Erfassung von Sprache im KFZ mit Störgeräuschreduktion im Nahfeld eines Mikrofon-Sparse-Arrays

Patrick Vicinus[1], Peter Kitzenmaier[1], Reinhold Orglmeister[2]

[1] *PEIKER acustic GmbH & Co. KG, 61381 Friedrichsdorf / Ts., Deutschland, patrick.vicinus@peiker.de*

[2] *Fachgebiet Elektronik und medizinische Signalverarbeitung, TU Berlin, 10587 Berlin, Deutschland*

Einleitung

Die Sprachsignalerfassung in einem KFZ ist von immer größerem Interesse. Neben der Freisprechtelefonie spielt dabei auch die sprachgestützte Bedienung von Geräten im Fahrzeug eine zunehmende Rolle. Sollen diese Anwendungen von allen Fahrzeuginsassen genutzt werden können, ergeben sich erhöhte Anforderungen für die Signalerfassung. Aufgrund der Direktivität der Sprache in Blickrichtung sowie der Abnahme der Schallintensität mit zunehmendem Abstand vom Mund und einem annähernd diffusen Störgeräusch innerhalb der Fahrzeugkabine ist für eine gute Erfassung aller Sprach-Quellensignale die Platzierung von mindestens einem Mikrofon in der Nähe und in ungefährer Blickrichtung jedes potentiellen Sprechers angebracht. In vielen Fällen kann die entstehende Mikrofonanordnung als sogenanntes Sparse-Array betrachtet werden, in dessen Nahfeld sich die Insassen befinden.

Für die Sprachsignalerfassung eines Sprechers, der sich an vorgegebener Position im Nahfeld eines Sparse-Arrays befindet, kann auf einen Nahfeld-Beamformerentwurf zurückgegriffen werden. Auch, wenn die Abstände zweier Mikrofone n und m bei einem Sparse-Array über der halben minimalen Wellenlänge des zu berücksichtigenden Frequenzbereiches 100...8000Hz liegen können und daher das Beampattern durch starke Nebenmaxima und reduzierte Direktivität geprägt ist [1], gelingt bei einer geeigneten Länge der Filter des Beamformers ein Null-Beamforming auf einen oder mehrere nicht stationäre, lokalisierbare Störer (Interferer). Zur Vermeidung einer Beeinträchtigung des Sprachsignals aufgrund von Kopfbewegungen wird die Fokusierung auf einen angemessenen Bereich erweitert [4].

In einer ersten Stufe wird das Störgeräusch des zu erfassenden Sprachsignals eines jeden Sprechers über unabhängige robuste Nahfeld-Beamformer reduziert. In der zweiten Stufe erfolgt die einkanalige Zusammenfassung der gewünschten Sprechersignale im Sinne einer SNR-Maximierung. Dies ermöglicht die Erfassung der Sprachsignale mehrerer Sprecher unter Berücksichtigung variierender Kopfpositionen und Halleinfluss, die Unterdrückung lokalisierbarer Störer (z.B. Störsprecher) und die Reduzierung diffuser Störung. Erfolgt die Schätzung der Letzteren am Ausgang der zweiten Stufe (z.B. über ein kohärenzbasiertes Verfahren [3]), ist das vorgestellte Verfahren eine optimale Vorverarbeitung für einen nachgeschalteten einkanaligen Bayes-Schätzer für die Sprachsignale [3].

Signalmodell

Es existieren D mögliche Sprecherpositionen. Die Impulsantwort des Sprechers d zu den N Mikrofonen in Form von FIR-Filtern der Länge L_h seien im *Array Mainfold Vector* $\mathbf{h}_d \in \mathbb{C}^{L_h N \times 1}$ konkartiniert und über

$$\mathbf{h}_d = \left[\mathbf{h}_0^T \dots \mathbf{h}_{N-1}^T\right]^T \text{ mit } \mathbf{h}_n = \left[h_0^{(d,n)}\, h_1^{(d,n)} \dots h_{L_h-1}^{(d,n)}\right]^T \quad (1)$$

beschrieben. Die idealisierte normierte Signalkovarianz wird über $\mathbf{Q}_d = \mathbf{h}_d \mathbf{h}_d^H$ berechnet. Die Koeffizienten $\mathbf{c}_d \in \mathbb{C}^{L_c N \times 1}$ des Beamformers, der auf die Position d fokusiert, seien über N FIR-Filter der Länge L_c festgelegt. Werden die zurückliegenden L_c Abtastwerte zum Zeitpunkt K in einem Vektor

$$\mathbf{x} = \left[\mathbf{x}_0^T \dots \mathbf{x}_{N-1}^T\right]^T \text{ mit } \mathbf{x}_n = \left[x_{L_c-1}^{(n)}\, x_{L_c-2}^{(n)} \dots x_0^{(n)}\right]^T \quad (2)$$

abgelegt, berechnet sich zum Zeitpunkt K das Beamformerausgangssignal zu $y_d(K) = \mathbf{c}_d^H \mathbf{x}$. Mit Hilfe des Erwartungswert-Operators $\mathrm{E}\{\cdot\}$ wird die Eingagskovarianzmatrix zu $\mathbf{R}_x = \mathrm{E}\left\{\mathbf{x}\mathbf{x}^H\right\}$ berechnet.

Ferner definiert man eine Vektor-zu-Matrix Operation $\mathbf{P}(\mathbf{a}, L_b)$, die einen Vektor $\mathbf{a}$ der Länge $N \cdot L_a$ in eine Matrix der Dimension $(N \cdot L_b) \times (L_a + L_b - 1)$ transformiert, zwei Vektoren $\mathbf{a} = [\mathbf{a}_0^T \dots \mathbf{a}_{N-1}^T]^T$ mit $\mathbf{a}_n \in \mathbb{C}^{L_a \times 1}$ und $\mathbf{b} = [\mathbf{b}_0^T \dots \mathbf{b}_{N-1}^T]^T$ mit $\mathbf{b}_n \in \mathbb{C}^{L_b \times 1}$, sowie den Faltungsoperator '$*$', sodass die Beziehung

$$\mathbf{b}^H \mathbf{P}(\mathbf{a}, L_b) = \sum_{n=0}^{N-1} [\mathbf{a}_n * \mathbf{b}_n^*] \quad (3)$$

gilt. Die Beeinflussung des Vorzugssignals durch Raumübertragung und den Beamformer wird über $\mathbf{g}^H = \mathbf{c}^H P(\mathbf{h}, L_c)$ zusammengefasst. Ein weiterer Operator $\mathcal{P}(\mathbf{R}, M)$ berechnet die Eigenvektoren der quadratischen Matrix $\mathbf{R}$ der Dimesion $L_R \times L_R$ korrespondierend zu den höchsten M Eigenwerten und gibt diese in einer Matrix der Dimension $L_R \times M$ zurück.

Robustes Nahfeld-Beamforming

Zunächst werden die Beamformer Kostenfunktion und eine energiebasierte (quadratische) robuste *Distortionless*-Bedingung definiert, die mögliche Schwankungen der Sprecher-Position modelliert:

$$\underset{\mathbf{c}_d}{\operatorname{argmin}}\ \mathrm{E}\left\{|\mathbf{c}_d^H \mathbf{x}|^2\right\} \quad \text{u.d.Bed. } \mathbf{c}_d^H \tilde{\mathbf{Q}}_d \mathbf{c}_d = 1 \quad . \quad (4)$$

Die Schätzung der Signalkovarianz $\tilde{\mathbf{Q}}_d$ berücksichtigt mögliche Kopfbewegungen des $d$ten Sprechers. Im Nahfeld eines Sparse-Arrays können diese Variationen nicht über einen Bereich möglicher Einfallsrichtungen beschrieben werden, geeigneter ist die Angabe einer maximalen

Abweichung ϵ_d zur idealen Signalkovarianz $\mathbf{Q}_d$ über die Frobenius-Norm. Dies führt zur in [4] hergeleiteten Beziehung $\tilde{\mathbf{Q}}_d = \mathbf{Q}_d - \epsilon_d \mathbf{I}$ (negative diagonal loading). Die Adaption der Beamformerkoeffizienten erfolgt entweder nach einem *spatial prewhitening* der Störung über die Abbildungsmatrix $\mathbf{W}_d = \tilde{\mathbf{Q}}_d^{-0.5}$ (sie transformiert die Nebenbedingung in die Form $\mathbf{c}_d^H \mathbf{W}_d \mathbf{Q} \mathbf{W}_d^H \mathbf{c}_d = \tilde{\mathbf{c}}_d^H \tilde{\mathbf{c}}_d = 1$) über den *scaled projection*-Algorithmus aus [2] oder im Falle der Existenz negativer Eigenwerte von $\tilde{\mathbf{Q}}$ über die Adaption des dominierenden Eigenvektors $\mathcal{P}(\mathbf{R}_x^{-1}\tilde{\mathbf{Q}}, 1)$. Zusätzlich zu dieser quadratischen Bedingung, wird über lineare Nebenbedingungen die Filterwirkung $\mathbf{g}_d^H$ in Form eines Allpasses gehalten:

$$\tilde{\mathbf{c}}_d^H \mathbf{P}(\mathbf{W}_d \mathbf{h}_d, L_c) = [0 \ \dots \ a_d \ 0 \ \dots] = \mathbf{g}_d^H \quad . \qquad (5)$$

Der Parameter a_d wird über die quadratische Nebenbedingung festgelegt und muss nicht in die linearen Bedingungen einbezogen werden. Die adaptive Implementierung von (4) unter Wahrung der $L_h + L_c - 2$ linearen Null-Bedingungen von (5) erfolgt in der Struktur eines *Generalized Sidelobe Cancellers*. Die D voneinander unabhängig berechneten Beamformerkoeffizienten $\mathbf{c}_d$ werden in der Matrix $\mathbf{C} \in \mathbb{C}^{L_c \times D}$ zusammengefasst.

Eine Abschätzung der notwendigen Filterlänge erfolgt über die maximale Anzahl J der zu unterdrückenden lokalisierbaren Störer. Erreicht der $j$te Störer die Mikrofone über die Raumimpulsantworten $\mathbf{f}_j$, führt dies über $\mathbf{c}_d^H \mathbf{Z} = [\mathbf{g}_d^H \ \mathbf{0} \ \dots \ \mathbf{0}]$ zu

$$\mathbf{c}_d^H = [\mathbf{g}_d^H \ \mathbf{0} \ \dots \ \mathbf{0}] \mathbf{R} \mathbf{Z}^H \left(\mathbf{Z}^H \mathbf{R} \mathbf{Z}\right)^{-1}$$
$$\text{mit } \mathbf{Z} = [\mathbf{P}(\mathbf{h}_d, L_c) \ \mathbf{P}(\mathbf{f}_1, L_c) \ \dots \ \mathbf{P}(\mathbf{f}_J, L_c)] \ , \qquad (6)$$

wobei $\mathbf{0}$ ein Nullvektor der Länge $(L_c + L_h - 1)$ und $\mathbf{R} \in \mathbb{C}^{NL_c \times NL_c}$ eine beliebige invertierbare Matrix ist. Die Matrix $\mathbf{Z}^H \mathbf{R} \mathbf{Z}$ ist nur dann ínvertierbar, wenn $NL_c \geq (J+1)(L_c + L_h - 1)$ ist, also eine Beamformer-Filterlänge von $L_c \geq \frac{(J+1)(L_h-1)}{N-J-1}$ verwendet wird.

Principal Component Beamforming

Die Aufgabe der zweiten Stufe, ist die Bestimmung eines nachgeschalteten Nahfeld-Beamformers, der die Ausgangskanäle der ersten Stufe SNR-maximierend einkanlig zusammenführt. Er ermöglicht die Fokusierung auf den/die aktiven Sprecher und die Minimierung der über ihre Kovarianzmatrix $\tilde{\mathbf{R}}_n = \mathbf{C}^H \mathbf{R}_n \mathbf{C}$ charakterisierten Störung, wobei die Störkovarianz $\mathbf{R}_n$ die Störung am Eingang der ersten Stufe charakterisiert. Da für jede der D Sprecherpositionen in der ersten Stufe bereits ein SNR-maximierendes FIR-Filter bestimmt wurde, kann in dieser Stufe das Filter auf eine Länge von eins reduziert und auf zusätzliche Allpass-Bedingungen verzichtet werden. Die optimalen Koeffizienten $\mathbf{w}$ der zweiten Stufe berechnen sich dann über [5]

$$\mathbf{w} = k\mathbf{u} \text{ mit } \mathbf{u} = \mathcal{P}\left(\tilde{\mathbf{R}}_n^{-1}\tilde{\mathbf{R}}_x, 1\right), \ k = \frac{\max\left(|\tilde{\mathbf{R}}_n \mathbf{u}|\right)}{\mathbf{u}^H \tilde{\mathbf{R}}_n \mathbf{u}}, \quad (7)$$

wobei sich $\tilde{\mathbf{R}}_x$ über $\mathbf{C}^H \mathbf{R}_x \mathbf{C}$ berechnet. Der dominierende Eigenvektor (*principal component*) kann über ein Gradientenverfahren [6] adaptiert werden. Zur Schätzung der Störkovarianz $\mathbf{R}_n$ kann neben einer im Vergleich zur Sprache gegebenen größeren Langzeitstationarität vor allem die Zeitinvarianz der Kohärenz ausgenutzt werden [3]. Die Störgeräuschreferenz $\sigma_n^2 = \mathbf{w}^H \mathbf{C}^H \mathbf{R}_n \mathbf{C} \mathbf{w}$ und die Ausgangsvarianz der zweiten Stufe sind bei Annahme gaußverteilter Eingangsprozesse optimale Eingangsparameter für einen nachgeschalteten einkanaligen Bayes-Schätzer für das Sprachsignal [3].

Ergebnisse

Setzt man die Konvergenz der adaptiven Filter voraus und lässt die einkanalige Nachverabeitung außen vor, beschreibt das Verfahren einen Nahfeld-MPDR-Beamformer (*Minimum Power Distortionless Response*), dessen Störgeräuschreduktion maßgeblich durch die erste Stufe bestimmt wird. Aufgrund der großen Mikrofonabstände führt das stationäre diffuse Störgeräusch im interessierenden Frequenzbereich zu einem nahezu inkohärenten Störfeld. Geht man von einem 10dB besseren SNR des dem Sprecher nächstliegendem Mikrofons aus, führt dies bei vier Mikrofonen zu einer optimistischen Störgeräuschreduktion von nur etwa $10\log_{10}(1 + 0.1 + 0.1 + 0.1) = 1.1dB$. Das Sparse-Array ermöglicht jedoch selbst mit Hall-Einfluss ein Null-Beamforming auf bis zu $N - 2$ lokalisierbare Störer, entsprechende Länge der FIR-Filter vorrausgesetzt. Vergleicht man die Ausgangsstörleistung der zweiten Stufe mit einem Verfahren, das lediglich die Ausgangssignale der ersten Stufe addiert (idealerweise ist nur in einem der Eingangskanäle ein Sprachsignal vorhanden), führt dies bei einer vierkanaligen Variante bei der Annahme von inkohärentem Rauschen zu einer Störgeräuschreduktion von 6dB für jeden Sprecher. Im Vergleich zu einer einkanaligen Variante reduziert sich im *worst case* (also keine lokalisierbaren Störer) der Gewinn für den nächstliegenden Sprecher auf vernachlässigbare 1.1dB, führt jedoch für die entfernten Sprecher zu einem Gewinn von 10dB. Aufgrund der niedrigen Diffusschall-Unterdrückung wird die Verwendung von unidirektionalen Mikrofonkapseln bzw. die Bildung von Array-Untergruppen kleiner Abmessungen im Verbindung mit Sparse-Arrays empfohlen.

Literatur

[1] T. D. Abhayapla, R. A. Kennedy, R. C. Williamson. Spatial Aliasing for Near-Field Sensor Arrays. In *Electronics Letters*, 35(10), Mai 1999.

[2] H. Cox, R. M. Zeskind, M. M. Owen. Robust Adaptive Beamforming. In *IEEE ASSP*, pages 1365-1376,

[3] S. Lefkimmiatis, P. Maragos. A Generalized Estimation Approach for Linear and Nonlinear Microphone Array Post-filters. In *Speech Communication 49*, Elsevier, 2007.

[4] S. Shahbazpanahi, A. B. Gershman, Z. Q. Luo, K. M. Wong. Robust Adaptive Beamforming for General-Rank Signal Models. In *IEEE Transactions on Signal Processing*, 51(9), September 2003.

[5] P. Vicinus, W. Baumann. Verteilte Mikrofone im Kraftfahrzeug. In *DEGA DAGA, 32. Jahrestagung für Akustik*, März 2006.

[6] E. Warsitz, R. Haeb-Umbach, D. H. Tran Vu. Blind Adaptive Principal Eigenvector Beamforming for Acoustical Source Separation. In *Interspeech, Antwerp, Belgium*, August 2007.