

Robust Sound Source Localization Using a Microphone Array on a Mobile Robot

(Jean-Marc Valin, François Michaud, Jean Rouat,
Dominic Létourneau)

Marc Pfeifer (Embedded Systems Engineering)

Proseminar – Algorithmen für Rechnernetze

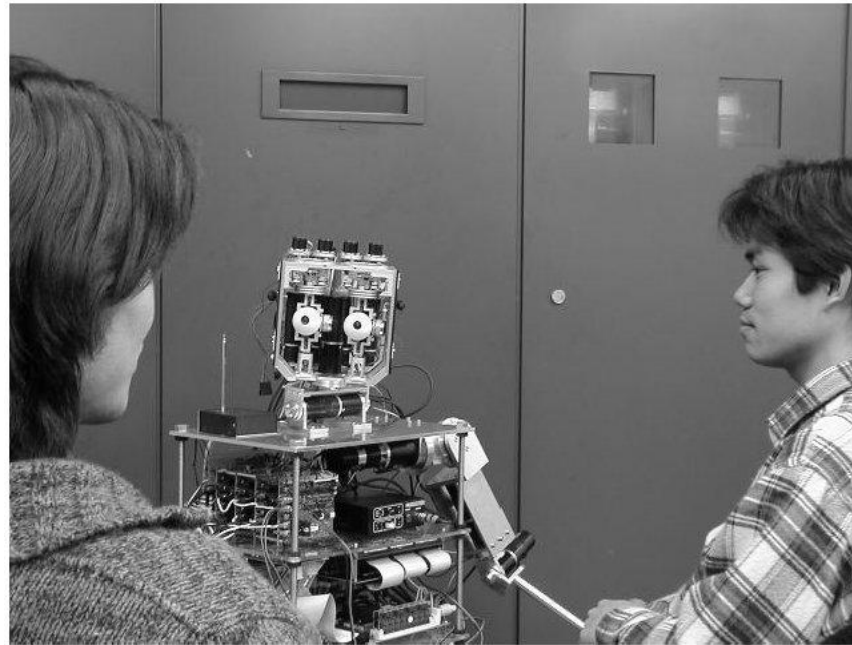
Dozenten: Prof. Dr. C. Schindelhauer

Albert-Ludwigs-Universität Freiburg



UNI
FREIBURG

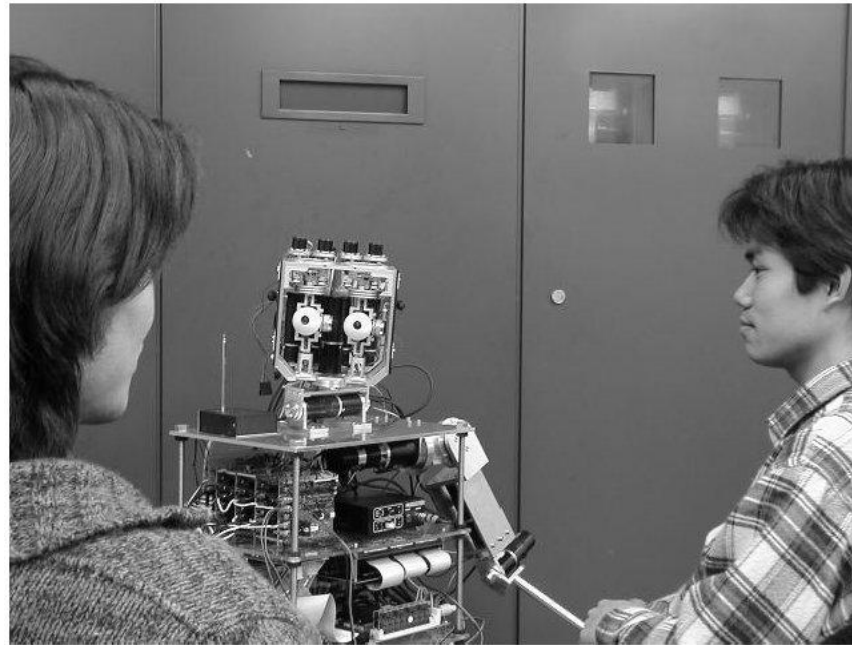
Menschliches Hören bei Robotern (1)



ROBITA [2]

- Probleme:
 - Ortung nur zweidimensional
 - Keine Unterscheidung zwischen hinten und vorne

Menschliches Hören bei Robotern (1)



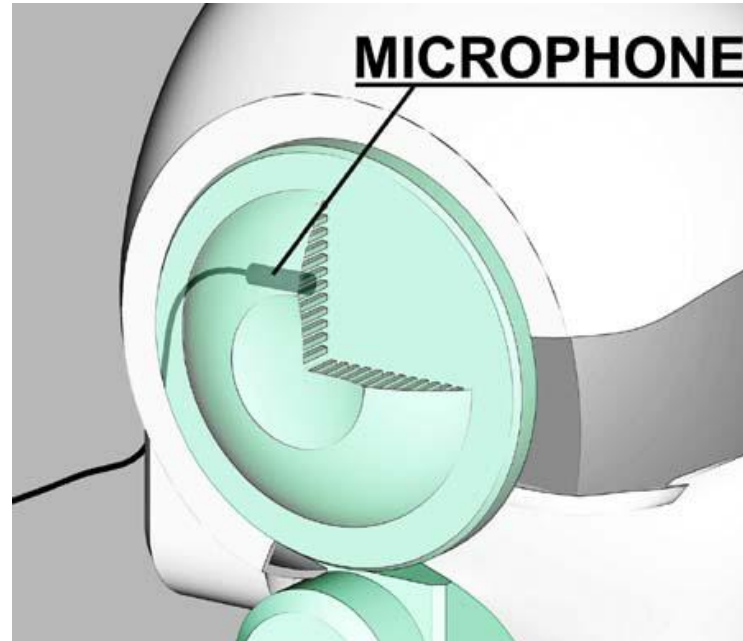
ROBITA [2]

- Probleme:
 - Ortung nur zweidimensional
 - Keine Unterscheidung zwischen hinten und vorne
- Lösung: Nutzung von Systemen, die dem menschlichen Gehör ähneln

Menschliches Hören bei Robotern (2)



SIG [3]



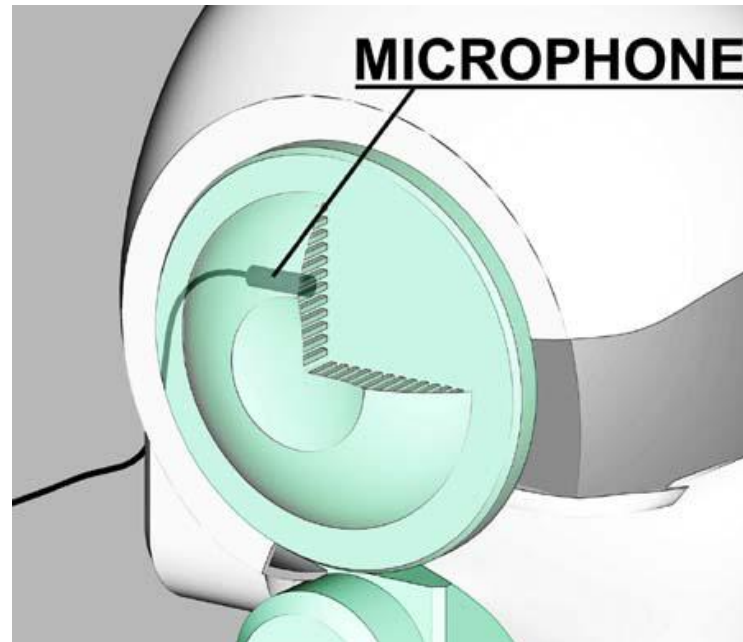
[3]

- Probleme:
 - Sehr komplexe Berechnung
 - Teure Mikrofone nötig

Menschliches Hören bei Robotern (2)



SIG [3]



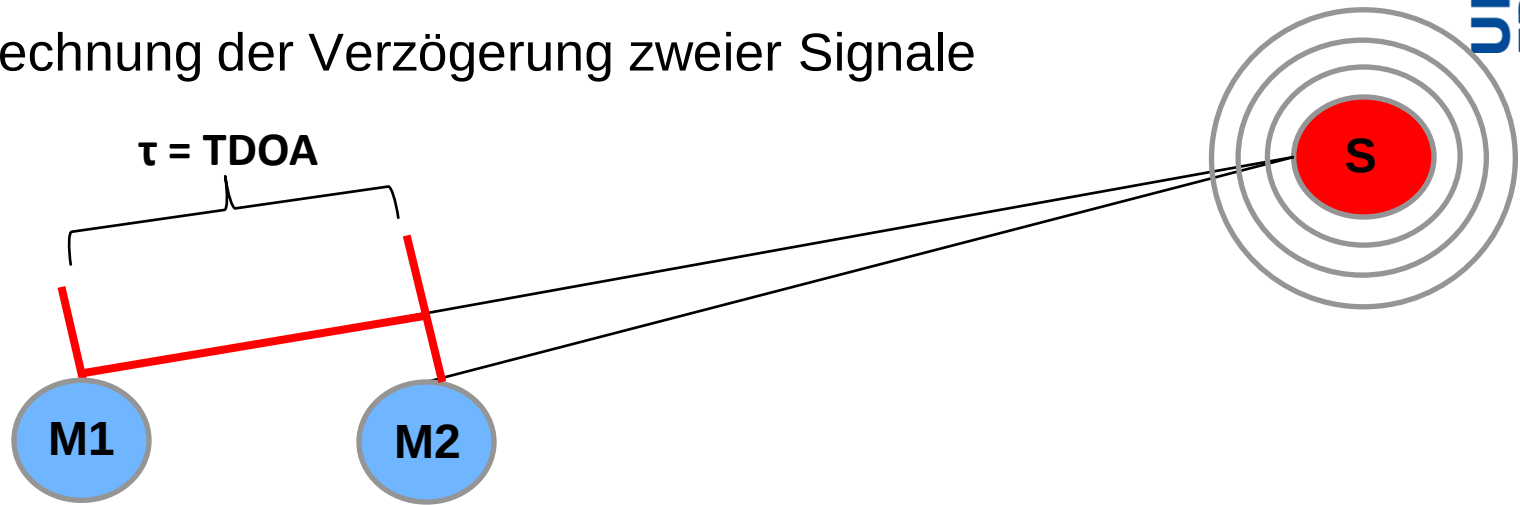
- Probleme:
 - Sehr komplexe Berechnung
 - Teure Mikrofone nötig
- Lösung: Nutzung von mehr als zwei Mikrofonen

- Motivation
- Theorie
 - Time Delay of Arrival (TDOA) – Abschätzung
 - Spektrale Gewichtung
 - TDOA - Abschätzung mit N Mikrofonen
 - Positions-Abschätzung
- Experimentelle Ergebnisse
- Zusammenfassung

Time Delay of Arrival - Abschätzung



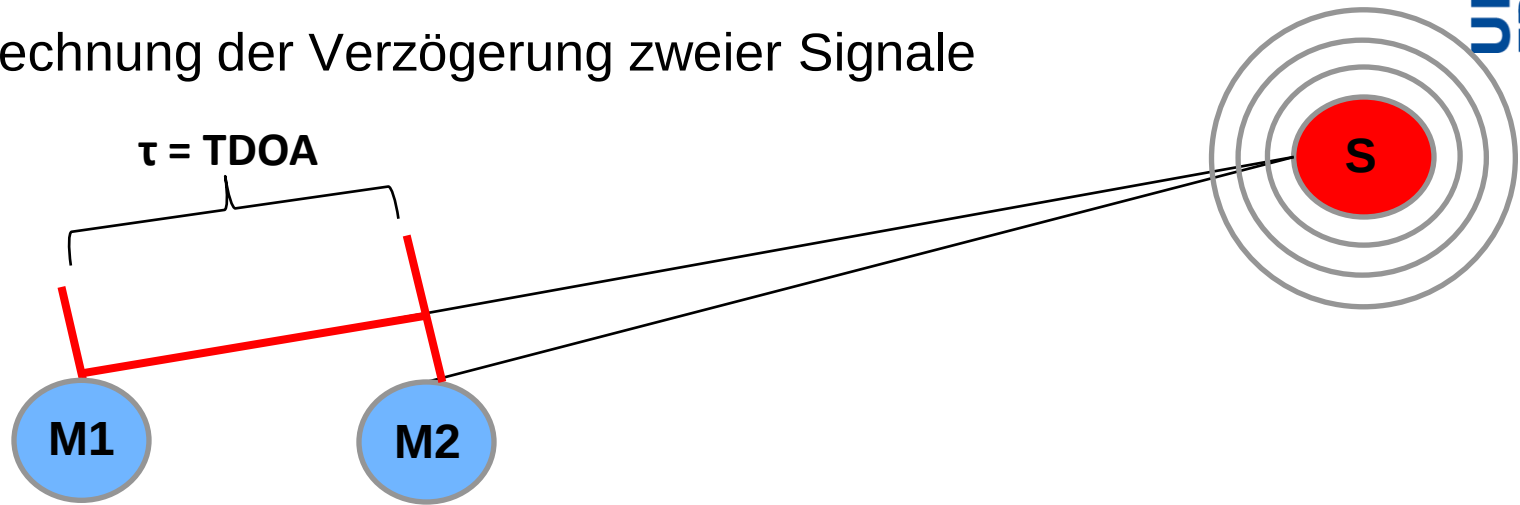
- Berechnung der Verzögerung zweier Signale



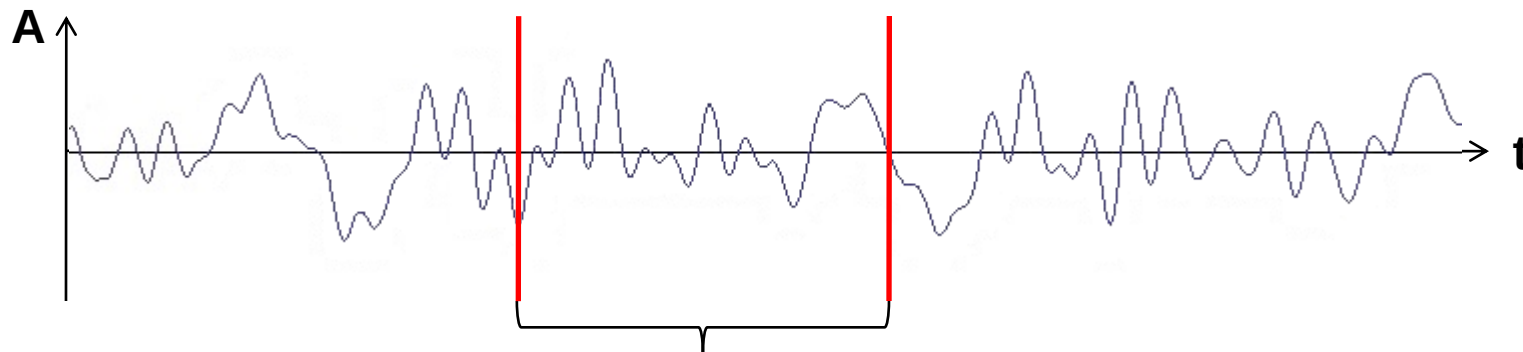
Time Delay of Arrival - Abschätzung



- Berechnung der Verzögerung zweier Signale



- Benötigt: Korrelations-Messung zweier Signale



Window mit N Vergleichspunkten

- Die Kreuzkorrelation mit zwei Mikrofonen kann beschrieben werden durch:

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n] \cdot x_j[n - \tau]$$

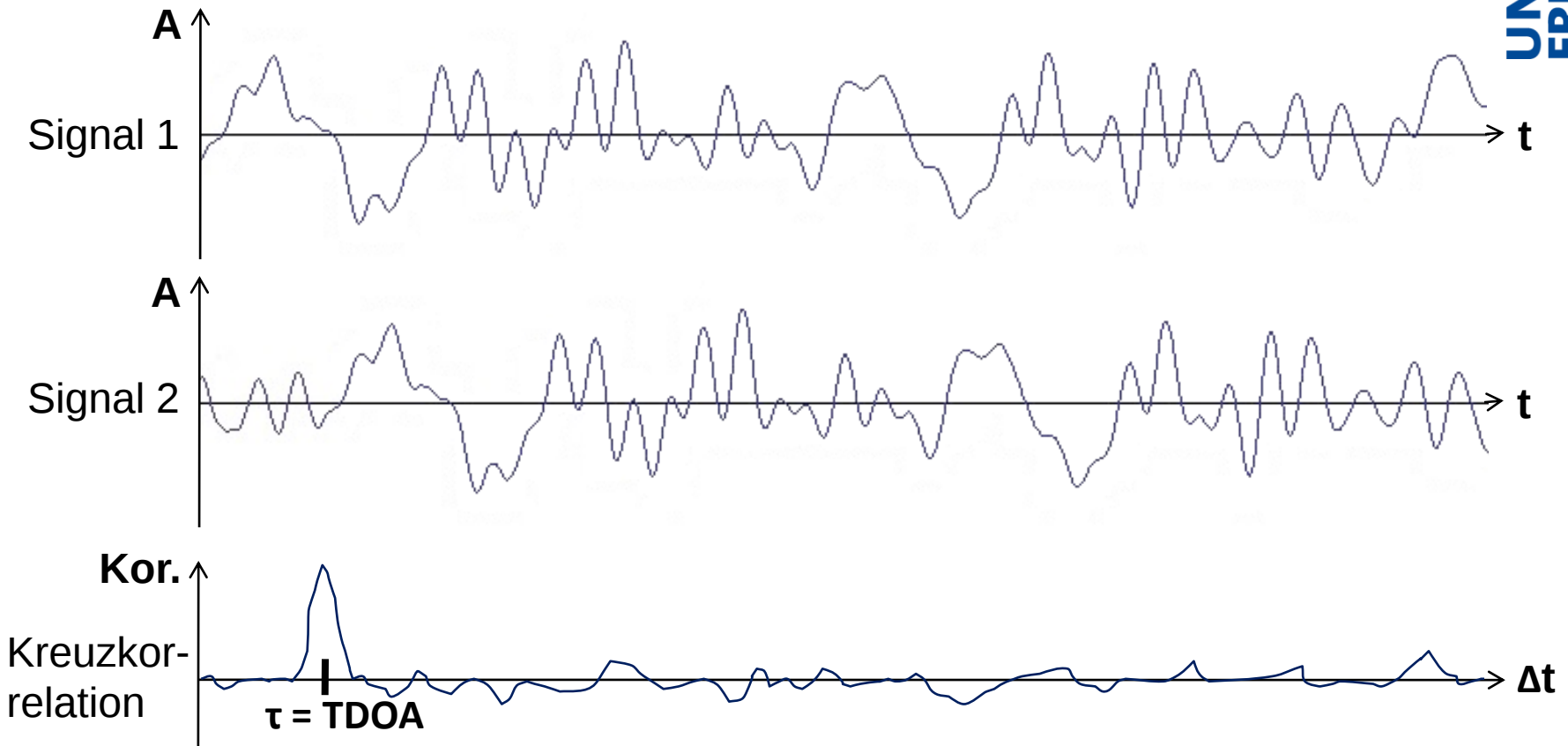
x_i = empfangenes Signal des ersten Mikrofons

τ = Verschiebung

- R_{ij} ist maximal, wenn τ der Verzögerung des Signals an den beiden Mikrofonen entspricht.

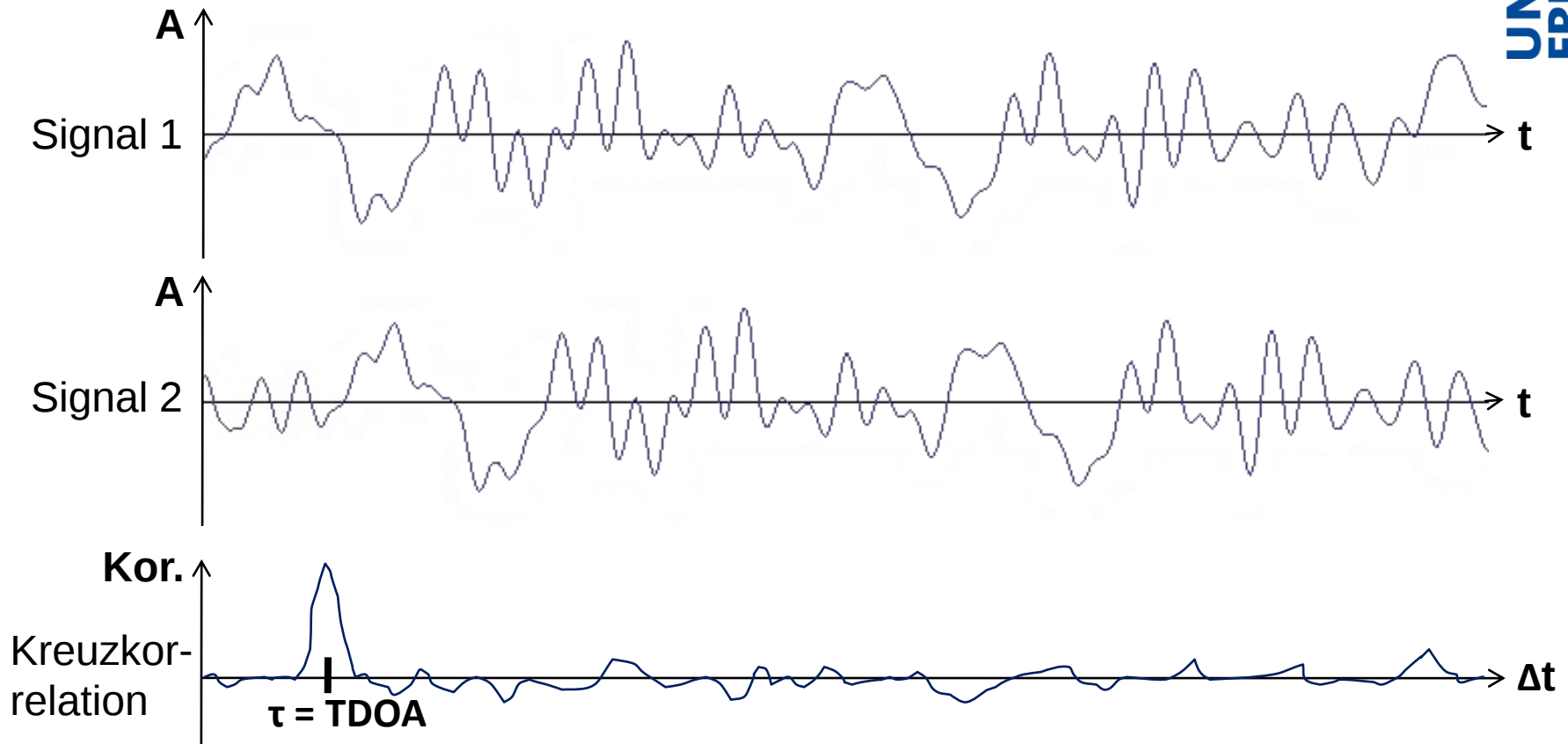
→ Dieses τ' ist in diesem Fall die gesuchte TDOA

Kreuzkorrelation – Ein Beispiel



$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n] \cdot x_j[n - \tau]$$

Kreuzkorrelation – Ein Beispiel



- Problem: Hohe Komplexität $\rightarrow O(N^2)$
- Lösung $\rightarrow$ Übertragung des Signal ins Frequenzspektrum

Verbesserung der Korrelation (1)



- Übertragung der zeitabhängigen Signale $x_i[n]$ und $x_j[n]$ ins Frequenzspektrum

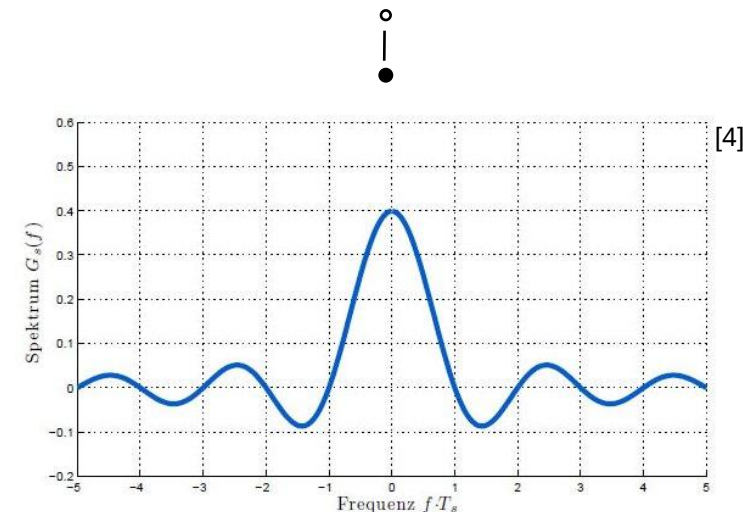
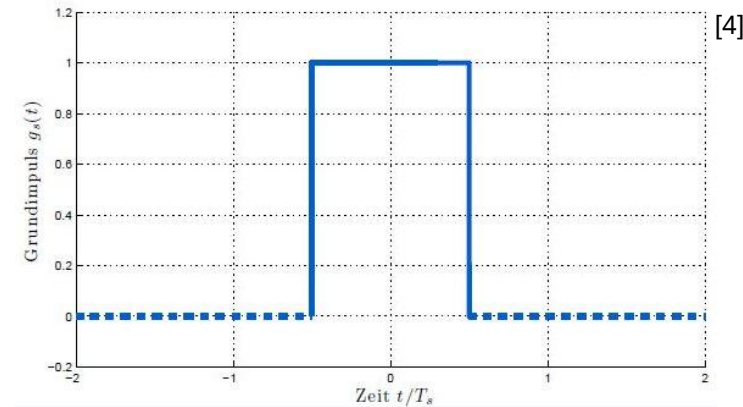
Fourier-Transf. [5]:
$$X(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-j2\pi ft} dt$$

- Diskrete Fourier-Transformation [5]:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \cdot e^{-j \frac{2\pi nk}{N}}$$

$k = 1, 2, \dots, N-1$

- → Frequenzspektren der Signale $x_i[n]$ und $x_j[n]$: $X_i[k]$ und $X_j[k]$



Verbesserung der Korrelation (2)



- Bildung des Kreuzleistungsspektrums:
- Rücktransformationen in ein zeitliches Signal
→ Inverse Fourier-Transformation [5]:

$$x(t) = \int_{-\infty}^{\infty} X(f) \cdot e^{j2\pi ft} df$$

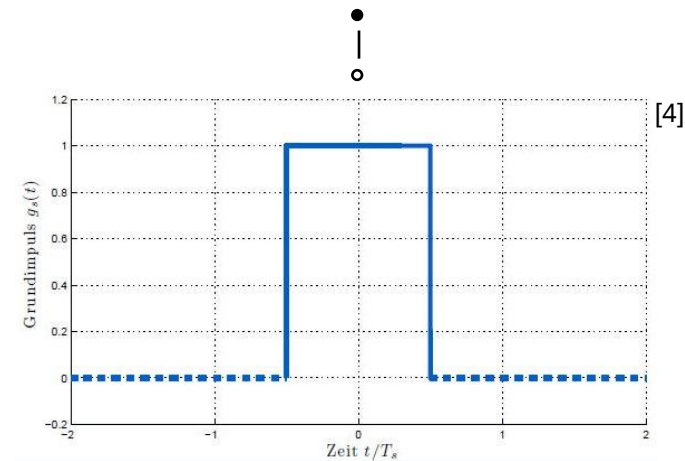
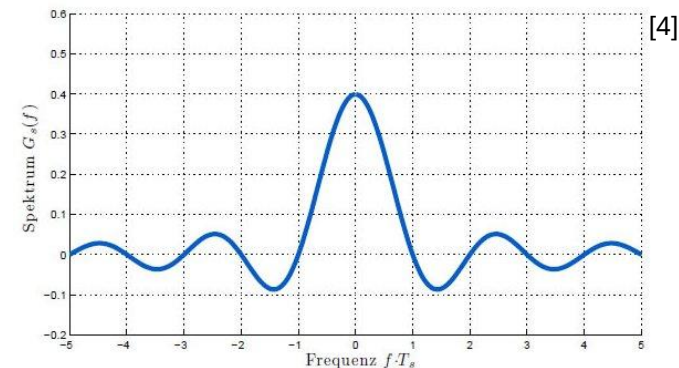
- Diskrete Anwendung auf das Kreuzleistungsspektrum:

$$R_{ij}(\tau) \approx \sum_{k=0}^{N-1} X_i(k) X_j(k)^* e^{j2\pi k\tau / N}$$

→ Zeitliches Signal der Verzögerung

→ Verbesserte Komplexität: $O(N \log_2 N)$

$X_i[k] X_j[k]^*$

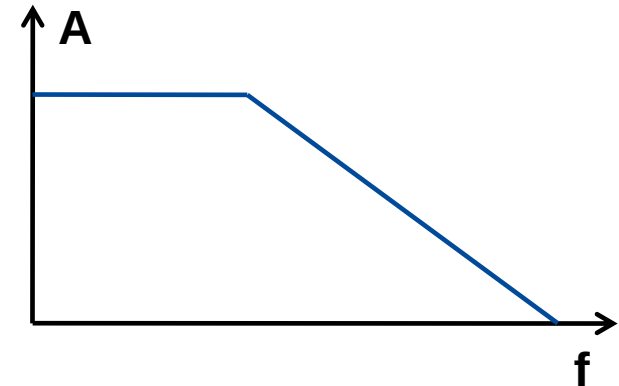


Ausbleichen des Spektrums (1)



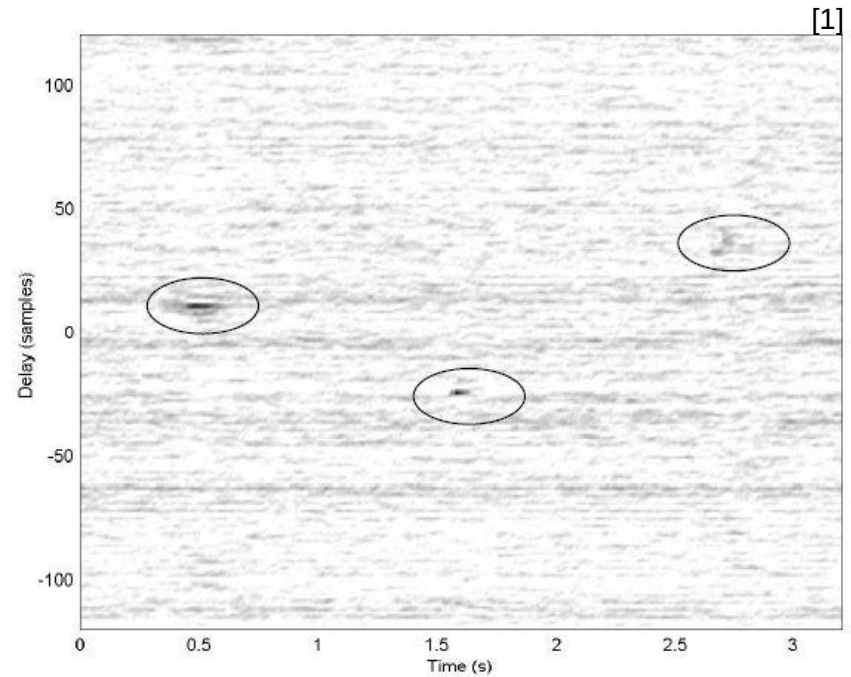
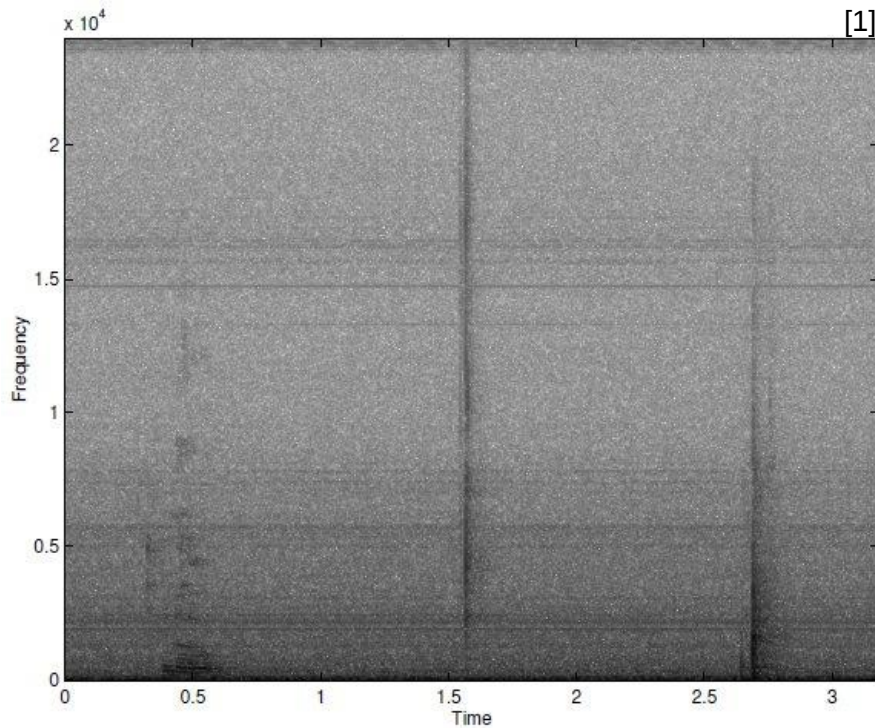
- Problem: Viele Geräusche Low-Pass → breite Kreuzkorrelations-Peaks bei aufeinanderfolgenden Geräuschen
- Lösung: Ausbleichen des Spektrums vor der Korrelationsberechnung:

$$R_{ij}(\tau) \approx \sum_{k=0}^{N-1} \frac{X_i(k)X_j(k)^*}{|X_i(k)||X_j(k)|} e^{j2\pi k\tau/N}$$

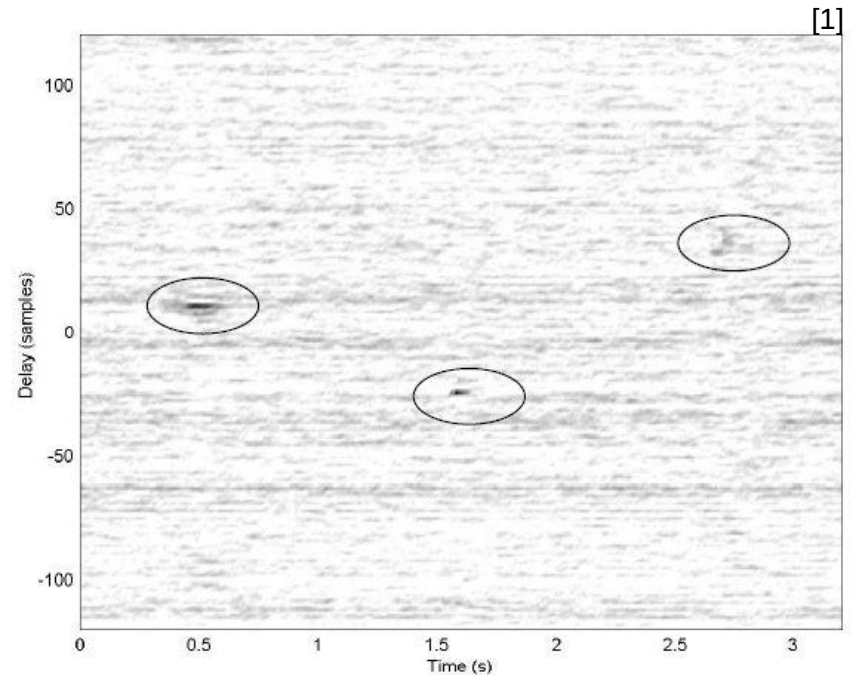
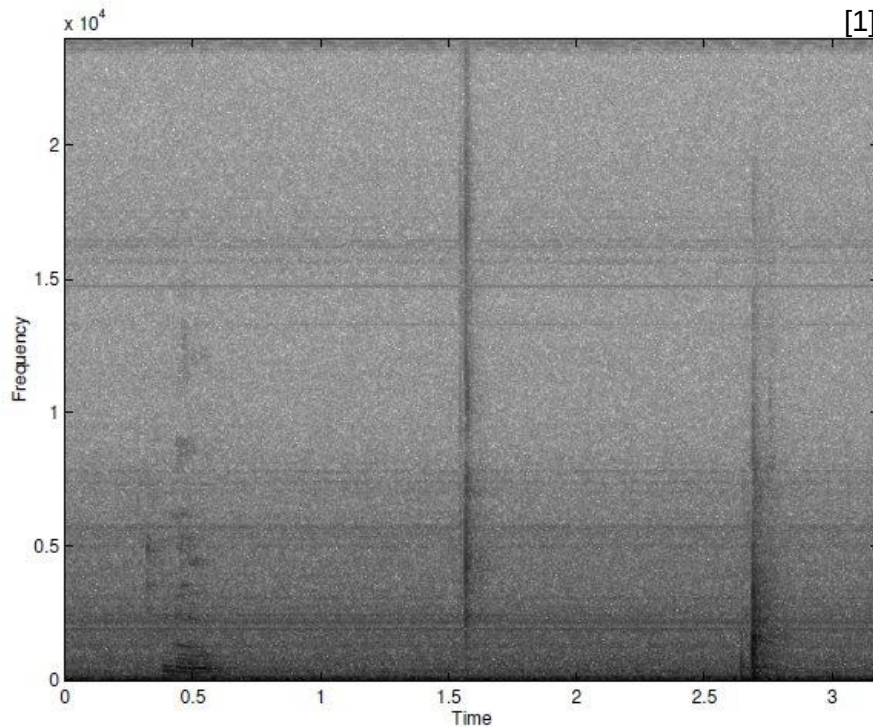


- → Normalisierung der Amplituden
- → Nur noch Phase ist wichtig
- → Jede Frequenz-Komponente bekommt gleiche Gewichtung

Ausbleichen des Spektrums (2)



Ausbleichen des Spektrums (2)



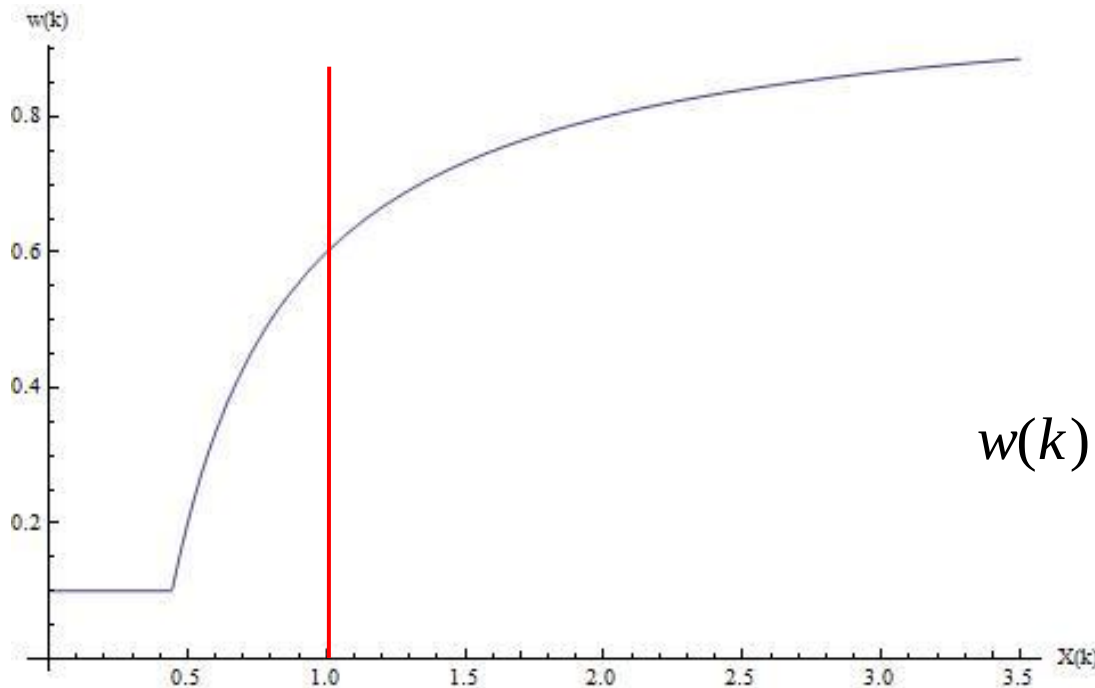
- → Problem: Auch leise Geräusche (Störgeräusche) tragen stark zur Korrelation bei
- → Geringere Robustheit

- Lösung der Autoren: Spektrale Gewichtung
- Regionen mit hohem Signal-Rausch-Verhältnis (SNR) werden stärker gewichtet.
- $X(k)$ ist die spektrale Leistungsdichte aller Mikrofone und $X_n(k)$ sei die Rauschabschätzung, basierend auf dem Zeitmittelwert von $X(k)$

$$w(k) = \max\left(0.1, \frac{X(k) - \alpha X_n(k)}{X(k)}\right)$$

- $\alpha < 1$ bestimmt den Einfluss der Rauschabschätzung

Spektrale Gewichtung (2)



- $X_n(k) = 1$

- $\alpha = 0,4$

$$w(k) = \max\left(0.1, \frac{X(k) - 0,4 \cdot 1}{X(k)}\right)$$

- $w(k) \rightarrow 0$ in rauschdominierten Regionen
- $w(k) \rightarrow 1$ in signaldominierten Regionen

Spektrale Gewichtung (3)

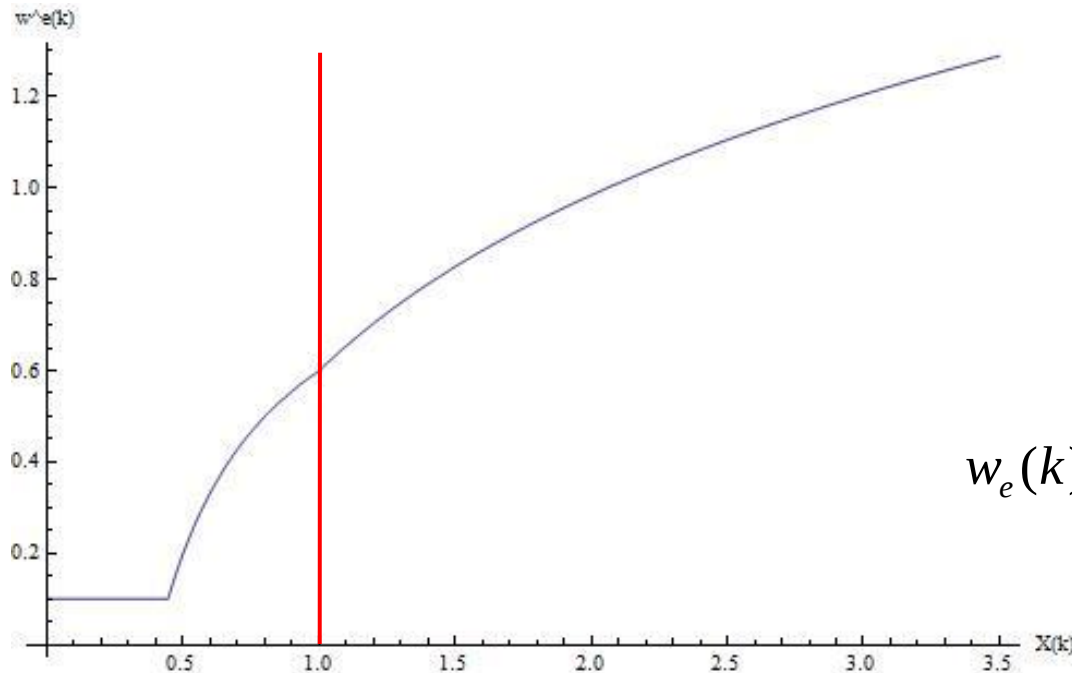


- Durch $w_e(k)$ werden Regionen mit hohem SNR stärker gewichtet:

$$w_e(k) = \begin{cases} w(k) & , X(k) \leq X_n(k) \\ w(k) \left(\frac{X(k)}{X_n(k)} \right)^\gamma & , X(k) > X_n(k) \end{cases}$$

- $0 < \gamma < 1$ entspricht dem Gewichtungsfaktor
- Empirische Setzung der Koeffizienten: $\alpha = 0,4$ $\gamma = 0,3$

Spektrale Gewichtung (4)



- $X_n(k) = 1$

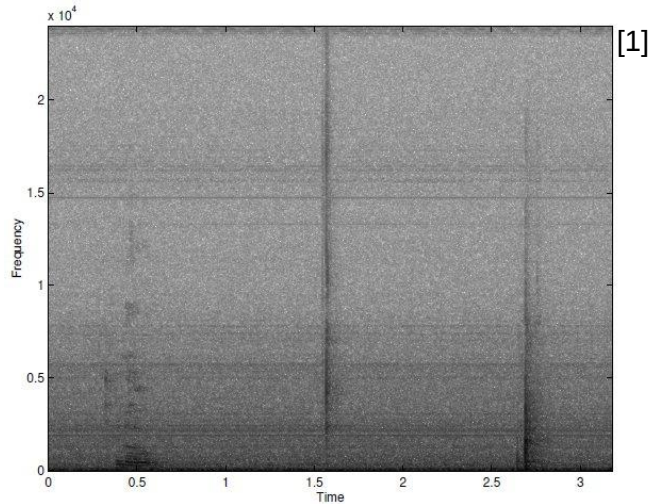
- $\gamma = 0,3$

$$w_e(k) = \begin{cases} w(k) & , X(k) \leq 1 \\ w(k) \left(\frac{X(k)}{1} \right)^{0,3} & , X(k) > 1 \end{cases}$$

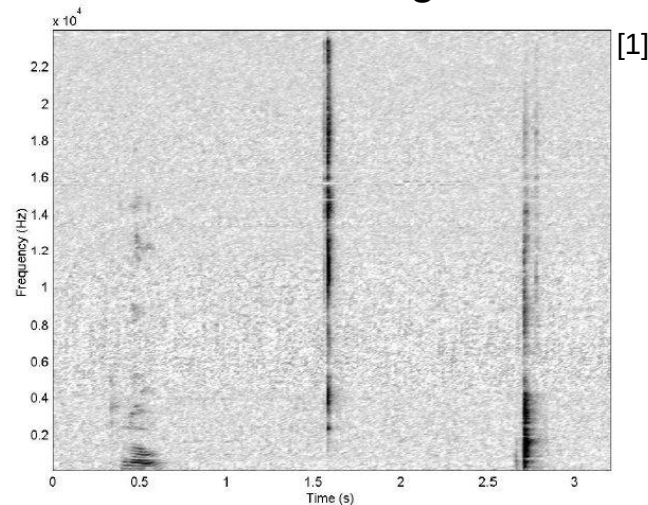
- Die gewichtete Korrelation:

$$R^{(e)}_{ij}(\tau) \approx \sum_{k=0}^{N-1} \frac{w_e^2(k) X_i(k) X_j(k)^*}{|X_i(k)| |X_j(k)|} e^{j2\pi k\tau / N}$$

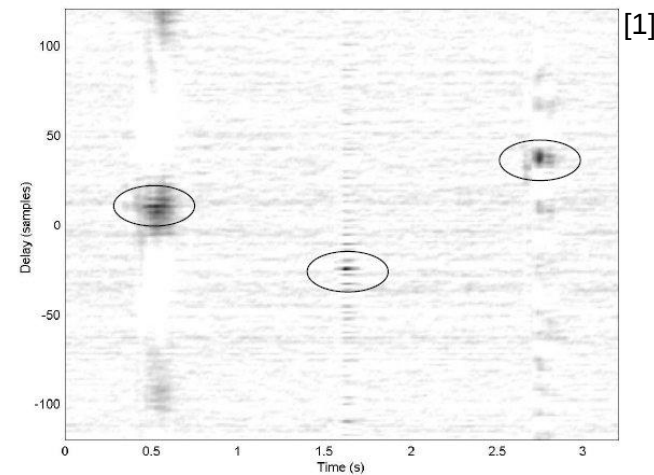
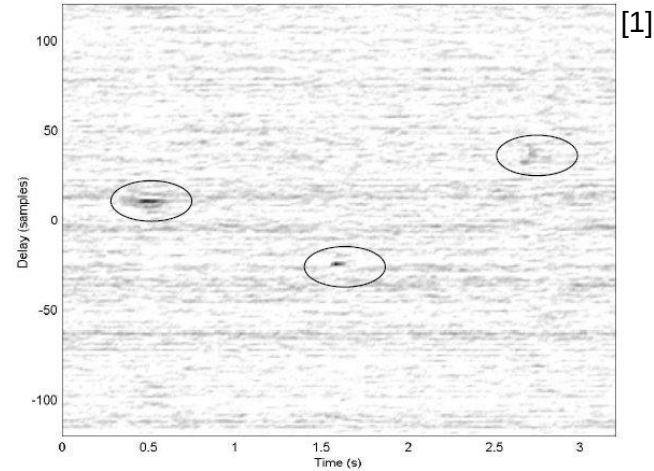
Spektrale Gewichtung - Ergebnis



Ohne Gewichtung



Mit Gewichtung



TDOA-Abschätzung mit N Mikrofonen



- Berechnung der TDOA zwischen den Mikrofonen i und j:

$$\Delta T_{ij} = \arg \max_{\tau} R^{(e)}_{ij}(\tau)$$

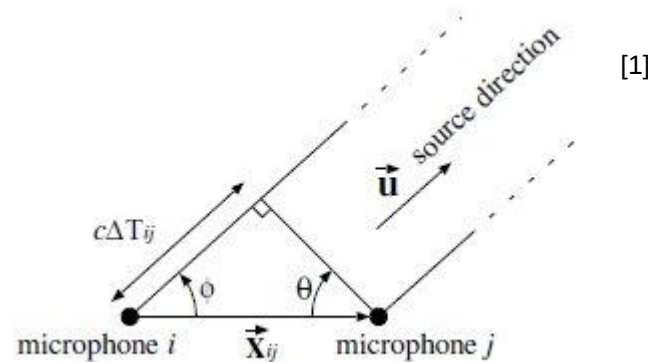
- N Mikrofone $\rightarrow$ $N(N-1)/2$ Verzögerungen
- N-1 unabhängig davon unabhängig
 $\rightarrow$ Verwendung der Verzögerungen ΔT_{1i} (ΔT_{12} bis ΔT_{18})
- Berechnung der restlichen Verzögerungen möglich mit:

$$\Delta T_{ij} = \Delta T_{1j} - \Delta T_{1i}$$

Positions-Abschätzung (1)



- Verwendetes Model:



- **Gegeben:** - $\mathbf{x}_{ij}$ = Vektor von Mikrofon i zu Mikrofon j
- ΔT_{ij} = TDOA
- **Gesucht:** $\mathbf{u}$ = Einheitsvektor in Richtung der Geräuschquelle

$$\cos \phi = \frac{\vec{u} \cdot \vec{x}_{ij}}{\|\vec{u}\| \|\vec{x}_{ij}\|} = \frac{\vec{u} \cdot \vec{x}_{ij}}{\|\vec{x}_{ij}\|}$$

$$\cos \phi = \sin \theta = \frac{c \Delta T_{ij}}{\|\vec{x}_{ij}\|}$$

$$\vec{u} \cdot \vec{x}_{ij} = c \Delta T_{ij}$$

Positions-Abschätzung (2)



- Gefundene Gleichung: $\vec{u} \cdot \vec{x}_{ij} = c\Delta T_{ij}$
- Mit $\mathbf{u} = (u, v, w)$ und $\mathbf{x}_{ij} = (x_j - x_i, y_j - y_i, z_j - z_i)$ (Position Mikrofon i (x_i, y_i, z_i)) folgt:

$$u(x_j - x_i) + v(y_j - y_i) + w(z_j - z_i) = c\Delta T_{ij}$$

- Für N Mikrofone ergibt sich so, folgende Matrix:

$$\begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) & (z_2 - z_1) \\ \vdots & \vdots & \vdots \\ (x_N - x_1) & (y_N - y_1) & (z_N - z_1) \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} c\Delta T_{12} \\ \vdots \\ c\Delta T_{1N} \end{bmatrix}$$

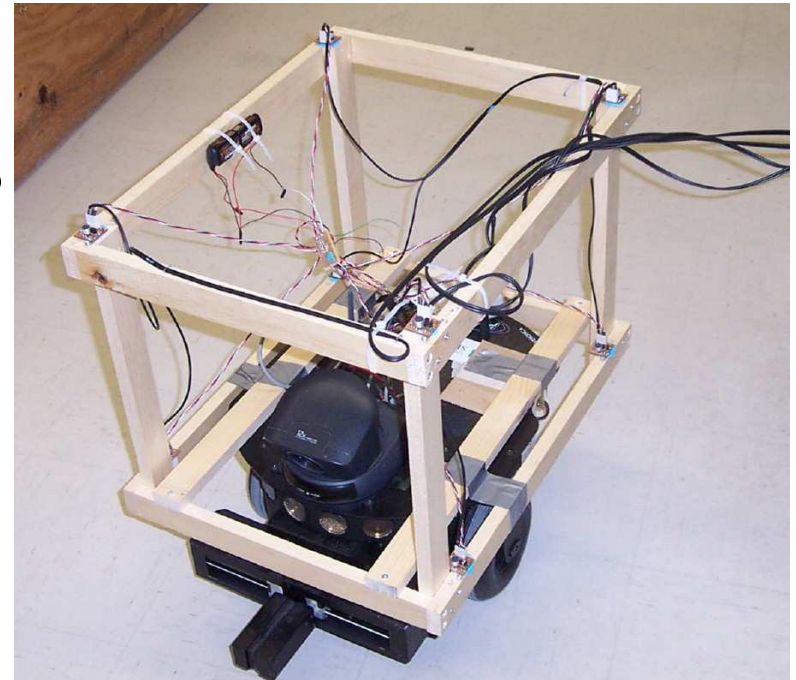
- Ab 4 Mikrofonen ist das System überbestimmt
→ Lösung kann gefunden werden, mittels Pseudoinverse

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$$

Versuchsaufbau (1)



- Acht Mikrofonen in den Ecken eines Rechtecks
- Montiert auf einem Activ Pioneer 2 Roboter
- Echtzeitberechnungen auf Athlon XP 2000+ Desktop PC
- Drehung des Roboters und vertikale Positionierung der Kamera mit Hilfe der berechneten Position
- Messungen mit fester Höhe/Distanz und variablem Winkel



[1]

Experimentelle Ergebnisse



Distanz, Höhe	Winkelfehler
3 m, -7°	$1,7^\circ$
3 m, 8°	$3,0^\circ$
1,5 m, -13°	$3,1^\circ$
0,9 m, 24°	$3,3^\circ$

- Fehler hauptsächlich durch Höhe $\rightarrow$ Vorteil
- Detektion ohne Drehung $\rightarrow$ Erkennung von kurzen Geräuschen
- Beste Funktion zwischen 3 m und 5 m Entfernung
- Tonlaute, aufgrund kleiner Bandbreite schwer detektierbar
- Geräusche mit großer Bandbreite (z.B. Papier rascheln) sehr gut detektierbar
- Erkennung von Stimmen normalerweise innerhalb der ersten zwei Silben

Vorteile des Systems mit acht Mikrofonen:

- Dreidimensionale Ortung
- Nur kurze Geräusche und keine Rauschunterdrückung nötig
- Genauigkeit: 3° auf 3 Meter Entfernung
- Robuste TDOA-Abschätzung
- Einsatz einfacher, günstiger Mikrofone

Weiterhin offene Problemstellungen:

- Keine Entfernungsabschätzung möglich
- Bei zwei gleichzeitigen Geräuschen wird nur lauterer wahrgenommen
- Hoher Rechenaufwand

Vielen Dank für die Aufmerksamkeit!



■ Quellen:

- [1] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau. Robust Sound Source Localization Using a Microphone Array on a Mobile Robot. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, pages 1228-1233, 2003
- [2] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface – a robot who communicate with multi-user. In *Proceedings EUROSPEECH*, pages 1732-1726, 1999.
- [3] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano. Active audition system and humanoid exterior design. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, pages 193-196, 2002.
- [4] <http://www.nhonnef.de/pics/spektrum.jpg> (08.07.2012)
- [5] K. Finger. Untersuchungen zur Kraftanregung durch die Verbrennung beim direkteinspritzenden Common-Rail Dieselmotor unter Berücksichtigung des Körperschallübertragungsverhaltens[Dissertation] .TU Darmstadt , pages 23 -26, 2001
- [6] L. Reindl. Skript – Messtechnische Grundlagen und Verfahren. Albert-Ludwigs-Universität Freiburg, Lehrstuhl für Elektrische Mess- Und Prüfverfahren, 2012
- [7] <http://de.wikipedia.org/wiki/Kreuzleistungsspektrum> (09.07.2012)
- [8] <http://de.wikipedia.org/wiki/Pseudoinverse> (09.07.2012)