SIGNALVERARBEITUNG FÜR AKUSTISCHE MENSCH/MASCHINE – SCHNITTSTELLEN

*Walter Kellermann, Wolfgang Herbordt und Herbert Buchner**

Lehrstuhl für Multimediakommunikation und Signalverarbeitung, Universität Erlangen–Nürnberg
Cauerstr.7, 91058 Erlangen
{wk,herbordt,buchner}@LNT.de

Abstract: Die Grundaufgaben der digitalen Signalverarbeitung bei mehrkanaligen Aufnahme- und Wiedergabesystemen an der akustischen Mensch/Maschine-Schnittstelle werden anhand von Matrizengleichungen diskutiert. Dabei zeigt sich, dass es sich dabei stets um Signaltrennungs- und Systemidenti£kationsaufgaben mit unterschiedlich anspruchsvollen Randbedingungen handelt. Der Stand der Forschung wird beispielhaft für Echokompensation und 'beamforming' als Kernalgorithmen zur Extraktion von Wunschanteilen aus den gestörten Mikrophonsignalen dargestellt. Schließlich werden für ein integriertes System beispielhafte Ergebnisse vorgestellt, die für Anwendungen in der Telekommunikation und der Spracherkennung kennzeichnend sind.

1 Einführung

Es wird eine akustische Mensch/Maschine–Schnittstelle nach Abb.1 mit mehrkanaliger Lautsprecherwiedergabe und mehrkanaliger Mikrophonaufzeichnung betrachtet, mit der mehrere lokale Sprecher bzw. Hörer erfasst werden können.

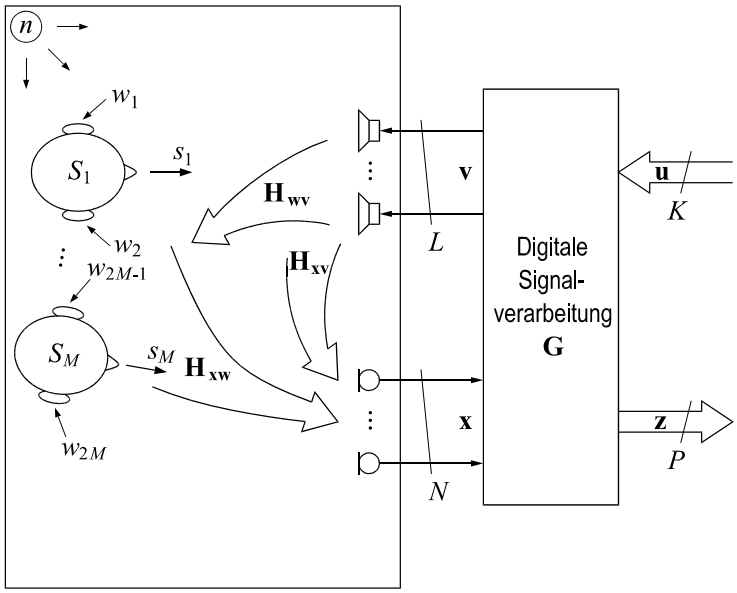

Abbildung 1 - Akustische Mensch/Maschine–Schnittstelle

Auf der Wiedergabeseite enthält Vektor $\mathbf{v}$ die von den L Lautsprechern wiederzugebenden Signale, die aus dem Quellensignalvektor $\mathbf{u}$ der Länge K abgeleitet werden. $\mathbf{w}$ enthält die tatsächlich an den $2M$ Ohren der Hörer anliegenden Signale, die im Idealfall einem gewünschten Höreindruck $\mathbf{w}_W$ entsprechen, während $\mathbf{s}$ die M aus der akustischen Umgebung aufzuneh-

*Teile dieser Arbeiten wurden unterstützt von der Fa. Intel Co., Santa Clara, CA, dem Bundesministerium f. Bildung und Forschung und der Fa. Grundig, und der EU-Kommission.

menden Signale erfasst. n repräsentiert eine Störquelle, die an den Ohren bzw. den Mikrophonen zu additiven Störvektoren $\mathbf{n_w}$, $\mathbf{n_x}$ führt. Aus den N Mikrophonsignalen $\mathbf{x}$ wird ein Vektor $\mathbf{z}$ berechnet, der im Idealfall $P \leq M$ gewünschte Quellensignale s_i enthält. Die Matrizen $\mathbf{H_{wv}}$,$\mathbf{H_{xv}}$, $\mathbf{H_{xs}}$ beschreiben die Übertragungssysteme zwischen den betreffenden Komponenten der Vektoren.

Typisch für das Szenario ist dabei, dass sich weder Lautsprecher noch Mikrophone in unmittelbarer Nähe der menschlichen Nutzer be£nden müssen, und diese sich zudem frei bewegen dürfen. Mit dieser allgemeinen Anordnung werden zahlreiche reale Situationen erfasst, bei denen natürliche oder künstliche akustische Umgebungen wiedergegeben werden sollen, und/oder Quellen für Kommunikation oder maschinelle Weiterverarbeitung aufgenommen werden sollen. Insbesondere gehören dazu Freisprech- und Lauthör-Einrichtungen bei Fahrzeugen, Multimedia-Endgeräten, und Telekonferenz, aber auch andere Telepräsenzsysteme, 'home theatres' und 'virtual reality'-Umgebungen fallen darunter. Von besonderem Interesse sind solche akustischen Schnittstellen auch für Spracherkennungs- und Sprachdialogsysteme.

Im weiteren wird untersucht, welche grundsätzlichen Signalverarbeitungsaufgaben mit der Erzeugung eines gewünschten Höreindrucks $\mathbf{w}_W$ und der Extraktion der gewünschten Quellensignale $\mathbf{z}$ aus $\mathbf{x}$ verbunden sind. Für zwei der Kernprobleme, die Kompensation akustischer Echos bei vielkanaliger Wiedergabe und die räumliche Filterung zur Signaltrennung bei Mikrophonarrays, werden jüngst publizierte Verfahren diskutiert. Schließlich illustrieren einige Ergebnisse für bereits realisierte Systeme den Stand der Forschung.

2 Grundaufgaben der Signalverarbeitung

Für das Folgende wird angenommen, dass die betrachteten Komponenten des Szenarios als lineare zeitdiskrete Systeme modelliert werden können, so dass sich die Funktion der Signalverarbeitung $\mathbf{G}$ durch Matrizengleichungen einfach darstellen lässt. Dies impliziert, dass $\mathbf{G}$ als lineares MIMO('multiple input/multiple output')–System nur lineare Faltungen auf den Zeitsignalen u_i, x_j ($i = 1, \ldots, K; j = 1, \ldots, N$) ausführt, die durch eine Matrix $\mathbf{G}$ mit den Teilmatrizen $\mathbf{G_{vu}}, \mathbf{G_{vx}}, \mathbf{G_{zu}}, \mathbf{G_{zx}}$ erfasst werden können[1]:

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \mathbf{G} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{G_{vu}} & \mathbf{G_{vx}} \\ \mathbf{G_{zu}} & \mathbf{G_{zx}} \end{pmatrix} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}. \tag{1}$$

Die durch die akustische Umgebung geprägten tatsächlichen Hörersignale $\mathbf{w}$ und Mikrophonsignale $\mathbf{x}$ setzen sich wie folgt zusammen:

$$\mathbf{w} = \mathbf{H_{wv}} * \mathbf{v} + \mathbf{n_w}, \tag{2}$$

$$\mathbf{x} = \mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}. \tag{3}$$

Aus dieser Darstellung lassen sich nun die Aufgaben der Signalverarbeitung $\mathbf{G}$ unmittelbar ableiten. Dabei darf vorausgesetzt werden, dass die Sprachsignale s_i und die Wiedergabesignale u_i zueinander und zu den Störvektoren $\mathbf{n_w}$, $\mathbf{n_x}$ unkorreliert sind.

[1]Bei der durch $\mathbf{y} = \mathbf{A} * \mathbf{x}$ de£nierten Faltung des Spaltenvektors $\mathbf{x}$ mit der Matrix $\mathbf{A}$ erhält man die Elemente $y_i(k)$ des Ergebnisvektors $\mathbf{y}$ gemäß $y_i(k) = \sum_{j=1}^{N} \sum_{n=-\infty}^{\infty} a_{ij}(k-n)x_j(n)$ unter der Annahme, dass die Impulsantwort $a_{ij}(k)$ zeitinvariant ist. Als Inverse $\mathbf{A}^{-1}$ wird diejenige Matrix bezeichnet, die die Beziehung $\mathbf{A}^{-1} * \mathbf{A} = \mathbf{I} \cdot \delta(k)$ erfüllt, wobei $\mathbf{I}$ die Einheitsmatrix und $\delta(k)$ der zeitdiskrete Einheitsimpuls sind. Bei nichtquadratischen Matrizen $\mathbf{A}$ bezeichnet $\mathbf{A}^{-1}$ die Pseudoinverse (siehe [1]).

2.1 Wiedergabe

Bei mehrkanaliger Wiedergabe zur Erzeugung eines gewünschten akustischen Höreindrucks $\mathbf{w}_W$ an den $2M$ Ohren M menschlicher Hörer soll erreicht werden:

$$\mathbf{w} \stackrel{!}{=} \mathbf{w}_W = \mathbf{H}_W * \mathbf{u}, \tag{4}$$

wobei die $2M \times K$-Matrix $\mathbf{H}_W$ die gewünschten, im allgemeinen zeitvarianten Impulsantworten $h_{ij}(k,l)$ zwischen Kanal u_j und Ohr i enthält. Um die gewünschte Wiedergabe ideal zu realisieren, muss mit Gln.1,2 demnach gelten:

$$\mathbf{H_{wv}} * (\mathbf{G_{vu}} * \mathbf{u} + \mathbf{G_{vx}} * \mathbf{x}) + \mathbf{n_w} \stackrel{!}{=} \mathbf{H}_W * \mathbf{u}. \tag{5}$$

Dies impliziert zwei Signalverarbeitungsaufgaben:

1. Entfaltung. Die Matrix $\mathbf{G_{vu}}$ muss den Ein¤uss der Raumimpulsantworten $\mathbf{H_{wv}}$ invertieren, wenn die Wiedergabesignalverarbeitung unabhängig vom gerade anliegenden Signalvektor $\mathbf{u}$ bleiben soll:

$$\mathbf{H_{wv}} * \mathbf{G_{vu}} * \mathbf{u} \stackrel{!}{=} \mathbf{H}_W * \mathbf{u} \Longrightarrow \mathbf{H_{vw}} * \mathbf{G_{uv}} \stackrel{!}{=} \mathbf{H}_W \Longrightarrow \mathbf{G_{vu}} \stackrel{!}{=} \mathbf{H_{wv}^{-1}} * \mathbf{H}_W \tag{6}$$

Abgesehen von der Sicherstellung der Kausalität von $\mathbf{G_{vu}}$ durch eine in $\mathbf{H}_W$ einzubringende Verzögerung, besteht offensichtlich das Hauptproblem hier darin, dass die Impulsantwortmatrix $\mathbf{H_{wv}}$ im allgemeinen nicht identi£ziert und damit auch nicht invertiert werden kann, da unter den geforderten Nutzungsbedingungen kein Referenzsignal an den Ohren der Hörer abgegriffen werden kann (”Blinde Entfaltung”). Ausnahmen bilden Fälle, in denen die Impulsantwort zum Kopf des Hörers und seine kopfbezogene Übertragungsfunktion bekannt sind (z.B. durch separate Messungen mit Referenzmikrophonen).

2. Störkompensation. Aus den Mikrophonsignalen $\mathbf{x}$ muss Referenzinformation über die Störsignale am Ohr gewonnen werden, die dann zur Kompensation verwendet werden kann:

$$\mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{x} + \mathbf{n_w} \stackrel{!}{=} \mathbf{0}. \tag{7}$$

Dies setzt voraus, dass die Störkomponenten an den Ohren $\mathbf{n_w}$ auch in den Mikrophonsignalen enthalten sind, so dass dort ein Störkomponentenvektor

$$\mathbf{n_{xw}} = \mathbf{H_{xw}} * \mathbf{n_w} \tag{8}$$

gemessen werden kann. Aufgabe der Signalverarbeitung ist es zunächst, $\mathbf{n_{x,w}}$ aus $\mathbf{x}$ zu extrahieren und dann ein Kompensationssignal über die Lautsprecher abzustrahlen, das sich wie folgt ergibt:

$$\mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{n_{xw}} = \mathbf{H_{wv}} * \mathbf{G_{vx}} * \mathbf{H_{xw}} * \mathbf{n_w} \stackrel{!}{=} -\mathbf{n_w}. \tag{9}$$

Neben der Extraktion von $\mathbf{n_x}$ aus $\mathbf{x}$ muss die Matrix $\mathbf{G_{vx}}$ demnach für signalunabhängige Störkompensation weiterhin leisten:

$$\mathbf{G_{vx}} \stackrel{!}{=} -\mathbf{H_{wv}^{-1}} * \mathbf{H_{xw}^{-1}}. \tag{10}$$

Offensichtlich ist $\mathbf{G_{vx}}$ nur dann kausal, wenn $\mathbf{H_{xw}^{-1}}$ die Nichtkausalität von $\mathbf{H_{wv}^{-1}}$ ausgleicht. Dies erfordert, dass $\mathbf{H_{xw}}$ nichtkausal ist, d.h., dass die Störquelle $n(t)$ näher an den Mikrophonen als an den Ohren liegt.

Während die Inversion von $\mathbf{H_{wv}}$ unter Punkt 1 als blindes Entfaltungsproblem mit unbekanntem Ausgangssignal $\mathbf{w}$ erkannt wurde, ist die Inversion von $\mathbf{H_{xw}}$ ein blindes Entfaltungsproblem, bei dem $\mathbf{w}$ als Eingangssignal des Systems nicht direkt messbar ist. Man beachte, dass Gl.7 ein aus anderem Zusammenhang bekanntes mehrkanaliges System zur aktiven Geräuschkompensation ('active noise control') beschreibt [2, 3], dass hier jedoch durch die für das Gesamtsystem gewollten größeren räumlichen Abstände zwischen den Sensoren und Aktoren einerseits und dem Ort der Störkompensation (Ohren) andererseits die Aufgabe noch erheblich erschwert wird.

Bisher eingesetzte Techniken lösen weder das erste noch das zweite Problem. Bei Stereo- oder 5-Kanal-Wiedergabe wird die lokale akustische Umgebung ($\mathbf{H_{wv}}, \mathbf{n_w}$ nicht berücksichtigt und die Matrix $\mathbf{G_{vu}}$ als Diagonalmatrix mit Verstärkungsfaktoren besetzt, so dass sich nur an einem Raumpunkt und nur in einem hallarmen Raum ohne Hintergrundstörung das gewünschte Höremp£nden einstellen kann. Bei der Wellenfeldsynthese [4] wird mit $L = 24, \ldots, 128$ dieser Raumpunkt ('sweet spot') durch einen größeren Raumbereich ersetzt, die Inversion der Raumakustik $\mathbf{H_{wv}}$ wurde jedoch dabei bisher nur ansatzweise für einen eng begrenzten Frequenzbereich und unter idealisierten Bedingungen erreicht. Realisierungen zur aktiven Störkompensation sind für dieses Szenario noch nicht bekannt.

2.2 Aufnahme

Das Ziel der Aufnahme ist ein Vektor $\mathbf{z}$ mit P getrennten und eventuell um k_0 verzögerten Quellensignalen $z_i(k) = s_j(k) * \delta(k - k_0), (i = 1, \ldots, P; j \in \{1, M\})$. Zur Gewinnung jedes Quellensignals $z_i(k)$ müssen von der Signalverarbeitung idealerweise die jeweils unerwünschten lokalen Nutzquellen sowie alle Störquellen unterdrückt, die akustischen Echos der Lautsprechersignale kompensiert, sowie Echos und Nachhall des gewünschten Quellensignals aus dem Mikrophonsignal entfernt werden. Zur formalen Vereinfachung wird im Folgenden $P = M$ angenommen, so dass man ausgehend von Gl.1 mit Gl.3 als Anforderung an die Aufnahme erhält:

$$\begin{aligned}
\mathbf{z} &= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * \mathbf{x} \\
&= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}) \\
&= (\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} + \mathbf{G_{zx}} * (\mathbf{H_{xs}} * \mathbf{s} + \mathbf{n_x}) \\
&\stackrel{!}{=} \mathbf{s} * \delta(k - k_0). \qquad (11)
\end{aligned}$$

Damit sind drei Anforderungen an die Signalverarbeitung verknüpft:

1. Echokompensation. Um das Übersprechen der Wiedergabesignale $\mathbf{u}$ auf die Aufnahmesignale $\mathbf{z}$ zu kompensieren, muss offensichtlich

$$(\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} = \mathbf{0} \qquad (12)$$

angestrebt werden. Aus der Sicht eines fernen Kommunikationspartners, der $\mathbf{u}$ sendet und $\mathbf{z}$ empfängt, entspricht dies der Kompensation des Echos der selbst abgegebenen Signale $\mathbf{u}$ ("Echokompensation"). Soll die Echokompensation unabhängig von $\mathbf{u}$ gewährleistet sein, dann ist

$$\mathbf{G_{zu}} = -\mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}} \qquad (13)$$

zu erfüllen. Dies entspricht einer mehrkanaligen Version der klassischen einkanaligen Systemidenti£kationsaufgabe, bei der Anregungs- und Ausgangssignal des zu identi£zierenden Systems beobachtet werden können. Man beachte, dass im Prinzip hier nur die Matrix $\mathbf{H_{xv}}$ identi£ziert werden muss, also die Matrix der Impulsantworten zwischen Lautsprechern und Mikrophonen.

2. Störunterdrückung. Zur Unterdrückung der lokalen Störungen muss idealerweise

$$\mathbf{G_{zx}} * \mathbf{n_x} = \mathbf{0} \tag{14}$$

erfüllt sein. Signalunabhängige Lösungen würden $\mathbf{G_{zx}} = \mathbf{0}$ erfordern, womit aber offensichtlich die Aufnahme der Quellensignale unmöglich wird. Entsprechend kann die Störunterdrückung deshalb nur dann ohne Beeinträchtigung der Quellensignale erfolgen, wenn sich die Störkomponenten in den Mikrophonsignalen von den Quellensignalen trennen lassen und separat unterdrückt werden können.

3. Quellentrennung und Enthallung. Schließlich müssen die Quellensignale getrennt und von Nachhall befreit werden, so dass

$$\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} = \mathbf{s} * \delta(k - k_0) \tag{15}$$

erfüllt wird. Dies bedeutet für signalunabhängige Lösungen, dass

$$\mathbf{G_{zx}} = \mathbf{H_{xs}^{-1}} * \delta(k - k_0) \tag{16}$$

angestrebt werden muss. Damit stellt sich für die Hauptdiagonalelemente von $\mathbf{G_{zx}} * \mathbf{H_{xs}}$ ein mehrkanaliges blindes Inversionsproblem und für die Nebendiagonalelemente ein Störunterdrückungsproblem ähnlich dem in Gl.14.

Wie bei der Wiedergabe lassen sich die Teilaufgaben zwei Kategorien zuordnen: Zum einen der Signaltrennung, zum anderen der Systemidenti£kation. Dabei ist die Trennung der einzelnen Signalkomponenten in $\mathbf{x}$ notwendige Voraussetzung für die Identi£kation der Komponenten $\mathbf{G_{zx}}$, $\mathbf{G_{zu}}$ und $\mathbf{G_{xv}}$. Grundsätzlich können die Signalkomponenten relativ einfach im Zeit- oder Frequenzbereich getrennt werden, falls sie in diesen Bereichen orthogonal sind. Dies ist jedoch bei den betrachteten Anwendung nicht immer gegeben, insbesondere können die Komponenten von $\mathbf{x}$ praktisch zu keiner Zeit und meist auch bei keiner Frequenz ohne irgendeine nichtverschwindende Störkomponente $\mathbf{n}$ betrachtet werden. Jedoch erlaubt die mehrkanalige Aufnahme gegenüber einkanaliger Aufnahme auch eine räumliche Selektivität entsprechend der räumlich abgetasteten Apertur, die durch die Mikrophongruppe realisiert wird ('Beamforming'). Damit können Quellen mit kohärenten Wellenfeldern aufgrund verschiedener Einfallsrichtungen getrennt werden. Schließlich können Signale auch auf der Basis statistischer Methoden getrennt werden, indem aus einer Mischung eine vorgegebene Anzahl von Signalen ermittelt wird, die untereinander statistisch unabhängig oder zumindest unkorreliert sind (blinde Quellentrennung).

Von den verschiedenen Systemidenti£kationsaufgaben ist die Echokompensation die einfachste, da hier Ein- und Ausgangssignale beobachtbar sind, wenn auch der Ausgangssignalvektor $\mathbf{H_{xv}} * \mathbf{v}$ in $\mathbf{x}$ nur gestört vorliegt. Die blinde Entfaltung gemäß Gl.16 ist aufgrund der Zeitvarianz und der Nichtminimalphasigkeit der Impulsantworten in $\mathbf{H_{xs}}$ und wegen des Fehlens eines geeigneten statistischen Modells für die Quellen s_i derzeit für realitätsnahe Szenarien noch ungelöst.

3 Echokompensation bei mehrkanalige Wiedergabe

Zur übersichtlichen Behandlung des Prinzips kann man annehmen, dass $\mathbf{G_{vu}} = \mathbf{I}_{K,K} \cdot \delta(k)$, und dass es genügt die Kompensation für ein einziges Mikrophon- und Ausgangssignal ($N = P = 1$) zu betrachten. Nach Gl.13 ergibt sich damit das Systemidenti£kationsproblem, $\mathbf{G_{zu}} = -\mathbf{H_{xv}}$

zu bestimmen, wobei die Matrizen zu Zeilenvektoren mit K (i.a. zeitabhängigen) Impulsantworten als Elementen werden:

$$\mathbf{G_{zu}} = (g_1(k), \ldots, g_K(k)), \tag{17}$$
$$\mathbf{H_{xv}} = (h_1(k), \ldots, h_K(k)). \tag{18}$$

Setzt man für die Nachbildung FIR-Filter gleicher Länge L_g (typischerweise $L_g = 1000, \ldots, 4000$) an, dann lässt sich das Schätzsignal für das Echo (vergleiche Abb.2) schreiben als

$$\widehat{y}(k) = \mathbf{g}^T(k)\mathbf{u}(k), \quad \text{wobei} \tag{19}$$
$$\mathbf{g}(k) = \left(\mathbf{g}_1^T(k), \ldots, \mathbf{g}_K^T(k)\right)^T, \quad \text{mit } \mathbf{g}_i(k) = \left(g_{i,0}(k), \ldots, g_{i,L_g-1}(k)\right)^T, \tag{20}$$
$$\mathbf{u}(k) = \left(\mathbf{u}_1^T(k), \ldots, \mathbf{u}_K^T(k)\right)^T, \quad \text{mit } \mathbf{u}_i(k) = (u_i(k), \ldots, u_i(k - L_g + 1))^T. \tag{21}$$

Als Schätzfehler wird betrachtet

$$e(k) = y(k) - \widehat{y}(k), \quad \text{wobei} \quad e(k) =: z(k)|_{\mathbf{s}=\mathbf{0}, n_x=0}, \quad y(k) =: x(k)|_{\mathbf{s}=\mathbf{0}, n_x=0}. \tag{22}$$

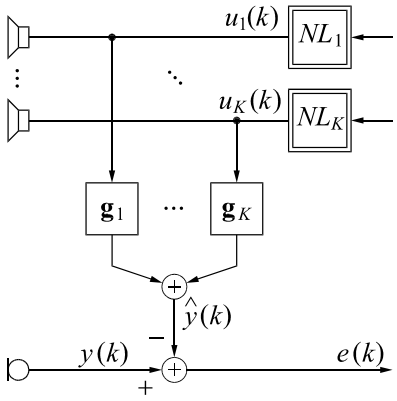

Abbildung 2 - Echokompensation bei K-kanaliger Wiedergabe

Wegen der Zeitvarianz der $h_i(k)$ werden zur iterativen Bestimmung des im Wiener'schen Sinne optimalen Vektors $\mathbf{g}(k)$ Adaptionsalgorithmen folgender Form eingesetzt:

$$\mathbf{g}(k) = \mathbf{g}(k-1) + \mathbf{K}(k)e(k), \tag{23}$$

wobei der Vektor $\mathbf{K}(k)$ die Richtung der Adaption bestimmt. Während für einkanalige Echokompensation ($K = 1$) einfache Adaptionsverfahren weit verbreitet sind (beispielsweise NLMS mit $\mathbf{K}(k) = \alpha\mathbf{u}/(\mathbf{u}^H\mathbf{u}), 0 < \alpha < 2$, siehe [5]), ist deren Konvergenzgeschwindigkeit für $K \geq 2$ unzureichend, da die K Kanäle $u_i(k)$ in der Regel stark kreuzkorreliert sind, weil sie von gemeinsamen Quellen erzeugt werden. Alternativ lässt der RLS('recursive least squares')-Algorithmus mit dem Kalman-Vektor $\mathbf{K}(k) = \mathbf{R}_{\mathbf{uu}}^{-1}(k) \cdot \mathbf{u}$ die schnellste Konvergenz erwarten. Allerdings muss auch hier die Konditionierung der zu $\mathbf{u}$ gehörigen Autokorrelationsmatrix $\mathbf{R_{uu}}$ verbessert werden, z.B. durch (möglichst unhörbare) Nichtlinearitäten NL_i (vgl. Abb.2). Da sich eine direkte Implementierung der Matrizeninversion $\mathbf{R}_{\mathbf{uu}}^{-1}(k)$ aus Aufwandsgründen verbietet, sind Näherungslösungen im DFT-Bereich besonders interessant. Jüngst wurde in [6] ein Algorithmus vorgestellt, der statt der Inversion der $(K \cdot L_G) \times (K \cdot L_G)$ Matrix nur L_G Matrizen der Größe $K \times K$ invertieren muss, und somit eine Echtzeitimplementierung eines $K = 5$-Kanal-Echokompensators mit $K \cdot L_G > 20000$ Filterkoef£zienten (Abtastfrequenz 12kHz) auf einem handelsüblichen PC (Intel 1.7GHz, 'dual processor board') erlaubt. In Abb.3 sind die Konvergenz des Systemabstands ($\propto \log_{10} \|\mathbf{G_{zu}} + \mathbf{H_{xv}}\|_2^2$) und der Echounterdrückung dargestellt.

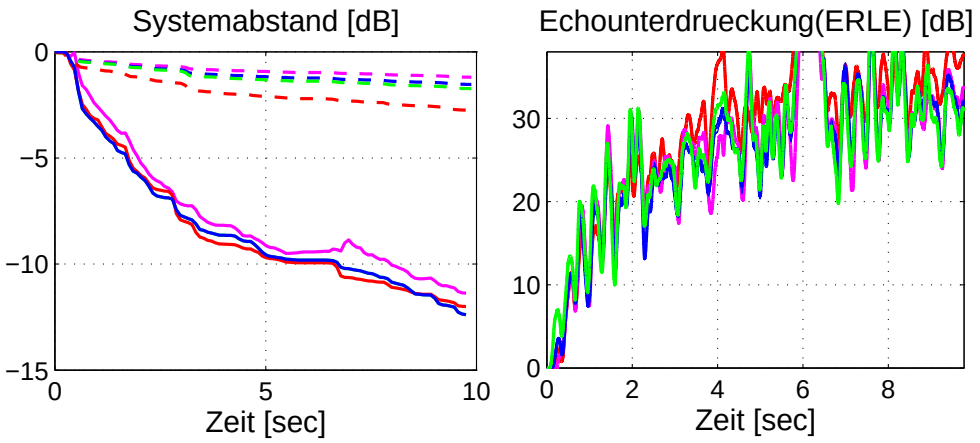

Abbildung 3 - Konvergenzverhalten des DFT-basierten Adaptionsalgorithmus nach [6] bei $K = 2, 3, 4, 5$-kanaliger Wiedergabe. Links Systemabstand relativ zum NLMS(gestrichelt), rechts Unterdrückung des Echosignals

4 Signaltrennung und Störunterdrückung mit Mikrophonarrays

Von den drei Aufgaben der Aufnahme führen Störunterdrückung einerseits und Quellentrennung bzw. Enthallung andererseits zu widersprüchlichen Anforderungen an $\mathbf{G_{zx}}$. Da in der Regel keine Trennung der Signalanteile von $\mathbf{x}$ im Zeit-/Frequenzbereich möglich ist, sondern eine Unterdrückung der Störung $\mathbf{n_x}$ nur auf Kosten einer gleichzeitigen Beeinträchtigung von $\mathbf{s}$ möglich ist, ist eine räumliche Trennung der Quellen durch 'beamforming' von besonderem Interesse. Eine besonders attraktive Struktur ist der in Abb.4 dargestellte 'Generalized Sidelobe Canceller', der ein Ausgangssignal z_i entsprechend einem Wunschsignal s_i ermittelt.

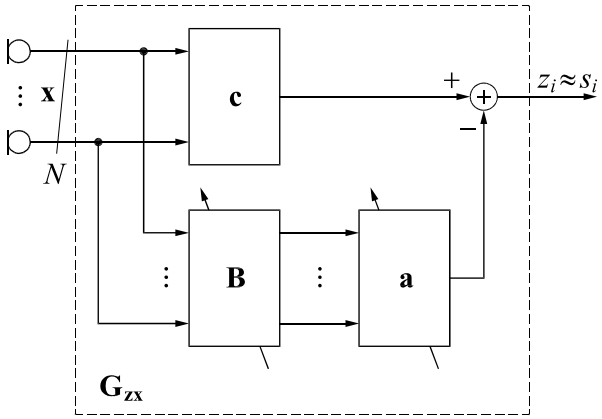

Abbildung 4 - Struktur eines 'Generalized Sidelobe Canceller'

Hierbei wird ein signalunabhängiger 'beamformer' $\mathbf{c}$ so entworfen, dass das aus einer bekannten Einfallsrichtung einfallende Wunschsignal unverzerrt bleibt, während Signale aus anderen Richtungen möglichst stark unterdrückt werden. Im unteren adaptiven signalabhängigen Pfad wird mit der Blockiermatrix $\mathbf{B}$ versucht, alle Wunschsignalanteile zu unterdrücken, so dass am Ausgang von $\mathbf{B}$ nur noch Störkomponenten anliegen. Der nachfolgende adaptive 'interference canceller' $\mathbf{a}$ versucht dann daraus ein Schätzsignal für die verbleibenden Störanteile im Ausgangssignal von $\mathbf{c}$ zu ermitteln. Die Adaption der Blockiermatrix $\mathbf{B}$ ist notwendig, um auch geringen Positionsänderungen der Quelle S_i zu folgen, da sonst in z_i das eigentliche Wunschsignal durch den 'interference canceller' $\mathbf{a}$ teilweise ausgelöscht wird. In realen Systemen entsprechen alle Elemente der Vektoren $\mathbf{a}, \mathbf{c}$ und der Matrix $\mathbf{B}$ adaptiven FIR-Filtern, die entweder im Zeitbereich oder im DFT-Bereich realisiert werden [7]. Typische Werte für die erreichte Störunterdrückung von unerwünschten Quellen gleichen Pegels ohne Verzerrung des Wunschsignals liegen bei ca. 25 dB.

Eine Alternative zum 'beamforming', bei dem die Position der Quellen als bekannt vorausgesetzt wird, stellt die blinde Quellentrennung dar, bei der mittels statistischer Methoden eine Filtermatrix $\mathbf{G_{zx}}$ ohne Kenntnis der Quellenpositionen so identi£ziert wird, dass am Ausgang statistisch unabhängige Signale z_i anliegen. Bei einem der bisher erfolgreichsten Verfahren wird dazu $\mathbf{G_{zx}}$ so bestimmt, dass die Kreuzkorrelation zwischen den Signalen z_i minimiert wird [8].
Eine explizite Enthallung nach Gl.16 ist mit beiden Verfahren nicht verbunden. Jedoch führt die räumliche Filterung des 'beamforming' in der Regel zu einer Verbesserung des Leistungsverhältnisses von Direktschall zu Nachhall.

5 Systemintegration

Für das allgemeine Szenario von Abb.1 ist das Zusammenwirken der einzelnen Signalverarbeitungskomponenten von entscheidender Bedeutung. Das Zusammenwirken von Echokompensation und 'beamforming' wurde z.B. in [9] untersucht, allgemeine Strategien bei vielkanaliger Wiedergabe mit $\mathbf{G_{vu}} \neq \mathbf{I}_{K,K} \cdot \delta(k)$ wurden in [10] vorgestellt.
Als Beispiel sei ein akustisches 'front-end' für Multimedia-Endgeräte angeführt, bei dem Stereo-Echokompensation ($K = L = 2$) und ein GSC-'beamformer' ($N = 8$ Mikrophone) kombiniert und im DFT-Bereich implementiert wurden [11]. Typischerweise werden etwa 15dB Störunterdrückung auch bei simultaner Aktivität von Wunschquelle und Störer erreicht (beides instationäre Sprachsignale), während gleichzeitig Lautsprecherechos um ca. 30dB gedämpft werden. Die Bedeutung der akustischen Vorverarbeitung für Sprachdialogsysteme wurde mit einem Diktiersystem ('Dragon System naturally speaking preferred') veri£ziert. Dabei ergaben sich die in Tabelle 1 dargestellten Erkennungsraten [7].

Akustische Umgebung	Einzelmikrophon	FBF	GSC	EK + GSC
Hallarmer Raum ($T_{60} = 50$msec)	32%	60%	92%	97%
Büroraum ($T_{60} = 300$msec)	30%	50%	86%	91%

Tabelle 1 - Worterkennungsraten in % bei einem Diktiersystem (Erkennungsrate mit Nahbesprechungsmikrophon:= 100%; Sprecherabstand zum Arrayzentrum 0.6m; FBF:= signalunabhängiger 'beamformer'; EK: Echokompensation)

6 Zusammenfassung und Ausblick

Von den verschiedenen Signalverarbeitungsaufgaben, die sich an der akustischen Mensch/Maschine-Schnittstelle ergeben, kann am ehesten die Echokompensation als gelöst betrachtet werden. Für Störunterdrückung und Quellentrennung mittels räumlicher Filterung bei der Aufnahme gibt es ebenfalls Verfahren, die in bestimmten Umgebungen gute Ergebnisse erzielen. Die Enthallung ist für praktische Anwendungen derzeit noch ungelöst. Andererseits erlaubt die Wellenfeldsynthese zwar die Erzeugung eines extern festlegbaren Schallfelds, Einfüsse der lokalen akustischen Umgebung können damit aber noch nicht kompensiert werden. Insgesamt werden die praktische Relevanz und die Schwierigkeit der ungelösten Probleme für die digitale Signalverarbeitung auf theoretischer wie praktischer Ebene auch in absehbarer Zukunft noch eine ernstzunehmende Herausforderung darstellen.

Literatur

[1] C. L. Lawson und R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[2] P.A. Nelson und S.J. Elliott, *Active Control of Sound*, Academic Press, London, 1992.

[3] S. M. Kuo und D. R. Morgan, *Active Noise Control Systems*, Wiley, New York, 1996.

[4] P.J. Berkhout, D. de Vries und P. Vogel, "Acoustic control by wave£eld synthesis," *J. Acoust. Soc. Am.*, Bd. 93, Nr. 5, S. 2764–2778, Mai 1993.

[5] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt und J. Tilp, "Acoustic echo control," *IEEE Signal Processing Magazine*, Bd. 16, Nr. 4, S. 42–69, Juli 1999.

[6] H.Buchner und W. Kellermann, "Improved Kalman gain computation for multichannel frequency-domain adaptive £ltering and application to acoustic echo cancellation," in *Proceedings Intern. Conference on Acoustics, Speech, and Signal Processing*, Orlando,FL, USA, Mai 2002, IEEE, S. 1909–1912.

[7] W. Herbordt und W. Kellermann, "Computationally ef£cient frequency-domain combination of acoustic echo cancellation and robust adaptive beamforming," in *EUROSPEECH)*, Aalborg, Dänemark, Sept. 2001, EURASIP, S. 1001–1004.

[8] L. Parra und C. Fancourt, "An adaptive beamforming perspective on convolutive blind source separation," in *Noise Reduction in Speech Applications*, G. Davis, Ed. CRC Press LLC, 2002.

[9] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein und D. Ward, Eds., Kap. 13, S. 281–306. Springer, Berlin, Mai 2001.

[10] H.Buchner, S.Spors, W. Kellermann und R. Rabenstein, "Full-duplex communication systems with loudspeaker arrays and microphone arrays," in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, Lausanne, Schweiz, Aug. 2002, IEEE.

[11] W. Herbordt und W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness," *European Transactions on Telecommunications (ETT)*, Bd. 13, Nr. 2, S. 123–132, März 2002.