"Natürliche" Akustische Mensch/Maschine-Schnittstellen – Eine Herausforderung für die digitale Signalverarbeitung

Walter Kellermann

Lehrstuhl für Multimediakommunikation und Signalverarbeitung, Universität Erlangen-Nürnberg, Email: wk@LNT.de

Einleitung

Bei herkömmlichen akustischen Mensch/Maschine–Schnittstellen kommt der Nutzer typischerweise mit technischem Gerät in physischen Kontakt oder er muss sich in dessen unmittelbarer Nähe aufhalten, will er die störenden Einflüsse seiner akustischen Umgebung gering halten. Ideale "natürliche" akustische Mensch/Maschine-Schnittstellen würden den Nutzern den Kontakt mit technischem Gerät ersparen und ihnen völlige Bewegungsfreiheit lassen. Dabei sollten diese Schnittstellen wohldefinierte Wunschsignalen an den Ohren der Hörer bereitstellen können, und die lokalen Quellen, insbesondere interessierende Sprecher oder Musikquellen, so aufnehmen als ob unmittelbar an der Quelle ein Mikrophon platziert wäre. Häufig ist auch die räumliche Anordnung der lokalen Quellen von Interesse, so dass Quellenlokalisierung als Aufgabe hinzukommt.

Die wiederzugebenden und die aufzunehmenden Signale sind typischerweise Sprachsignale oder andere Audiosignale, wie sie in den hier interessierenden Anwendungen auftreten, beispielsweise bei Freisprecheinrichtungen für Mobiltelephone, 'personal computers', Kraftfahrzeuge, Telekonferenzen, bei der Gerätebedienung im Heim- oder Industriebereich, in 'Virtual reality'-Umgebungen oder in Studios und auf Bühnen. Maßgebend für die Qualitätsbeurteilung ist bei der Wiedergabe das menschliche Hörempfinden, auf der Aufnahmeseite können dies auch objektiv messbare Leistungsparameter beispielsweise von Spracherkennungssystemen oder Quellencodierverfahren sein.

Die gewünschte räumliche Distanz zwischen Mensch und Maschine impliziert drei durch die Akustik bedingte Probleme, die mithilfe digitaler Signalverarbeitung wieder behoben werden sollen: Zum einen werden akustische Echos der Lautsprechersignale in die Mikrophone zurückgekoppelt, zweitens enthalten die an den Ohren der Hörer und in den Mikrophonen ankommenden Signale jeweils Echos und Nachhall aus der lokalen Umgebung, und drittens werden die Hörersignale und die Mikrophonsignale durch zusätzliche Störungen aus der lokalen Umgebung beeinträchtigt.

Dabei wachsen die Probleme für die Signalverarbeitung mit zunehmenden räumlichen Distanzen zwischen Lautsprechern und Hörern bzw. Quellen und Mikrophonen, mit zunehmendem Störpegel, und mit wachsender zu verarbeitender Signalbandbreite, mit der meist auch steigende Qualitätsansprüchen einhergehen.

Wir gehen im Folgenden davon aus, dass für Aufnahme und Wiedergabe jeweils mehrere Lautsprechern beziehungsweise Mikrophone zur Verfügung stehen ('Arrays'), und es werden entsprechend Signalverarbeitungsmethoden betrachtet, die die Mehrkanaligkeit ausnutzen und damit auch eine Wiedergabe bzw. Aufnahme der räumlichen Dimension einer akustischen Szene erlauben.

Aufgaben der Signalverarbeitung

In [1, 2] wurde anhand systemtheoretischer Betrachtungen gezeigt, dass die grundsätzlichen Probleme für die digitale Signalverarbeitung an der akustischen Mensch/Maschine-Schnittstelle im Wesentlichen entweder als Signaltrennungs- oder als Systemidentifikationsprobleme aufgefasst werden können. Die Schwierigkeit der jeweiligen Aufgabe hängt dann vor allem von der verfügbaren Referenzinformation ab. In Abhängigkeit davon werden die Probleme als überwacht oder nicht überwacht ('blind') klassifiziert.

Bei der **Wiedergabe** wohldefinierter Signale an den Ohren eines Hörers stellen sich aus systemtheoretischer Sicht zwei Aufgaben: Zum einen ist die Übertragungscharakteristik von den Lautsprechern zu den Ohren auszugleichen, und zum zweiten sind die lokalen Störungen zu kompensieren. Solange keine Mikrophone an den Ohren vorgesehen sind, ist dort kein Referenzsignal beobachtbar, so dass die Identifikation der Übertragungskanäle von den Lautsprechern zu den Ohren offensichtlich ein blindes Entzerrungsproblem ist. Dies ändert sich auch dann nicht, wenn die kopfbezogenen Übertragungsfunktionen ('head related transfer functions', HRTFs) bekannt sind, solange nicht gleichzeitig die Position und Orientierung der Hörer und die benötigten Raumimpulsantworten bekannt sind. Auch für die aktive Kompensation der Störanteile fehlt das Referenzsignal am Ohr, so dass auch die Störkompensation hier ein blindes Problem ist [1, 2]. Darüberhinaus führt die Forderung nach Bewegungsfreiheit der Hörer offensichtlich dazu, dass Kanalentzerrung und Störkompensation nicht punktweise, sondern für einen kontinuierlichen Raumbereich erzielt werden müssen und außerdem der Zeitvarianz der akustischen Szene Rechnung tragen müssen.

Möchte man mehrere Hörer gleichzeitig mit unterschiedlichem Wiedergabesignalen versorgen, dann müssen diese Probleme raumselektiv so gelöst werden, dass für die einzelnen Hörerpositionen im Raum nicht nur jeweils die Übertragungscharakteristik von den Lautsprechern zum Hörer entzerrt werden und gleichzeitig die Störquellen kompensiert werden müssen, sondern es muss durch unterschiedliche räumliche Filterung der verschiedenen Wiedergabesignale auch dafür gesorgt werden, dass diese

an den Hörerpositionen nicht interferieren.

Für die **Aufnahme** stellen sich drei Probleme, von denen die Kompensation der akustischen Echos der Lautsprechersignale am einfachsten erscheint: Hier gilt es die akustischen Pfade von den Lautsprechern zu den Mikrophonen nachzubilden, so dass die mit diesen Modellen erzeugten Schätzsignale die Echokomponenten in den Mikrophonsignalen kompensieren können. Da hierzu sowohl die Lautsprecher- wie auch die Mikrophonsignale beobachtbar sind, handelt es sich um ein überwachtes Systemidentifikationsproblem. Darüberhinaus sollen jedoch auch unerwünschte lokale Quellen unterdrückt werden und erwünschte Quellen voneinander getrennt werden. Je nachdem, ob dazu Referenzinformation über die Quellenpositionen bereitsteht, wird dies als ein überwachtes oder ein blindes Signaltrennungsproblem behandelt. Möchte man schließlich Echos und Nachhall aus den interessierenden Quellensignalen entfernen, dann sind die Übertragungskanäle von den Quellen zu den Mikrophonen zu entzerren, was wegen der fehlenden Mikrophone am Quellenort ein blindes Systemidentifikationsproblem darstellt. Man beachte, dass die in den Mikrophonsignalen enthaltene Referenzinformation für Echokompensation, Störunterdrückung und Enthallung jeweils durch eine Signaltrennung im Zeit- oder Frequenzbereich oder im räumlichen Bereich gewonnen werden muss.

Die **Quellenlokalisierung** mithilfe der Mikrophonsignale kann ebenfalls als blindes Problem angesehen werden, da keine Referenz für das Quellensignal vorliegt. Je nach Art des Verfahrens wird es als Signaltrennungs- oder Systemidentifikationsproblem angegangen (siehe unten).

Einige aktuelle Ergebnisse und Herausforderungen

Im Weiteren wird versucht, den Stand der Technik anhand kurzer Beschreibungen aktueller Beispiele zu illustrieren, und es werden einige aus der Sicht des Autors wesentliche Herausforderungen für die weitere Forschung aufgezeigt.

Kompensation akustischer Echos. Dieses überwachte Systemidentifikationsproblem, mit Lautsprecher- und Mikrophonsignalen als Referenzinformation, erfordert selbst für den Fall einkanaliger Wiedergabe immer noch ein adaptives FIR-Filter mit mehreren Hundert bis zu mehreren Tausend Koeffizienten zur Modellierung des akustischen Lautsprecher-Raum-Mikrophon-Systems. Dabei liegt die Schwierigkeit bei der praktischen Realisierung außer in der Rechenkomplexität vor allem in der Adaptionssteuerung, da die Adaption nicht nur durch lokales Hintergrundgeräusch, sondern vor allem auch durch alle lokalen Quellen gestört wird [3].

Für den Fall mehrkanaliger Wiedergabe wird das Problem auch theoretisch anspruchsvoller, da wegen der üblicherweise ausgeprägten Kreuzkorrelation zwischen den Wiedergabekanälen eine eindeutige Identifikation der jeweiligen Echopfade nicht mehr ohne Weiteres möglich ist, oder zumindest ein schlecht konditioniertes Systemidentifikationsproblem darstellt [4]. Zur notwendigen Verringerung der Kreuzkorrelation bieten sich drei Optionen an: Einfügen verschiedener Nichtlinearitäten in die Wiedergabekanäle [4, 5], Addition von kanalweise statistisch unabhängigem Rauschen [6] oder kanalweise verschiedene zeitvariante Signalverarbeitung, z.B. Phasenmodulation [7]. Ein jüngst vorgeschlagenes Verfahren zur zeitvarianten Phasenmodulation mit frequenzabhängigem Phasenhub [8] hat sich als bisher bestes Verfahren bezüglich optimaler Audio-Wiedergabequalität bei gleicher Konvergenzgeschwindigkeit der adaptiven Filter erwiesen. Mehrkanal-Echokompensation, wie Sie beispielsweise für die Sprachsteuerung einer 'Home Theatre'-Anlage im Freisprechbetrieb eingesetzt werden kann [9], kann derzeit mit bis zu fünf Wiedergabekanälen und insgesamt mehr als 20000 nachgebildeten Impulsantwortkoeffizienten auf üblichen PC-Plattformen implementiert werden [10].

Für die Wiedergabe von noch mehr Kanälen, insbesondere bei der Wellenfeldsynthese, wird die Echopfadnachbildung im Wellenbereich attraktiv ('Wave domain adaptive filtering' (WDAF), [11]). Hierbei wird durch die Transformation des abgestrahlten Schallfelds mittels arraygeometrieabhängiger Eigenfunktionen das Echokompensationsproblem dahingehend orthogonalisiert, dass bei einer Anordnung mit N Wiedergabe- und N Aufnahmekanälen statt N^2 Echopfaden nur N Echopfade identifiziert werden müssen. Für eine breiteren Einsatz des WDAF-Konzeptes sind insbesondere Transformationen für allgemeinere Arraygeometrien von Interesse, so dass solche Systeme unauffällig in üblichen Abhörumgebungen installiert werden können.

Interferenz- und Geräuschunterdrückung. Die unerwünschten Quellen in der lokalen akustischen Umgebung sind in der Regel als Punktquellen oder diffuse Quellen (oder als Mischformen) zu modellieren. Soll ihr Einfluss auf die interessierenden Signale minimiert werden, dann bietet sich im Fall mehrkanaliger Aufnahmen eine raumselektive Filterung an [12]. Beim datenunabhängigen 'Beamforming' wird dazu ohne Berücksichtigung der Signalstatistik der beteiligten Quellen eine Richtkeule erhöhter Empfindlichkeit auf eine erwünschte Quelle ausgerichtet und alle anderen Richtung werden soweit möglich unterdrückt. Als einfachste Version gilt dabei der 'Delay and Sum Beamformer', der die räumliche Richtwirkung allein durch geeignete Verzögerung und konstruktive Überlagerung der Mikrophonsignale erreicht.

Die Raumselektivität wird dabei entscheidend von der geometrischen Ausdehnung der Mikrophonanordnung ('Mikrophonarray') relativ zur Wellenlänge und der Anzahl der Mikrophone zur Abtastung des Wellenfelds bestimmt. Wegen der Ausdehnung des Audiofrequenzbereichs über etwa drei Dekaden erfordert eine gleichmäßige räumliche Selektivität für alle Frequenzen gleichzeitig eine große Apertur für niedrige Frequenzen und enge Mikrophonabstände für hohe Frequenzen. Dem kann man

mit 'geschachtelten' Arrays entgegenkommen [13], bei denen die Mikrophonabstände vom Zentrum zum Rand der Anordnung hin zunehmen. Dabei kann mit frequenzabhängiger Gewichtung der Mikrophonsignale eine näherungsweise konstante Keulenbreite angenähert werden ('Filter and Sum Beamformer'). Typischerweise muss jedoch wegen geometrischer Randbedingungen die Arrayausdehnung und damit für niedrige Frequenzen auch die Raumselektivität begrenzt bleiben. Um die Beschränkung der Selektivität bei niedrigen Frequenzen aufzuheben, kann man sogenannte differentielle ('superdirektive') Mikrophonarrays einsetzen [14], bei denen die räumliche Selektivität durch geeignete Differenzbildung zwischen den Sensorsignalen statt durch deren Addition erzielt wird. Damit lassen sich insbesondere räumliche Nullstellen in Richtung unerwünschter Punktquellen realisieren. Naturgemäß bedingen differentielle Beamformer eine Verstärkung räumlich unkorrelierter Störanteile ('white noise gain') wie etwa Sensorrauschen, so dass differentielle Arrays insbesondere bei höherer Ordnung nicht zuletzt eine sehr sorgfältige Mikrophonkalibrierung erfordern [15]. In der Regel wird die räumliche Selektivität bei niedrigen Frequenzen verringert, um die Rauschverstärkung zu begrenzen.

Bei aktuellen Verfahren für den Entwurf von datenunabhängigen Beamformern kann man von einer gegebenen Mikrophonanordnung ausgehen und dann einen Satz von FIR-Filtern für die Mikrophonsignale so ermitteln, dass man eine nahezu konstante Keulenbreite für die Nutzsignalrichtung erzielt und dabei zusätzliche Nebenbedingungen für die anderen Blickrichtungen einhält. Dabei ergeben sich bei üblichen Arraygeometrien Beamformer, die im unteren Frequenzbereich differentiell (superdirektiv) und im höheren Frequenzbereich wie konventionelle 'Filter and Sum'-Beamformer wirken [16].

Steht zusätzlich zu dem Wissen über die Quellenpositionen auch Information über die Signalstatistik der beteiligten Quellen zum Entwurf des Beamformers zur Verfügung, dann lassen sich statistisch optimale Beamformer einsetzen. In den am meisten verbreiteten Konzepten, LCMV ('Linearly Constrained Minimum Variance')- und MMSE('Minimum Mean Square Error')-Beamformern wird dazu lediglich Statistik zweiter Ordnung in Form von zeitlicher und räumlicher Korrelation der Mikrophonsignale ausgewertet [17].

In realistischen Szenarios muss die Nichtstationarität der Quellensignale und die Zeitvarianz der akustischen Umgebung berücksichtigt werden, so dass statistisch optimale Beamformer notwendigerweise adaptiv sein müssen und die Signalstatistiken während des Betriebs schätzen müssen. Als ein für Audio-Anwendungen besonders erfolgreiches Konzept hat sich eine Weiterentwicklung des sogenannten 'Generalized Sidelobe Cancellers' [18] (eine effiziente Realisierung des MVDR-Beamformers, der wiederum ein Spezialfall eines LCMV-Beamformers ist) erwiesen, die besonders hinsichtlich Robustheit gegenüber Bewegungen der Nutzsignalquelle optimiert ist ('Robust Generalized Sidelobe Canceller' (RGSC), [19, 20]). Die bisher besten Ergebnisse erhält man bei Implementierung des RGSC im DFT-Bereich, weil dann eine frequenzselektive Adaptionskontrolle eine optimale Interferenzunterdrückung ohne gleichzeitige Beeinträchtigung des Nutzsignals gewährleisten kann [21].

Bei den oben angeführten Verfahren zur raumselektiven Filterung wird stets davon ausgegangen, dass die Position der Nutzsignalquelle bekannt ist und teilweise wird auch die Position der Störer als bekannt vorausgesetzt. Durch Parallelschaltung mehrerer solcher Beamformer können prinzipiell auch mehrere gleichzeitig aktive Nutzquellen aus der akustischen Umgebung extrahiert werden, jedoch wird dann bei adaptiven signalabhängigen Beamformern die Schätzung der jeweiligen individuellen Signalstatistiken erheblich erschwert.

Für das Szenario mehrerer gleichzeitig aktiver Nutzquellen wurden in jüngerer Vergangenheit verstärkt Algorithmen zur blinden Quellentrennung ('Blind Source Separation', BSS) untersucht. Hierbei wird davon ausgegangen, dass jedes Mikrophon eine Mischung der mit den Raumimpulsantworten gefalteten Quellensignale enthält, wobei alle Quellensignale als statistisch unabhängig angenommen werden dürfen. Idealerweise erscheinen als Ausgangssignale des BSS-Algorithmus' die voneinander getrennten Quellensignale in einer Form, die bis auf eine Filterung mit den Originalquellen übereinstimmt. Zur Bestimmung der sogenannten Entmischmatrix werden Iterationsverfahren eingesetzt, die die statistische Unabhängigkeit der Ausgangssignale erzwingen sollen. Viele dieser Verfahren sind heuristisch entstanden oder basieren auf heuristischen Kostenfunktionen, die auf verschiedene Weise die statistischen Bindungen zwischen den Ausgangssignalen beschreiben. Im allgemeinen TRINICON-Konzept [22, 23, 24] finden sich viele dieser Verfahren als Spezialfälle wieder [25]. Das TRINICON-Konzept stellt dabei einen allgemeinen Rahmen für mehrkanalige blinde Signalverarbeitung bereit, wobei die Kostenfunktion auf der multivariaten Verbunddichte aller Ausgangskanäle beruht und für jeden Ausgangskanal mehrere zeitlich aufeinanderfolgende Abtastwerte berücksichtigt werden. Für die blinde Quellentrennung wird dann im allgemeinen Fall die Kullback-Leibler-Divergenz zwischen dem Produkt der multivariaten Dichten der einzelnen Ausgangskanäle und der Verbunddichte aller Kanäle der Kostenfunktion zugrundegelegt und daraus ein Gradientenverfahren abgeleitet, das die FIR-Filter der Entmischmatrix bestimmt. Ein populärer alternativer Ansatz sieht vor, die Faltungsmixturen als skalare Mischungen im DFT-Bereich zu modellieren ('Frequency-domain BSS', [26]), und dann individuell in jedem DFT-Bin die Quellen zu trennen. Die Vereinfachung auf skalare Mixturen erlaubt eine einfache Anwendung von BSS-Algorithmen, die auch Statistik höherer Ordnung ausnutzen. Dabei ist dann jedoch zusätzlich das sogenannte 'interne Permutationsproblem' zu lösen: Die jeweils getrennten Anteile aus verschiedenen DFT-Bins müssen wieder derselben Quelle zugordnet werden. Die dazu notwendigen Reparaturmaßnahmen müssen - teilweise mit Hilfe von geometrischer Zusatzinformation - wieder eine Bindung zwischen den DFT-

Bins erzwingen, die andererseits bei strenger Anwendung oder geeigneter Approximation des TRINICON-Kriteriums stets beibehalten wird [25]. Sowohl recheneffiziente 'Frequency-domain BSS'-Verfahren [27, 28, 29] als auch die leistungsfähigeren TRINICON-basierten Algorithmen [30] konnten mittlerweile in Echtzeitsystemen realisiert werden.

Im Vergleich zu den vorgenannten Beamformern können BSS-Verfahren auch als Systeme zur Interferenzunterdrückung verstanden werden, die gleichzeitig mehrere Nutzsignale extrahieren und dabei die bezüglich eines bestimmten Ausgangskanals unerwünschten Signale unterdrücken ('Blind Beamforming' [31]). Die Blindheit drückt sich in der Tatsache aus, dass dazu die Positionen der Quellen nicht bekannt sein müssen und - falls nicht wie bei 'Frequency-domain BSS' geometrische Zusatzinformation benötigt wird - auch die Arraygeometrie nicht bekannt sein muss.

Sowohl datenabhängiges Beamforming als auch BSS-Algorithmen basieren auf der Annahme, dass die gewünschten Quellen Punktquellen sind, deren Wellenfeld aus nur einer Einfallsrichtung auf das Array einfällt, und Signale aus anderen Richtungen hierzu unkorreliert sind. Wird diese Annahme zum Beispiel durch Echos allzusehr verletzt, dann ist mit Degradation zu rechnen. In der weiteren Verbesserung der Robustheit gegenüber halligen Umgebungen liegt deshalb eine wesentliche Herausforderung für die zukünftige Entwicklung dieser Algorithmen.

Enthallung. Die üblichen Optimierungskriterien für Beamforming, egal ob signalunabhängig oder statistisch optimal und/oder adaptiv, streben ebensowenig wie die der blinden Quellentrennung eine Enthallung der Quellensignale an. Für ideale Enthallung ist eine Entfaltung des Quellensignals mit den Raumimpulsantworten notwendig. Beamformer und BSS können jedoch aufgrund der damit einhergehenden raumselektiven Filterung auch ohne Entfaltung eine enthallende Wirkung haben, da Reflexionen von Wänden aus bestimmten Einfallsrichtungen unterdrückt werden. Dieser Effekt ist jedoch in der Regel nicht ausreichend, um beispielsweise akzeptable Erkennungsraten für Spracherkenner bei Sprechern in einem Abstand außerhalb des Hallradius' zu erzielen.

Während bei einkanaligen Enthallungsverfahren die Enthallung strenggenommen mit einem Filter erfolgen müsste, das die Raumimpulsantwort invertiert, und damit eine Übertragungsfunktion realisieren müsste, die Hunderte bis Tausende von Nullstellen nahe am Einheitskreis der z-Ebene invertieren müsste, können mehrkanalige Verfahren auf MINT ('Multiple-input/output INverse Theorem') [32]) zurückgreifen, das eine Inversion mit endlich langen Filtern garantiert, wenn die Länge des zu invertierenden mehrkanaligen Systems bekannt ist. Bei der Enthallung von Sprachsignalen ist zusätzlich zu beachten, dass die Entfaltung nicht etwa – wie bei der blinden Entfaltung zur Kanalentzerrung in der Übertragungstechnik – sämtliche zeitlichen Korrelation aus dem beobachteten Signal entfernen soll, sondern dass die vom Vokaltrakt herührende Korrelation erhalten bleiben muss, damit ein 'Whitening'-Effekt vermieden wird. Für eine solche partielle Mehrkanalentfaltung wurden mittlerweile mehrere erfolgreiche Algorithmen vorgestellt [24, 33, 34], die jedoch noch nicht als Echtzeitverfahren realisiert sind. Das TRINICON-basierte Verfahren [24] ergibt sich hier als eine direkte Erweiterung des BSS-Verfahrens einfach durch eine modifizierte Kostenfunktion und erlaubt die simultane Trennung und Enthallung mehrerer Quellen. Für die zukünftige Forschung stellt insbesondere die Robustheit der Verfahren in realen Szenarios mit bewegten Quellen noch eine erhebliche Herausforderung dar.

Quellenlokalisierung. Die Lokalisierung von Schallquellen in akustischen Umgebungen war lange Zeit auf die Auswertung von Kreuzkorrelationsfunktionen zwischen den Mikrophonsignalen mit verschiedenen Varianten in Zeit- und Frequenzbereich beschränkt. Dem zugrunde liegt stets das Modell der Schallausbreitung im Freifeld, so dass der Laufzeitunterschied zwischen den Mikrophonsignalen einen direkten Rückschluss auf den Einfallswinkel in der betrachteten Ebene und damit auf die Quellenposition erlaubt [35]. Offensichtlich ist dieses Modell in geschlossenen Räumen mit ausgeprägten Reflexionen nur noch bedingt sinnvoll und liefert entsprechend unbefriedigende Ergebnisse. Alternativ kann man mit Beamformern die akustischen Umgebungen absuchen und die Orte, von denen maximale Signalenergie empfangen wird, als Quellenpositionen annehmen (SRP [35]). Auch hier ist bei mehreren Quellen und signifikanten Reflexionen mit Fehlentscheidungen zu rechnen. Ein neues leistungsfähiges Lokalisierungsverfahren für mehrere Quellen, das noch dazu mit sehr kleinen Arrayausdehnungen auskommt, wurde ausgehend vom Konzept der Wellenfeldzerlegung in Eigenfunktionen entwickelt, wie es auch dem 'Eigenmike' ([36]) zugrundeliegt: Die Wellenfeldzerlegung erlaubt dabei eine Anwendung der für Schmalbandsignale entwickelten Unterraum-Methoden auf Breitbandsignale, so dass man mit den resultierenden 'Eigenbeams' ohne großen Rechenaufwand beispielsweise mit einem zirkularen Array mit zehn Mikrophonen bis zu fünf Quellen gleichzeitig lokalisieren kann [37].

Um Echos und Nachhall der akustischen Umgebung Rechnung zu tragen, wurde in [38] ein Ansatz zur Systemidentifikation zur Lokalisierung eingesetzt, bei dem letztlich die Impulsantworten einer Quelle zu zwei Mikrophonen identifiziert werden ('Adaptive Eigenvalue Decomposition') und aus der relativen zeitlichen Lage der Maxima der Impulsantworten auf den Laufzeitunterschied zu den Mikrophonen geschlossen wird. Dieses Prinzip kann mithilfe TRINICON-basierter blinder Quellentrennung auch zur gleichzeitigen Lokalisierung mehrerer Quellen verwendet werden, da auch hier die Ausbreitungspfade von den Quellen zu den Mikrophonen durch die Entmischmatrix nachgebildet werden [39]. Bei der gleichzeitigen Lokalisierung in mehreren Dimensionen mittels mehrerer BSS-Systeme kann durch Korrelationsbildung zwischen den Ausgangssignalen der BSS-

Systeme die richtige Zuordnung der Ausgangssignale und der Laufzeitunterschiede zu den jeweiligen Quellen erhalten werden [40].

Schallfeldwiedergabe. Die Mehrkanalwiedergabe über mehrere Lautsprecher ist seit Einführung der Stereophonie etabliert und jüngere Verfahren versprechen einen raumgetreuen Höreindruck mit typischerweise fünf bis sieben Wiedergabekanälen. Dieser wird jedoch nur im sogenannten 'sweet spot' erzielt und auch dort werden Raumakustik der Abhörumgebung und zusätzliche Störungen nicht phasenrichtig entzerrt beziehungsweise kompensiert. Die Beschränkungen des 'sweet spot' können mit Schallfeldsyntheseverfahren aufgehoben werden, bei denen versucht wird, das akustische Wellenfeld in jedem Punkt eines bestimmten Raumbereichs festzulegen. Die aktuelle Forschung widmet sich dazu vor allem der Wellenfeldsynthese (WFS, [41, 42]), die auf dem Huygen'schen Prinzip und dem Kirchhoff-Helmholtz-Integral basiert und mindestens mehrere Dutzend Lautsprecherkanäle erfordert. Dieses Verfahren erlaubt prinzipiell auch die Entzerrung der Abhörumgebung [42] und die Kompensation von lokalen Geräuschquellen [43]. Die dazu vorgeschlagenen adaptiven Verfahren [44] zur notwendigen Verfolgung der Zeitvarianz der Umgebung und zur Einbeziehung beweglicher Streukörper (nicht zuletzt der Hörer selbst) sind jedoch bisher noch nicht bis zu robusten Implementierungen für praktische Systeme entwickelt.

In diesem Zusammenhang sei betont, dass bei der Wiedergabe letztlich das menschliche Gehör über die Qualität entscheidet, und damit die bisher hauptsächlich systemtheoretisch motivierten Optimierungskriterien der Signalverarbeitung vor allem auch psychoakustischen Gesichtspunkten Rechnung tragen sollten. Insbesondere kann man aufgrund der Erfahrungen bei der Audiocodierung [45] hoffen, dass die Anforderungen an die Signalverarbeitung zur raumgetreuen Wiedergabe verringert werden, wenn durch psychoakustische Untersuchungen noch besser geklärt ist, inwiefern die systemtheoretischen Kriterien aufgeweicht werden können, ohne das Hörerlebnis zu beeinträchtigen.

Perspektiven

Von den verschiedenen Problemen auf der Aufnahmeseite wird die Kompensation akustischer Echos theoretisch als weitgehend gelöst betrachtet, Forschungsbedarf besteht hier vor allem, wenn es darum geht, recheneffiziente und gleichzeitig schnell konvergierende Verfahren für viele Lautsprecherkanäle bereitzustellen.

Die Interferenzunterdrückung und die Geräuschreduktion bleibt eine Herausforderung, solange die Behandlung halliger Umgebungen mit weit von den Mikrophonen entfernten Nutzquellen ungelöst ist. Der Nachhall ist dann auch bei BSS-Algorithmen problematisch, da eine erhöhte Filterlänge der Entmischfilter nicht notwendig zu besserem Verhalten in halliger Umgebung führt.

Die Enthallung als ein noch vergleichsweise junges Forschungsgebiet hat noch nicht die Robustheit für die Anwendung im hier behandelten Szenario mit zeitvarianten akustischen Umgebungen und bewegten Quellen erreicht, so dass entsprechende Algorithmen für realitätsnahe Echtzeitimplementierungen noch ausstehen.

Die vorgestellten Lokalisierungsverfahren sind dagegen auch für mehrere Quellen schon vielfach in Echtzeitsystemen implementiert und in realen Szenarien erprobt. Eine spezielle Herausforderung liegt hier vor allem in der Lokalisierung von Schallereignissen sehr kurzer Dauer.

Auf der Wiedergabeseite befinden sich bei der Wellenfeldsynthese die Ansätze zur Entzerrung der Raumakustik und der Kompensation von Störungen noch in den Anfängen der Entwicklung und bieten hochinteressante Fragestellungen im Spannungsfeld von Signalverarbeitung und Psychoakustik.

Aus systemtheoretischer Sicht fällt auf, dass bei allen Teilproblemen mittels adaptiver Filter Signale getrennt oder Systeme modelliert werden. Zur Identifikation dieser Filter wird bisher meist nur Statistik zweiter Ordnung ausgenutzt und damit werden Lösungen erzielt, die lediglich für Gauss-Prozesse optimal sind. Die akustischen Signale sind jedoch in der Regel nicht normalverteilt, so dass die Berücksichtigung von Statistik höherer Ordnung für zukünftige Optimierungskriterien der adaptiven Systeme attraktiv erscheint. Erste Schritte in Richtung robuste Statistik wurden bereits für die Steuerung von Echokompensation [46, 47] und Beamformern [48] unternommen. 'Frequency-domain- BSS'-Algorithmen berücksichtigen Statistik höherer Ordnung zur Trennung skalarer Mixturen. Es ist zu erwarten, dass auch für die Trennung von Faltungsmixturen die Statistik höherer Ordnung bald intensiver genutzt wird und sie auch für die überwachte Adaption von linearen Filtern zunehmend Gegenstand der Forschung wird.

Über die Lösung der Teilprobleme hinaus wird sich die effiziente Integration der Einzellösungen in ein Gesamtsystem für die Mensch/Maschine-Schnittstelle zu einem zunehmend wichtigen Forschungsgebiet entwickeln. Am Beispiel der Kombination von Beamforming und Echokompensation [49] oder von BSS und Geräuschreduktion [50] kann man erkennen, dass die Kombination insbesondere mehrerer adaptiver Systeme mit eigenen Schwierigkeiten verbunden sein kann, aber auch wertvolle Synergien erlauben kann [51].

Schlussbemerkung

Die Signalverarbeitung an der akustischen Mensch/Maschine-Schnittstelle hat seit etwa dreißig Jahren ein ständig wachsendes Interesse in der Wissenschaftsgemeinde und in der Industrie erfahren, wobei mehrkanalige Verfahren in den letzten Jahren zunehmend ins Zentrum der Aufmerksamkeit rückten. Die natürliche Komplexität akustischer Szenarios bleibt auch heute noch eine Herausforderung für die Signalverarbeitung, die ihre Faszination vor allem auch deshalb wahrt, weil durch zunehmend leistungsfähige

Hardware immer neuartige und zunehmend komplexere Algorithmen gewinnbringend eingesetzt werden können.

Literatur

[1] Kellermann, W.: Signalverarbeitung für akustische Mensch-Maschine-Schnittstellen. Tagungsband 13. Konf. elektronische Sprachsignalverarbeitung (ESSV), 2002, 49-57

[2] Kellermann, W. et al.: Multichannel Acoustic Signal Processing for Human/Machine Interfaces - Fundamental Problems and Recent Advances. Proc. Int. Conf. on Acoustics (ICA), Kyoto, Japan, Apr. 2004

[3] Breining, C. et al.: Acoustic Echo Control. IEEE Signal Processing Magazine (1999), Nr.4, 42-69

[4] Benesty, J., et al.: A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation. IEEE Trans. on Speech and Audio Processing 6(1998) 156-165

[5] Morgan, D.R.; Hall, J.L.; Benesty, J.: Investigation of Several Types of Nonlinearities for Use in Stereo Acoustic Echo Cancellation. IEEE Trans. on Speech and Audio Processing 9(2001)6, 686- 696

[6] Gaensler, T.; Eneroth, P.: Influence of Audio Coding on Stereophonic Acoustic Echo Cancellation. Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98), 1998, 3649-3652.

[7] Ali, M.: Stereophonic Acoustic Echo Cancellation System Using Time-Varying All-Pass Filtering for Signal Decorrelation Proc. of Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98), 1998, 3689-3692

[8] Herre, J.; Buchner, H.; Kellermann, W.: Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'07), I-17 – I-20, 2007

[9] DICIT - Distant talking Interfaces for Control of Interactive TV. EU-Projekt FP6 IST-034624. `http://dicit.itc.it`

[10] Buchner, H.; Kellermann, W.: Improved Kalman Gain Computation for Multichannel Frequency-Domain Adaptive Filtering and Application to Acoustic Echo Cancellation. Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02), 2002, 1909-1912

[11] Buchner, H.; Spors, S.; Kellermann, W.: Wave-Domain Adaptive Filtering: Acoustic Echo Cancellation for Full-Duplex Systems Based on Wave-Field Synthesis. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04), IV-117 – IV-120, 2004

[12] Brandstein, M.; Ward, D. (Eds.): Microphone Arrays. Signal Processing Techniques and Applications. Springer, Berlin, 2001.

[13] Flanagan, J.L., et al: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am., 78(1985)5, 1508-1518

[14] Elko, G.: Microphone array systems for hands-free telecommunication. Speech Communication 20(1996), 229-240

[15] Elko, G.: Differential Microphone Arrays. In Huang, Y.; Benesty, J. (Eds.) *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Kluwer 2004, 2-65

[16] Parra, L.C.: Steerable frequency-invariant beamforming J. Acoust. Soc. Am., 119(2006)6, 3839-3847

[17] Bitzer, J. and Simmer, K.: Superdirective Microphone Arrays. In Brandstein, M.; Ward, D. (Eds.) *Microphone Arrays. Signal Processing Techniques and Applications*. Springer, Berlin, 2001, 19-38

[18] Griffiths, L.J.; Jim, C.W.: An alternative approach to Linear Constrained Adaptive Beamforming. IEEE Trans. Antennas and Propagation 30(1982)1, 27-34

[19] Hoshuyama, O.; Sugiyama, A.; Hirano, A.: A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96), 1996, 925-928

[20] Herbordt, W.; Kellermann, W.: Computationally Efficient Frequency-Domain Robust Generalized Sidelobe Canceller. International Workshop on Acoustic Echo and Noise Control (IWAENC), 2001, 51-54

[21] Herbordt, W.: Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing. Lecture Notes in Control and Information Sciences, vol. 315, Springer, Heidelberg, 2005

[22] Buchner, H.; Aichner, R.; Kellermann, W.: A Generalization of a Class of Blind Source Separation Algorithms for Convolutive Mixtures. Proc. IEEE Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA), 2003, 945-950

[23] Buchner, H.; Aichner, R.; Kellermann, W.: Blind Source Separation for Convolutive Mixtures Exploiting Nongaussianity, Nonwhiteness, and Nonstationarity. Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC), 2003, 275-278,

[24] Buchner, H.; Aichner, R.; Kellermann, W.: TRINICON: A Versatile Framework for Multichannel Blind Signal Processing. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04), 2004, III-889 – III-892

[25] Kellermann, W.; Buchner, H.; Aichner, R.: Separating Convolutive Mixtures with TRINICON. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'06), 2006, V-961 – V-964

[26] Makino, S., et al.: Blind source separation of convolutive mixtures of audio signals in frequency domain. In E. Haensler and G. Schmidt, Eds. *Topics in Acoustic Echo and Noise Control* Springer, 2006

[27] Parra, L.; Spence, C.: Convolutive blind source separation of non-stationary sources. IEEE Trans. on Speech and Audio Processing 8(2000)3, 320-327.

[28] Mukai, R., et al.: Real-time blind source separation and DOA estimation using small 3-D microphone array. Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC), 2005, 45-48

[29] Mori, Y., et al.: Real time implementation of two-stage blind source separation combining SIMO-ICA and binary masking. Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC), 2005, 229-232

[30] Aichner, R., et al.: Real-Time Convolutive Blind Source Separation based on a Broadband Approach. Fifth Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA), 2004

[31] Cardoso, J.; Soloumiac, A.: Blind beamforming for non-Gaussian signals. IEE ProceedingsF, 140(1993)46, 362–370

[32] Miyoshi, M.; Kaneda, Y.: Inverse filtering of room acoustics, IEEE Trans. Acoustics, Speech, and Signal Processing, 36(1988)2, 145-152

[33] Delcroix, M.; Hikichi, T.; Miyoshi, M.: Precise dereverberation using multi-channel linear prediction, IEEE Trans. Acoustic, Speech and Language Processing, 15(2007)2, 430-440

[34] Furuya, K.; Kataoka, A.: Hybrid Dereverberation Using Blind Deconvolution and Spectral Subtraction to Compensate for Motion of Source. Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC), 2006,

[35] DiBiase, J.H.; Silverman, H.F.; Brandstein, M.S.: Robust localization in reverberant rooms. In Brandstein, M.; Ward, D. (Eds.) *Microphone Arrays. Signal Processing Techniques and Applications.* Springer, Berlin, 2001, 157-180

[36] Meyer, J.; Elko, G.: A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'06), 2006, II-1781 – II-1784

[37] Teutsch, H.; Kellermann, W.: Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays. J. Acoust. Soc. Am., 120(2006)5, 2724-2736

[38] Benesty, J.: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am., 107(2000) 384-391

[39] Buchner, H., et al.: Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'05), 2005, III-97 – III-100

[40] Lombard, A.; Buchner, H.; Kellermann, W: Multidimensional Localization of Multiple Sound Sources Using Blind Adaptive MIMO System Identification. Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2006.

[41] Berkhout, A.J.: A holographic approach to acoustic control. J. Audio Engineering Soc., 36(1988), 977-985

[42] Spors, S., et al.: Sound Field Synthesis. In Huang, Y.; Benesty, J. *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer, 2004, 323-342

[43] Kuntz, A.; Rabenstein, R.: An Approach to Global Noise Control by Wave Field Synthesis. Proc. 12th European Signal Processing Conference (EUSIPCO), 2004

[44] Spors, S.; Buchner, H.; Rabenstein, R.: A Novel Approach to Active Listening Room Compensation for Wave Field Synthesis using Wave-Domain Adaptive Filtering. Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04), 2004, IV-29 – IV-32

[45] Faller, C.: Parametric Multichannel Audio Coding: Synthesis of Coherence Cues. IEEE Trans. Speech and Audio Processing, 14(2006)1, 299-310

[46] Gaensler, T.: A double-talk resistant subband echo canceller. Signal Processing 65(1998)1, 89-101

[47] Buchner, H., et al.: Robust extended multidelay filter and double-talk-detector for acoustic echo cancellation. IEEE Trans. Audio, Speech, Lang. Proc., 14(2006)5, 1633-1644

[48] Herbordt, W., et al.: Multichannel bin-wise robust frequency-domain adaptive filtering abd its application to adaptive beamforming. IEEE Trans. Audio, Speech, Lang. Proc., 15(2007)4, 1340-1351

[49] Kellermann, W.: Acoustic echo cancellation for beamforming microphone arrays. In Brandstein, M.; Ward, D. (Eds.) *Microphone Arrays. Signal Processing Techniques and Applications.* Springer, Berlin, 2001, 281-306

[50] Aichner, R., et al.: Post-processing for Convolutive Blind Source Separation. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'06), 2006

[51] Herbordt, W.; Buchner, H.; Kellermann W.: An acoustic human-machine front-end for multimedia applications. European Journal on Applied Signal Processing, (2003)1, 1-11