Mehrkanalige Sprachsignalverarbeitung durch adaptives Eigenbeamforming für Freisprecheinrichtungen im Kraftfahrzeug

Ernst Warsitz, Reinhold Häb-Umbach

Universität Paderborn, Inst. für Elektrotechnik und Informationstechnik, Fachgebiet Nachrichtentechnik, 33098 Paderborn

Email: {warsitz,haeb}@nt.uni-paderborn.de

Einleitung

Die Verwendung von mehrkanaligen Ansätzen zur Sprachsignalverarbeitung bei Freisprecheinrichtungen bietet im Vergleich zu einkanaligen Verfahren den Vorteil, das Sprachsignal idealerweise unverfälscht zu belassen. Die wohl bekanntesten Verfahren sind hierbei der Delay-and-Sum Beamformer (DSB), dessen Erweiterung zu einem Generalized Sidelobe Canceller (GSC) nach Griffith-Jim [1] und die Weiterentwicklung nach Hoshuyama [2]. Problematisch bei diesen Verfahren ist jedoch, dass sie eine Schätzung für die Sprecherrichtung benötigen. Außerdem ist im Falle von so genannten diffusen Hintergrundrauschen keine signifikante Rauschunterdrückung speziell im unteren Frequenzbereich zu erzielen [3].

Theoretisch läßt sich zeigen, dass ein diffuses Schallfeld durch unendlich viele unabhängige Schallquellen verursacht wird. Der Schall fällt dabei aus allen Raumrichtungen mit gleicher Intensität ein, und die räumliche Korrelationseigenschaft (Kohärenz) weist dadurch einen speziellen Charakter auf. Das Geräuschfeld in der Fahrgastzelle eines Kraftfahrzeugs (Kfz) ist ebenfalls auf unterschiedliche Schallquellen und Übertragungswege zurückzuführen. Maßgeblich hierbei sind Motor-, Reifenabroll- und Windgeräusche die über die Luft, aber auch als Körperschall übertragen werden. Daher läßt sich das Geräuschfeld innerhalb eines Kfz gut durch ein diffuses Schallfeld beschreiben.

Eigenbeamforming

Es soll hier ein Beamformingverfahren beschrieben werden, welches auf der Lösung eines verallgemeinerten Eigenwertproblems (Generalized Eigenvalue, GEV) basiert. Dieser Beamformer benötigt keine explizite Sprecherrichtungsbestimmung und ermöglicht auch bei niedrigen Frequenzen in einem diffusen Schallfeld eine Störgeräuschreduzierung.

Das i-te Mikrofonsignal $X_i(k)$ soll im Frequenzbereich für jede Spektralkomponente k aus der Überlagerung des Sprachsignals $S_i(k)$ und dem Rauschen $N_i(k)$ bestehen: $X_i(k) = S_i(k) + N_i(k), i = 1,..,M$. Die M gefilterten Signale ergeben sich dann am Beamformer-Ausgang zu

$$Y(k) = \sum_{i=1}^{M} F_i^*(k) \cdot X_i(k) \; = \mathbf{F}^H(k) \cdot \mathbf{X}(k), \qquad (1)$$

wobei $\mathbf{X}(k) = (X_1(k),..,X_M(k))^T$ und die Filterkoeffizienten $\mathbf{F}(k) = (F_1(k),..,F_M(k))^T$ in Vektornotation dargestellt sind; $(\cdot)^T$ bedeutet transponiert und $(\cdot)^H$ hermitisch transponiert. Da Sprach- und Störsignal als unkorreliert zueinander anzunehmen sind, kann das Leistungsdichtespektrum (LDS) am Beamformer-Ausgang wie folgt angegeben werden:

$$\Phi_{YY}(k) = \mathbf{F}^H(k)\mathbf{\Phi_{SS}}(k)\mathbf{F}(k) + \mathbf{F}^H(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}(k). \qquad (2)$$

Die Maximierung des frequenzabhängigen Signal-zu-Rauschverhältnisses

$$\mathrm{SNR}(k) = \frac{\mathbf{F}^H(k)\mathbf{\Phi_{XX}}(k)\mathbf{F}(k)}{\mathbf{F}^H(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}(k)} - 1 \qquad (3)$$

führt schließlich zu einer verallgemeinerten Eigenwertzerlegung bezüglich $\mathbf{\Phi_{XX}}(k)$ (LDS von Sprache-plus-Rauschen) und $\mathbf{\Phi_{NN}}(k)$ (LDS des reinen Rauschsignals). Dabei ergeben sich die optimalen Filterkoeffizienten gerade durch den Eigenvektor korrespondierend zu dem größten Eigenwert, welcher die obere Grenze des in (3) geschriebenen Rayleigh Quotienten angibt [4].

GEV-Adaptionsalgorithmus

Mit Hilfe des Gradientenanstiegsverfahrens und einer speziellen Anwendung der Methode nach Lagrange haben wir in [4] einen robusten Adaptionsalgorithmus zur Ermittlung und Verfolgung des gesuchten Filterkoeffizientenvektors entwickelt:

$$\begin{aligned}\mathbf{F}(m+1) = \mathbf{F}(m) &+ \frac{C - \mathbf{F}^H(m)\mathbf{\Phi_{NN}}\mathbf{F}(m)}{2\mathbf{F}^H(m)\mathbf{\Phi_{NN}}\mathbf{\Phi_{NN}}\mathbf{F}(m)}\mathbf{\Phi_{NN}}\mathbf{F}(m) \\ &+ \mu\left[\,\mathbf{\Phi_{XX}}\mathbf{F}(m) - \frac{\mathbf{F}^H(m)\mathbf{\Phi}^{(XN)}\mathbf{F}(m)}{2\mathbf{F}^H(m)\mathbf{\Phi_{NN}}\mathbf{\Phi_{NN}}\mathbf{F}(m)}\mathbf{\Phi_{NN}}\mathbf{F}(m)\right], \qquad (4)\end{aligned}$$

mit $\mathbf{\Phi}^{(XN)} = \mathbf{\Phi_{XX}}\mathbf{\Phi_{NN}} + \mathbf{\Phi_{NN}}\mathbf{\Phi_{XX}}$. Der Frequenzindex k wurde aufgrund der Übersichtlichkeit weggelassen und m beschreibt hier den Iterationsindex, welcher gleichzeitig den Blockindex der segmentweisen Signalverarbeitung darstellt. Über die Konstante C wird die Konvergenz gewährleistet und eine zu erzielende Störgeräuschdämpfung pro Frequenz vorgegeben, da sich die Störleistung im stationären Fall gerade zu $\mathbf{F}^H(m)\mathbf{\Phi_{NN}}\mathbf{F}(m) \stackrel{!}{=} C$ ergibt. Die benötigten Kovarianzmatrizen können rekursiv in Sprachpausen mit $\mathbf{\Phi_{NN}} \approx \hat{\mathbf{\Phi}}_{\mathbf{NN}}(m) = \alpha \cdot \hat{\mathbf{\Phi}}_{\mathbf{NN}}(m-1) + (1-\alpha) \cdot \mathbf{N}(m)\mathbf{N}^H(m)$ bzw. durch $\mathbf{\Phi_{XX}} \approx \hat{\mathbf{\Phi}}_{\mathbf{XX}}(m) = \beta \cdot \hat{\mathbf{\Phi}}_{\mathbf{XX}}(m-1) + (1-\beta) \cdot \mathbf{X}(m)\mathbf{X}^H(m)$ während Sprachaktivität geschätzt werden $(0 < \beta < \alpha < 1)$.

Experimentelle Ergebnisse

Die im folgenden beschriebenen Experimente basieren jeweils auf 4-kanaligen mit 16 kHz abgetasteten Signalen. Die 4 Mikrophone zur Aufzeichnung der Sprachsignale und Autofahrgeräusche wurden mittig auf der Armaturenkonsole eines BMW E46/2 äquidistant im Abstand von $d = 5$ cm angeordnet. So konnten jeweils von dem Fahrer und dem Beifahrer in verschiedenen Fahrsituationen bei 60, 80, 100 und 120 kmh Audiodaten aufgenommen werden. Bei allen Experimenten betrug die Filterlänge des GEV-Beamformers und der Leacky-Adaptive Blockingmatrix [2] jeweils 128 und bei den Adaptive Interference Cancelers 256. Die SNR-Messungen wurden für alle Verfahren jeweils nach Konvergenz der entsprechenden Filterkoeffizienten durchgeführt.

Frequenzabhängiges SNR

In einem ersten Experiment soll gezeigt werden, wie der theoretische SNR-Gewinn zwischen Ein- und Ausgang des Beamformers bei einem idealen diffusen Schallfeld für $M = 4$ ausfällt. Dazu wird die Kovarianzmatrix der Störung wie folgt

besetzt: $\{\mathbf{\Phi_{NN}}\}_{i,j} = \mathrm{si}(2\pi f_k d_{ij}/c)$ und mit einem Regularisierungsterm δ erweitert $\mathbf{\Phi_{NN}} := \mathbf{\Phi_{NN}} + \delta\mathbf{I}$, wobei $\mathbf{I}$ die Einheitsmatrix, f_k die diskrete Frequenz und c die Schallgeschwindigkeit definiert. Die Auswertung von (3) mit optimalen Filterkoeffizienten für ein unverhalltes Nutzsignal mit einem Einfallswinkel von 39° ist in Bild 1 oben dargestellt. In der unteren Abbildung ist zum Vergleich das Ergebnis für reale, mit 150 Hz hochpass gefilterte Aufnahmen im Kfz abgebildet. Dabei sind für den GEV-Beamformer in dünnen blauen Linien beispielhaft Einzelmessungen für unterschiedliche Fahrsituationen und in dicken Linien die gemittelten Verläufe für den GEV-Beamformer und den DSB aufgetragen. Anhand der Ergebnisse in Bild 1 und der Auswertung der Kohärenz der gemachten Aufnahmen kann gesagt werden, dass das Autofahrgeräusch bei unseren Aufnahmen kein ideales diffuses Schallfeld aufwies.

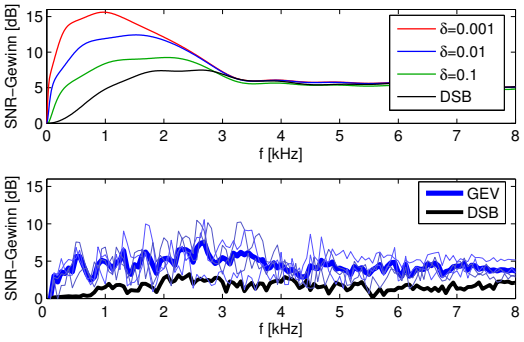

Abbildung 1: SNR-Gewinn für GEV-Beamformer und DSB: theoretischer Verlauf für diffuses Schallfeld (oben) und gemessener Verlauf in realer Fahrsituation (unten).

Synthetische Überlagerung

In einem weiteren Experiment wurde untersucht, ob der erzielbare SNR-Gewinn vom Eingangs-SNR abhängt. Dazu wurden reine Sprachsignale synthetisch mit Autofahrgeräuschen überlagert, wobei das Sprachsignal unverhallt mit einem Einfallswinkel von 29° auf das Array einfiel. In Bild 2 links ist das Ausgangs- über dem Eingangs-SNR aufgetragen. Hierbei bezeichnet 'Mik' das Mikrofonsignal, 'HP' das hochpass-gefilterte Signal, 'GJ' den Ausgang am Griffith-Jim Beamformer und 'Ho' den Ausgang am Beamformer nach Hoshuyama. Dabei zeigte sich eine leichte Tendenz einer Steigerung des SNR-Gewinns bei größer werdendem Eingangs-SNR. Aufgrund nicht vorhandenem Hall und einer perfekten Ausrichtung zeigte sich kein Unterschied zwischen dem Griffith-Jim Beamformer und dem Verfahren nach Hoshuyama. Obgleich diese Beamformer eine deutliche Verbesserung erzielten, konnte der GEV-Beamformer diese Ergebnisse noch übertreffen. In Bild 3 sind beispielhaft die Spektrogramme für das reine Sprachsignal, das Signal nach der Hochpass-Filterung, nach dem Griffith-Jim- und dem GEV-Beamformer abgebildet. Beim Vergleich der Spektrogramme ist die relativ gute Rauschunterdrückung bei mittleren und tiefen Frequenzen durch den GEV-Beamformer zu erkennen.

Reale Überlagerung

Da sich bei den Aufnahmen von real überlagerten Sprach- und Autofahrgeräuschen bei den vier untersuchten Geschwindigkeiten keine eindeutige Zuordnung zwischen Geschwindigkeit und Eingangs-SNR durchführen ließ, also bei steigender Fahrgeschwindigkeit aufgrund des Lombardeffekts kein Absinken des SNR festzustellen war, wurden die Ergebnisse für jedes Verfahren über alle Fahrgeschwindigkeiten gemittelt. Dabei kamen für jede der 4 Geschwindigkeiten jeweils 4 Sprachsequenzen von dem Fahrer und Beifahrer in die Auswertung hinein. Das sich so ergebende Gesamtergebnis ist in Bild 2 rechts dargestellt. Obwohl die SNR-Gewinne aufgrund der realen Bedingungen geringer ausfallen als bei der synthetischen Überlagerung, läßt sich ein vergleichbares Verhältnis der Verfahren zueinander erkennen. Allerdings kann hier,

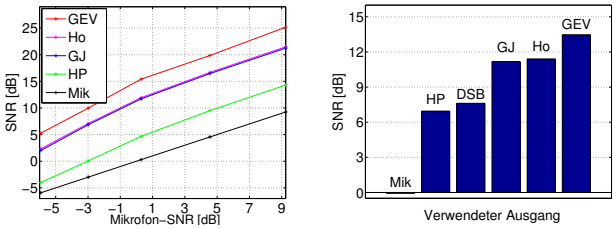

Abbildung 2: SNR-Gewinn für künstlich überlagerte Sprach- und Autofahrgeräusche (links) und für reale Fahrsituation (rechts).

wie bei den vorherigen Experimenten wiederum festgestellt werden, dass aufgrund des zusätzlichen SNR-Gewinns des GJ-Beamformers gegenüber dem DSB kein ideales diffuses Störgeräuschfeld vorlag [3].

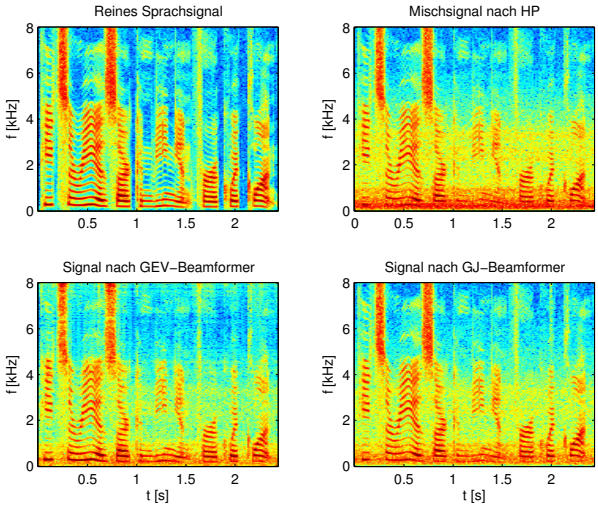

Abbildung 3: Spektrogramme für ein synthetisch überlagertes Sprach- und Fahrgeräusch.

Zusammenfassung

Es konnte gezeigt werden, dass das vorgestellte Verfahren, welches auf einer verallgemeinerten Eigenwertzerlegung basiert, eine blinde Adaption auf den Sprecher in der Fahrgastzelle eines Kfz unter realen Fahrbedingungen durchführt. Desweiteren wurde der theoretisch erzielbare SNR-Gewinn für ein diffuses Störschallfeld mit praktischen Ergebnissen im Kfz verglichen. Bei den SNR-Messungen des GEV-Beamformers zeigten sich durchweg bessere Ergebnisse im Vergleich zu den Beamforming-Verfahren nach Griffith-Jim und Hoshuyama.

Literatur

[1] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27-34, Jan. 1982.

[2] O. Hoshuyama and A. Sugiyama, "Robust adaptive beamforming", in *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, 2001.

[3] J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement", in *Proc. IEEE ICASSP*, Phoenix, May 1999.

[4] R. Haeb-Umbach and E. Warsitz, "Adaptive Filter-and-Sum Beamforming in Spatially Correlated Noise", in *Proc. IWAENC*, Eindhoven, Netherlands, Sep. 2005.