Vergleich akustischer Quellenortungsverfahren für Sprachsignale in Räumen

Thomas Fehér[1], Matthias Lippmann[2]

[1] *Institut für Akustik und Sprachkommunikation, 01062 Dresden, Deutschland, Email: thomas.feher@ias.et.tu-dresden.de*

[2] *Institut für Akustik und Sprachkommunikation, 01062 Dresden, Deutschland, Email: matthias_lippmann@web.de*

Einleitung

In diesem Beitrag werden drei verschiedene akustische Quellenortungsverfahren für mehrkanalige Mikrofonanordnungen in realen Räumen miteinander verglichen. Die verwendeten Verfahren sind Generalized-Cross-Correlation (GCC), Steered Response Power (SRP) und Multiple Signal Classification (MUSIC). Es werden verschiedene Szenarien simuliert und im Anschluss ein Vergleich der Algorithmen mit realen Signalen vorgenommen. Die Anordnung der Mikrofone ist bei Simulation und Messung gleich, um eine gute Vergleichbarkeit zu gewährleisten. Als Quellsignale kommen aufgezeichnete Sprachsignale zum Einsatz. Für die Messung wurden die selben Sprachsignale mit gemessenen Raumimpulsantworten gefaltet.

Die Algorithmen

Generalized Cross Correlation (GCC)

Bei diesem Verfahren wird mit Hilfe der Laufzeitdifferenzen τ_{12} eines Signals $x(t)$ zwischen den verschiedenen Mikrofonpaaren eines Arrays der Ort der zugehörigen Schallquelle besimmt. Um τ_{12} so genau wie möglich bestimmen zu können, wird die Kreuzkorrelationsfunktion (Cross Correlation)

$$s_{12}(\tau) = \int_{-\infty}^{+\infty} x_1(t)x_2(t+\tau)dt \tag{1}$$

verwendet, deren Maximum den jeweiligen Laufzeitunterschied des Mikrofonpaares anzeigt. Unter Einbeziehung der Koordinaten der Mikrofone können damit die Parameter a und b für Hyperbelgraphen (Gleichung 2) berechnet werden. Die Schallquelle befindet sich am Schnittpunkt der Hyperbeln aller Mikrofonpaare.

$$\frac{x^2}{a} + \frac{y^2}{b} = 1 \tag{2}$$

Steered Response Power (SRP)

Beim SRP-Algorithmus [1] handelt es sich um ein delay and sum Beamformingverfahren. Dabei werden den Mikrofonen des Arrays Verzögerungsglieder Δ_n nachgeschalten (delay), mit deren Hifle ein virtueller Hohlspiegel erzeugt werden kann, dessen Fokus sich auf jeden beliebigen Punkt $\boldsymbol{q}$ im Raum lenken lässt. Im Anschluss werden die Signale aller n Mikrofone aufsummiert.(sum)

$$y(t, \boldsymbol{q}) = \sum_{n=1}^{N} x_n(t + \Delta_n) \tag{3}$$

Durch eine Transformation von $y(t, \boldsymbol{q})$ in den Frequenzbereich und der Integration über diesen, erhält die Gleichung:

$$P(\boldsymbol{q}) = \sum_{n=1}^{N} \int_{-\infty}^{+\infty} \left| X_n(\omega) e^{j\omega\Delta_n} \right|^2 d\omega \tag{4}$$

Da $P(\boldsymbol{q})$ proportional zur abgestrahlten Leistung am jeweiligen Ort $\boldsymbol{q}$ ist, kann ein Quellort durch eine Extremstellenbestimmung gefunden werden.

PHAse-Transform-Wichtung (PHAT)

Um den Einfluss von Störungen, wie zB. Reflexionen, auf das Ergebnis der Quellenortung zu minimieren, wird dem SRP-Algorithmus ein Wichtungsfaktor hinzugefügt.[2] Bei dem daraus resultierenden SRP-PHAT-Verfahren wird eine Normierung des Amplitudenkreuzspektrums vorgenommen, wobei die für die Lokalisierung wichtigen Phaseninformationen erhalten bleiben. Aus Gleichung(4) wird in diesem Fall:

$$P(\boldsymbol{q}) = \sum_{k=1}^{N} \sum_{l=1}^{N} \int_{-\infty}^{+\infty} \frac{X_k(\omega) X_l^*(\omega)}{|X_k(\omega) X_l^*(\omega)|} e^{j\omega(\Delta k - \Delta l)} d\omega \tag{5}$$

Multiple Signal Classification (MUSIC)

MUSIC [3] ist ein Verfahren bei dem der Ort einer Schallquelle über eine Signal-Noise-Subspace-Analyse der aufgenommenen Mikrofonsignale bestimmt wird. Für jeden Abtastpunkt im Raum wird dafür die Matrix der Kovarianzen erstellt, welche anschließend einer Eigenwertanalyse unterzogen wird. Die dabei errechneten Eigenvektoren können mit Hilfe der zugehörigen Eigenwerte in Signal- und Rauscheigenvektoren getrennt werden, die Signal- und Rauschunterräume (subspaces) bilden. Bei MUSIC wird davon ausgegangen, das die Energie des Rauschens an dem Ort am geringsten ist, an dem eine Quelle befindt, da an dieser Stelle das Rauschen vom Signal im jeweiligen Frequenzbereich überdeckt wird. Wenn $a(\boldsymbol{q})$ der Steeringvektor des Arrays und $\boldsymbol{V}_N$ die Matrix der jeweiligen Noiseeigenvektoren ist, kann der Pegel nach MUSIC $M(\boldsymbol{q})$ mit folgender Gleichung berechnet werden:

$$M(\boldsymbol{q}) = \frac{1}{\mathbf{a}^*(\boldsymbol{q}) \mathbf{V}_N \mathbf{V}_N^* \mathbf{a}(\boldsymbol{q})} \tag{6}$$

Befindet sich am Ort $\boldsymbol{q}$ eine Quelle, so hat $M(\boldsymbol{q})$ an dieser Stelle ein Maximum.

Vergleichende Untersuchung der Algorithmen

Um die drei verschiedenen Quellenortungsverfahren vergleichen zu können, wurden zunächst verschiedene Szenarien mit ein und zwei Quellen simuliert und anschließend mit den Messungen realer Sprachaufnahmen verglichen.

Messaufbau

Um Sprachaufnahmen einer realen Umgebung zu erhalten wurden 8 Mikrofone nah der Wände im Audiostudio des Institutes (Abb.: 1) aufgestellt. Der Raum hat eine Grundfläche von ca. $30\,m^2$, ein Volumen von ca. $90\,m^3$ und eine Nachhallzeit von ca. $0,2\,s$. Um stehende Wellen im Raum zu vermeiden, sind die jeweils gegenüberliegenden Wände nicht parallel, sondern leicht schräg zueinander angeordnet. Im Raum wurden 9 verschiedene Quellenpositionen definiert. Als Quellsignal diente ein logarithmischer Sinus-Sweep aus dem danach die Raumimpulsantworten berechnet werden konnten. Diese wurden mit den in der Simulation eingestetzen Sprachsignalen gefaltet, um Messergebnisse zu erhalten die mit denen der Simulation gut vergleichbar sind.

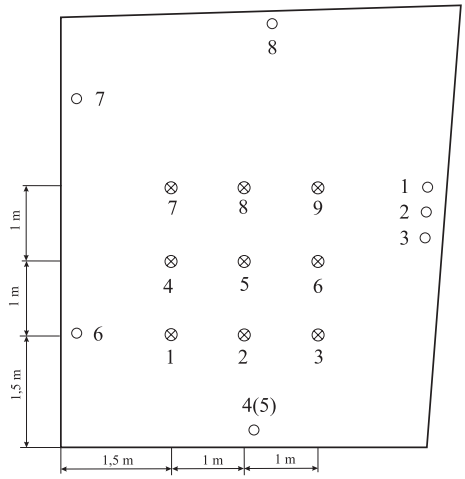

Abbildung 1: Messumgebung für die Aufnahme der Impulsantworten

Ergebnisse aus Simulation und Messung

Da es sich bei GCC um ein rein geometrisches Verfahren zur Quellenlokalisierung handelt, ist ein numerischer Vergleich mit SRP und MUSIC nicht möglich. Dennoch kann festgestellt werden, dass SRP und MUSIC gegenüber Reflexionen robuster und für die Ortung mehrerer Quellen deutlich besser geeignet sind als das GCC-Hyperbelverfahren. Da für die Berechnung der Laufzeitunterschiede das Maximum der Kreuzkorrelation ausschlaggebend ist, wird mit dieser Methode meist nur die lauteste der Quellen geortet. Es kann auch vorkommen, wenn sich die verschiedenen Quellen nah an verschiedenen Empfängern befinden, dass nur noch Phantomquellen geortet werden, bzw. die Schnittpunkte der Hyperbeln über den gesamten Raum verteilt sind. Um SRP mit MUSIC zu vergleichen wurden zwei Größen eingeführt. Zum einen das Durchschnittsniveau der Pegel an allen Orten im Verhältniss zum Maximum D und die Welligkeit der Nebenstrukturen w.

$$D = \frac{\overline{P(\boldsymbol{q})}}{max(P(\boldsymbol{q}))} \tag{7}$$

$$w = \overline{|D - P(\boldsymbol{q})|} \tag{8}$$

Damit konnte nachgewiesen werden, dass mit MUSIC die Orte der Quellen in den meisten Situationen deutlicher zu erkennen waren als bei SRP und SRP-PHAT und demzufolge für die Lokalisierung von Sprechern in Räumen zu empfehlen ist.

Anz. Quellen	Agorithmus	D in %	w in %
1	SRP-PHAT	30,6	4,7
	MUSIC	24,9	1,7
2	SRP-PHAT	37,2	5,5
	MUSIC	34,1	2,3

Tabelle 1: Vergleich der Simulationsergebnisse von SRP-PHAT und MUSIC im Freifeld

Anz. Quellen	Agorithmus	D in %	w in %
1	SRP-PHAT	55,6	12,6
	MUSIC	56,2	1,7
2	SRP-PHAT	64,9	10,3
	MUSIC	90,5	1,7

Tabelle 2: Vergleich der Messergebnisse von SRP-PHAT und MUSIC

Literatur

[1] Michael Brandstein, D. W.: Microphone Arrays. Springer-Verlag, 2001. – ISBN 3–540–41953–5

[2] Knapp, C. ; Carter, G. : The generalized correlation method for estimation of delay. In: IEEE Trans. Acuost. Speech Signal Process., vol. ASSP-24 pp.320-327 (August 1976)

[3] Schmidt, R. O.: Multiple Emitter Location and Signal Parameter Estimation. In: IEEE Transaction on Antennas and Propagation, Vol.AP-34, No.3 (March 1986)