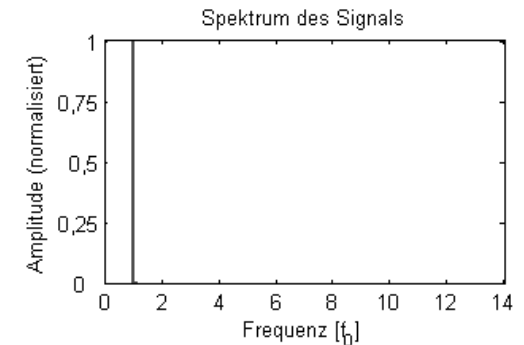
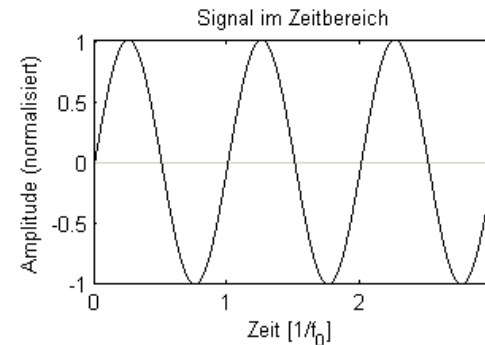
- Berücksichtigung der Phasenlage
 - Es wird **eine** Quelle zur Betrachtung ausgewählt
 - Ein Schwellwert der maximalen Phasendifferenz wird festgelegt
 - Alle Bins, deren Phase diesen Grenzwert unterschreiten, werden der betrachteten Quelle zugeordnet





Fundamentalfrequenzanalyse

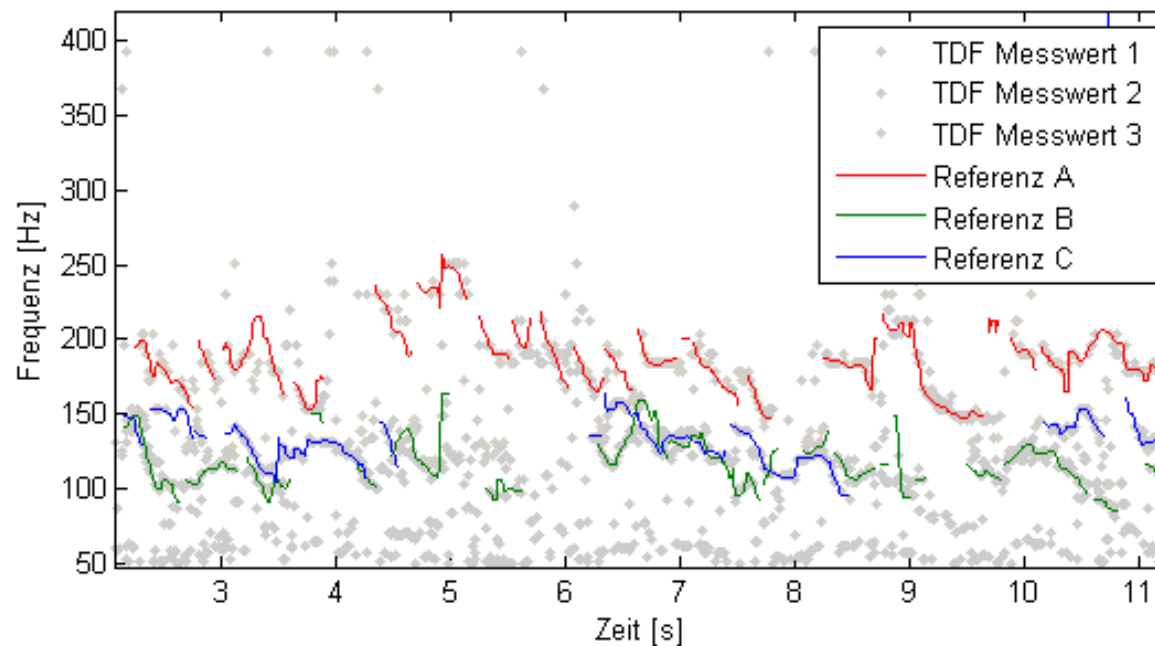
- Menschlicher Stimmapparat erzeugt harmonische Signale
- Korespondierende Fundamentalfrequenz variiert jedoch kontinuierlich beim Sprechen
- Zur Analyse muss die Fundamentalfrequenz nachgeführt werden





Fundamentalfrequenzanalyse

- Bei mehreren Sprechern müssen die jeweiligen Frequenzen getrennt verfolgt werden
- Problem entstehen hierbei bei Überlappungen und Kreuzungen

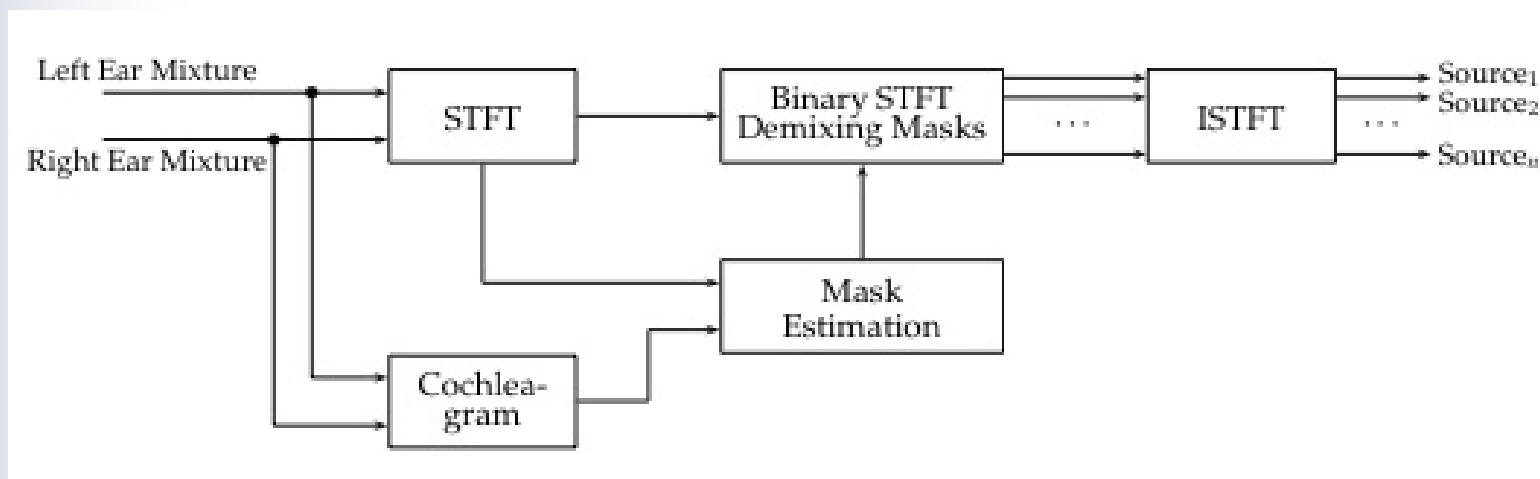




Quellentrennung

- Binary masks
 - Signalanteile, die von einer Quelle um mehr als T gegenüber anderen dominiert werden, werden dieser Quelle zugesprochen

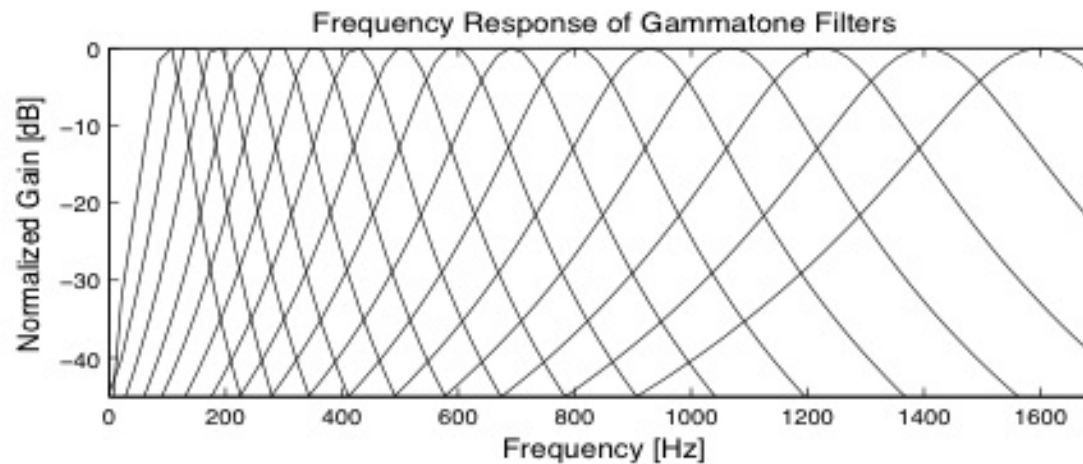
$$\Omega_i(t, f) = \begin{cases} 1 & s_i(t, f) - n_j(t, f) > T \quad \forall j \\ 0 & \text{else} \end{cases}$$










Quellentrennung

- Das menschliche Ohr „hört logarithmisch“
 - Nachbildung durch eine Gammatone Filterbank (Cochleagramm)





Quellentrennung

- Mixture Links 
- Mixture Rechts 
- Source 1 
- Source 1 ideal binary 
- Source 1 ideal non binary 



Zusammenfassung

- Binaurale Analyse und Trennung von Sprachsignalen
 - Orthogonalität
 - Trennung im Frequenzbereich (STFT, DUET)
 - Analyse im Cochleagramm
- Ergebnisse reflektieren „kognitive“ Signaltrennung
 - Anwendung in digital Health
 - Anwendung Multimodal Computing & Interaction
- Algorithmen ermöglichen gute Signaltrennung
 - Aber immer noch aktuelle Forschung!!