Enthallung und Trennung von Sprachsignalen mittels blinder adaptiver MIMO-Filterung

Herbert Buchner

Deutsche Telekom Laboratories, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin
herbert.buchner@telekom.de

Abstract: In diesem Beitrag stellen wir eine Klasse von neuen Algorithmen zur blinden Enthallung von Sprachsignalen und gleichzeitiger Trennung von Signalgemischen (blind source separation, BSS) vor, basierend auf TRINICON, einem allgemeinen Konzept für breitbandige adaptive MIMO-Signalverarbeitung. Um alle grundlegenden stochastischen Signaleigenschaften von Sprache für den Enthallungs- und Separationsprozess ausnutzen zu können und um die bekannten Whitening-Artefakte in bisherigen Verfahren zu vermeiden, schlagen wir die Einführung eines speziell entworfenen Signalmodells vor, welches auf einer Entwicklung mittels multivariater Tschebyscheff-Hermite-Polynome basiert. Das multivariate Modell beinhaltet auch inhärent eine lineare Prädiktion, welche bekanntermaßen direkt mit der Modellierung des menschlichen Vokaltrakts in Verbindung gebracht werden kann. Das vorgestellte Konzept ist anwendbar sowohl für Einzelsprecher-Szenarien, als auch für mehrere, simultan aktive Sprecher. Im letzteren Fall beinhaltet es zusätzlich zur Enthallung auch eine blinde Quellentrennung (BSS).

1 Einleitung

Bei der breitbandigen Signalaufnahme mittels Sensor-Arrays, wie z.B. in modernen Sprachkommunikationssystemen, werden die ursprünglichen Quellensignale $s_q(n)$, $q = 1, \ldots, Q$ durch ein lineares System mit mehreren Eingängen und mehreren Ausgängen (MIMO-System), z.B. den halligen Raum, gefiltert, bevor sie als Sensorsignale $x_p(n)$, $p = 1, \ldots, P$ verarbeitet werden. In diesem Beitrag beschreiben wir dieses MIMO-Mischsystem durch FIR-Filtermodelle entsprechend Abb. 1, wobei $h_{qp,\kappa}$, $\kappa = 0, \ldots, M-1$, die Koeffizienten des Modells von der q-ten Signalquelle zum p-ten Sensor (Signal $x_p(n)$) bezeichnet. Darüberhinaus nehmen wir in diesem Beitrag an, dass $Q \leq P$. Diese Fälle werden als überbestimmt bzw. bestimmt bezeichnet. Dabei ist zu beachten, dass die Quellen $s_q(n)$ nicht notwendigerweise alle

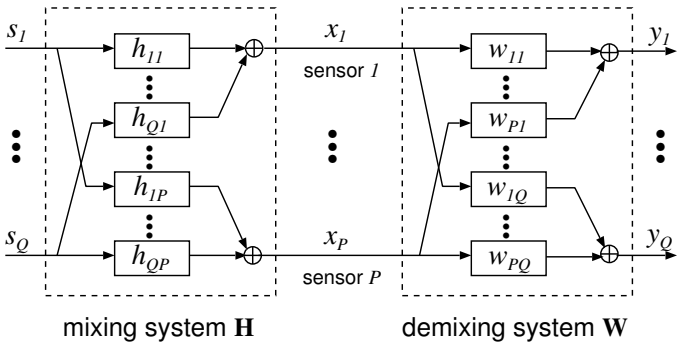

Abbildung 1 - Anordnung für blinde MIMO-Signalverarbeitung.

gleichzeitig zu einem bestimmten Zeitpunkt aktiv sein müssen. Wir sind daran interessiert, mittels adaptiver Signalverarbeitung optimale FIR-Entmischsysteme mit den Koeffizienten $w_{pq,\kappa}$ und der Länge L zu finden, welche die Ausgangssignale $y_q(n)$ liefern. Da in praktischen akustischen Szenarien weder die ursprünglichen Quellensignale $s_q(n)$, noch das Mischsystem direkt zugänglich sind, muß die Adaption des Entmischsystems *blind* erfolgen.

Basierend auf dieser MIMO-Struktur können wir die folgenden blinden Signalverarbeitungsprobleme für die Verbesserung der Ausgangssignale $y_q(n)$ unterscheiden:

(a) Signalseparation: Auslöschen aller Kreuzpfade in der Kaskade aus Misch- und Entmischsystem.

(b) Entfaltung/Enthallung: Zusätzlich zur Signalseparation sind die "trockenen" Quellen bis auf eine Verzögerung und einen Skalierungsfaktor zu ermitteln.

Im Gegensatz zum Separationsproblem (a), welches eng mit der blinden Systemidentifikation zusammenhängt [2], macht das wesentlich anspruchsvollere Problem (b) auch eine *Inversion* des Mischsystems notwendig (welches typischerweise nichtminimalphasige Impulsantworten beinhaltet). Es kann gezeigt werden, dass eine ideale signal-unabhängige breitbandige Separationslösung für $Q \leq P$ existiert [2], während für die ideale Lösung des inversen Problems (b) der überbestimmte Fall $Q < P$ nötig ist [4]. In diesem Beitrag behandeln wir das Problem der Enthallung von Sprachsignalen (b) mittels TRINICON ('TRIple-N ICA for CONvolutive mixtures'), einem generischen Konzept für breitbandige adaptive MIMO-Filterung [7, 5], basierend auf der Technik der independent component analysis (ICA), siehe z.B., [1].

2 Adaptive MIMO-Signalverarbeitung basierend auf TRINICON

In diesem Abschnitt geben wir zunächst einen kurzen Überblick über die wesentlichen Elemente von TRINICON für die Koeffizientenadaption. Dabei beschränken wir hier die Präsentation auf einfache gradientenbasierte Koeffizientenupdates im Zeitbereich.

2.1 Optimierungskriterium

Es existieren verschiedene Ansätze, um die Entmischmatrix $\mathbf{W}$ durch Ausnutzung der folgenden grundlegenden Quellsignaleigenschaften zu schätzen [1], welche alle in TRINICON vereint werden:
(i) Nichtgaußheit wird ausgenutzt durch die Verwendung von Statistik höherer Ordnung für ICA. Die minimierung der Transinformation zwischen den Ausgangssignalen kann als allgemeinster Ansatz für Separationsprobleme angesehen werden [1]. Um darüberhinaus einen auch für inverse Probleme geeigneten Schätzer zu erhalten, nutzen wir die sog. Kullback-Leibler-Divergenz (KLD) [9] zwischen einer bestimmten Wunsch-Verbunddichte (welche letztlich ein hypothetisiertes stochastisches Quellenmodell representiert, wie weiter unten gezeigt) und der Verbunddichte der tatsächlich geschätzten Ausgangssignale.
(ii) Nichtweissheit wird ausgenutzt durch simultane Minimierung von Ausgangs-Kreuzbeziehungen über mehrere zeitliche Verschiebungen. Wir betrachten deshalb multivariate Wahrscheinlichkeitsdichten, d.h. Dichten, welche ein Gedächtnis über D zeitliche Verschiebungen aufweisen.
(iii) Nichtstationarität wird ausgenutzt durch simultane Minimierung von Ausgangs-Kreuzbeziehungen zu unterschiedlichen Zeitpunkten. Wir nehmen Ergodizität innerhalb Blöcken der Länge N an, so dass der Scharmittelwert durch Zeitmittelwerte über diese Blöcke ersetzt werden kann.
In diesem Abschnitt formulieren wir dieses Konzept ohne Beschränkung der Allgemeinheit für $Q = P$. In der Praxis ist dabei ein Variieren der Zahl der aktuell gleichzeitig aktiven Quellen erlaubt, und nur die Bedingungen $Q \leq P$ (für reine Separation) bzw. $Q < P$ (für Entfaltung) müssen eingehalten werden.
Um einen Algorithmus für breitbandige Verarbeitung von Faltungsmixturen einzuführen, formulieren wir zuerst die Faltung des FIR-Entmischsystems der Länge L in der folgenden Matrizendarstellung [5]:

$$\mathbf{y}(n) = \mathbf{W}^{\mathrm{T}}\mathbf{x}(n), \tag{1}$$

wobei n den Zeitindex bezeichnet, und

$$\begin{aligned}
\mathbf{x}(n) &= [\mathbf{x}_1^{\mathrm{T}}(n), \ldots, \mathbf{x}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, && (2)\\
\mathbf{y}(n) &= [\mathbf{y}_1^{\mathrm{T}}(n), \ldots, \mathbf{y}_P^{\mathrm{T}}(n)]^{\mathrm{T}}, && (3)\\
\mathbf{x}_p(n) &= [x_p(n), \ldots, x_p(n-2L+1)]^{\mathrm{T}}, && (4)\\
\mathbf{y}_q(n) &= [y_q(n), \ldots, y_q(n-D+1)]^{\mathrm{T}}. && (5)
\end{aligned}$$

Der Parameter D in (5), $1 \leq D < L$, bezeichnet die Zahl der berücksichtigten zeitlichen Verschiebungen, um die Nichtweissheit der Quellensignale auszunutzen, wie weiter unten gezeigt. $\mathbf{W}_{pq}$, $p = 1, \ldots, P$, $q = 1, \ldots, P$ bezeichnet *Sylvestermatrizen* der Dimension $2L \times D$, welche alle Koeffizienten der entsprechenden Filter in jeder Spalte durch sukzessive Verschiebung beinhalten, d.h. die erste Spalte lautet $\left[\mathbf{w}_{pq}^{\mathrm{T}}, 0, \ldots, 0\right]^{\mathrm{T}}$, die zweite Spalte $\left[0, \mathbf{w}_{pq}^{\mathrm{T}}, 0, \ldots, 0\right]^{\mathrm{T}}$, usw. Schliesslich kombiniert die $2PL \times PD$-Matrix $\mathbf{W}$ alle Sylvestermatrizen $\mathbf{W}_{pq}$.
Basierend auf der KLD wurde die folgende Kostenfunktion in [5] eingeführt, welche alle drei grundlegenden Signaleigenschaften (i)-(iii) berücksichtigt:

$$\mathcal{J}(m, \mathbf{W}) = -\sum_{i=0}^{\infty} \beta(i,m) \frac{1}{N} \sum_{j=iN_L}^{iN_L+N-1} \left\{ \log(\hat{p}_{s,PD}(\mathbf{y}(j))) - \log(\hat{p}_{y,PD}(\mathbf{y}(j))) \right\}, \tag{6}$$

wobei $\hat{p}_{s,PD}(\cdot)$ und $\hat{p}_{y,PD}(\cdot)$ die angenommene, bzw. die geschätzte PD-variate Quellenmodell- (d.h. Wunsch-) Dichte und die Ausgangsdichte darstellen. Der Index m bezeichnet den Blockindex für einen Block von N Ausgangswerten, welche jeweils relativ zum vorhergehenden Block um L Werte verschoben sind. Weiterhin ist β eine Fensterfunktion, welche die Formulierung von Online-, Offline-, oder Block-Online-Algorithmen erlaubt [7].

2.2 Gradientenbasiertes Koeffizientenupdate

Aus Gründen der Kürze und Einfachheit konzentrieren wir uns in diesem Aufsatz auf iterative gradientenbasierte Block-Online-Koeffizientenupdates im euklidischen Raum, welche in der allgemeinen Form

$$\begin{aligned} \check{\mathbf{W}}^0(m) &:= \check{\mathbf{W}}(m-1), && \text{(7a)}\\ \check{\mathbf{W}}^\ell(m) &= \check{\mathbf{W}}^{\ell-1}(m) - \mu\Delta\check{\mathbf{W}}^\ell(m), \;\; \ell = 1,\ldots,\ell_{\max}, && \text{(7b)}\\ \check{\mathbf{W}}(m) &:= \check{\mathbf{W}}^{\ell_{\max}}(m) && \text{(7c)} \end{aligned}$$

geschrieben werden können. Dabei ist μ ein Schrittweitenparameter, und der hochgestellte Index ℓ bezeichnet einen Iterationsparameter, um mehrfache Iterationen ($\ell = 1,\ldots,\ell_{\max}$) innerhalb eines jeden Blocks m zu ermöglichen. Das nach unten zeigende Symbol auf $\mathbf{W}$ in (7) dient der Unterscheidung der der *kompakten* $PL\times Q$ Entmischmatrix $\check{\mathbf{W}}$, welche es letztlich zu optimieren gilt, von der entsprechenden größeren Sylvester matrix $\mathbf{W}$ in der Kostenfunktion. Die matrix $\check{\mathbf{W}}$ besteht aus der ersten Spalte jeder Untermatrix $\mathbf{W}_{pq}$ ohne die L Nullen.

Bei der expliziten Berechnung des Gradienten von $\mathcal{J}(m,\mathbf{W})$ bezüglich $\check{\mathbf{W}}$ stößt man offensichtlich auf das Problem der unterschiedlichen Matrizenformulierungen $\mathbf{W}$ und $\check{\mathbf{W}}$. Die grösseren Dimensionen von $\mathbf{W}$ sind eine direkte Konsequenz der Einbeziehung der Signaleigenschaft der Nichtweissheit durch die Wahl $D > 1$. Die strenge Unterscheidung zwischen diesen unterschiedlichen Matrixstrukturen ist ein wesentlicher Aspekt des allgemeinen TRINICON-Konzepts und führt zu einem wichtigen Element im Koeffizientenupdate, dem sog. *Sylvester constraint* ($\mathcal{SC}$), formal eingeführt in [7], dessen tatsächliche Implementierungsweise grundlegend für die Eigenschaften des resultierenden Algorithms' sind. Mittels des Sylvesterconstraint-Operators kann das Gradientenabstiegs-Update geschrieben werden als

$$\Delta\check{\mathbf{W}}^\ell(m) = \mathcal{SC}\left\{\nabla_{\mathbf{W}}\mathcal{J}(m,\mathbf{W})\right\}|_{\mathbf{W}=\mathbf{W}^\ell(m)}. \tag{8}$$

Abhängig von der jeweiligen Realisierung von ($\mathcal{SC}$) können wir sowohl aus der Literatur bereits bekannte, als auch neue und verbesserte Adaptionsalgorithmen spezifizieren [10]. In [2] wurde eine explizite Formulierung eines *generischen* Sylvesterconstraints hergeleitet, um das Konzept weiter zu formalisieren.

Es kann gezeigt werden [3], dass wir durch die Berechnung des Gradienten von $\mathcal{J}(m)$ bzgl. der Entmischfiltermatrix $\check{\mathbf{W}}(m)$ entsprechend (8) die folgende generische gradientenabstiegs-basierte TRINICON-Updateregel erhalten:

$$\Delta\check{\mathbf{W}}^\ell(m) = \frac{1}{N}\sum_{i=0}^{\infty}\beta(i,m)\,\mathcal{SC}\left\{\sum_{j=iN_L}^{iN_L+N-1}\left[\mathbf{x}(j)\boldsymbol{\Phi}_{s,PD}^{\mathrm{T}}(\mathbf{y}(j)) - \left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\right)^{+}\right]\right\}, \tag{9a}$$

mit $\cdot^+$ als der Bezeichung für die Pseudoinverse einer Matrix, und mit der generalisierten Score-Function

$$\boldsymbol{\Phi}_{s,PD}(\mathbf{y}(j)) = -\frac{\partial\log\hat{p}_{s,PD}(\mathbf{y}(j))}{\partial\mathbf{y}(j)} - \frac{1}{N}\sum_{r}\sum_{i_1,i_2,\ldots}\frac{\partial\mathcal{G}^{(r)}_{s,i_1,i_2,\ldots}}{\partial\mathbf{y}}\sum_{j=iN_L}^{iN_L+N-1}\frac{\partial\hat{p}_{s,PD}}{\partial\mathcal{Q}^{(r)}_{s,i_1,i_2,\ldots}}, \tag{9b}$$

resultierend aus dem hypothetisierten Quellenmodell $\hat{p}_{s,PD} = \hat{p}_{s,PD}(\mathbf{y},\boldsymbol{\mathcal{Q}}_s^{(1)},\boldsymbol{\mathcal{Q}}_s^{(2)},\ldots)$ mit bestimmten stochastischen Modellparametern $\boldsymbol{\mathcal{Q}}_s^{(r)}$, $r = 1,2,\ldots$ (Die kaligraphischen Symbole bezeichnen mehrdimensionale Arrays), gegeben durch ihre Elemente $\mathcal{Q}^{(r)}_{s,i_1,i_2,\ldots}$ in der generischen Form $\mathcal{Q}^{(r)}_{s,i_1,i_2,\ldots}(i) = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\mathcal{G}^{(r)}_{s,i_1,i_2,\ldots}(\mathbf{y}(j))\right\}$ mit bestimmten nichtlinearen Funktionen $\mathcal{G}^{(r)}_{s,i_1,i_2,\ldots}(\mathbf{y})$, $r = 1,2,\ldots$. Ein wohlbekannter Spezialfall einer solchen Parametrisierung ist die Schätzung der Korrelationsmatrix $\mathbf{R}_{\mathbf{yy}}(i) = \frac{1}{N}\sum_{j=iN_L}^{iN_L+N-1}\left\{\mathbf{y}(j)\mathbf{y}^{\mathrm{T}}(j)\right\}$. Die Filterkoeffizienten und die stochastischen Modellparameter werden in alternierender Weise geschätzt.

3 Anwendung von TRINICON auf blinde Enthallung

Das hypothetisierte Quellenmodell $\hat{p}_{s,PD}(\cdot)$ in (9b) wird entsprechend des gewünschten Signalverarbeitungsproblems gewählt. Beispielsweise liefert eine Faktorisierung von $\hat{p}_{s,PD}(\cdot)$ bezüglich der Quellen die

Verfahren der BSS [7, 5], i.e.,

$$\hat{p}_{s,PD}(\mathbf{y}(j)) \overset{(\mathrm{BSS})}{=} \prod_{q=1}^{P} \hat{p}_{y_q,D}(\mathbf{y}_q(j)), \tag{10}$$

während eine komplette Faktorisierung bzgl. der PD Dimensionen der multivariaten Dichte $\hat{p}_{s,PD}(\mathbf{y}(j))$ zum traditionellen, sogenannten MultiChannel Blind Deconvolution (MCBD)-Ansatz führen. D.h., gewöhnliche ICA-basierte MCBD-Algorithmen (z.B. [6]) nehmen *i.i.d.* -*Quellenmodelle* an. In anderen Worten: Zusätzlich zur Trennung von statistich unabhängigen Quellen führen MCBD-Algorithmen auch zu einem zeitlichen *Whitening* der Ausgangssignale, so dass dieser Ansatz für Sprach- und Audiosignale nicht direkt anwendbar ist.

Signalquellen, welche nicht i.i.d. sind, sollen am Ausgang der blinden adaptiven Verarbeitung nicht i.i.d. werden. Deshalb müssen ihre zeitlichen statistischen Abhängigkeiten geschützt werden. Im Adaptionsalgorithms muss also unterschieden werden zwischen den statistischen Abhängigkeiten innerhalb der Quellensignale und den statistischen Abhängigkeiten, welche durch das Mischsystem $\check{\mathbf{H}}$, d.h., den halligen Raum, eingeführt wurden, so dass nur der Einfluss der Raumakustik minimiert wird. Wir bezeichnen die entsprechende Verallgemeinerung des traditionellen MCBD-Ansatzes als *MultiChannel Blind Partial Deconvolution* (MCBPD) [5, 3]. Die Gleichungen (9) beinhalten bereits inhärent ein statistisches Quellenmodell (Signaleigenschaften (i)-(iii) in Abschnitt 2), ausgedrückt durch die multivariaten Dichten, und bieten deshalb bereits alle notwendigen Voraussetzungen für den MCBPD-Ansatz.

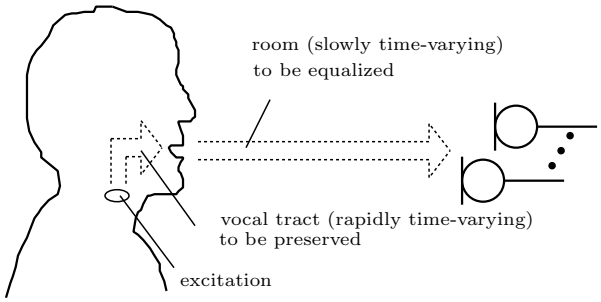

Abbildung 2 - Illustration der Sprach-Enthallung als eine MCBPD-Anwendung (nach [5]).

Für die Unterscheidung zwischen dem erzeugenden System der Quellensignale und der Raumakustik können wir wiederum alle drei der in Abschnitt 2 erwähnten grundlegenden Signaleigenschaften ausnutzen:

(i) Nichtweissheit. Die Struktur der Autokorrelation der Sprachsignale kann berücksichtigt werden. Während die Raumakustik alle Nebendiagonalen der $PD \times PD$-Ausgangskorrelationsmatrix $\mathbf{R}_{\mathbf{yy}}$ beeinflusst, konzentriert sich die Auswirkung des Vokaltrakts in den wenigen ersten Nebendiagonalen um die Hauptdiagonale. Im einfachsten Falle werden nun diese ersten Nebendiagonalen von $\mathbf{R}_{\mathbf{yy}}$ übernommen als Wunsch-Korrelationsmatrix mit einer Bandstruktur, wie in [5] vorgeschlagen. Es ist zu beachten, dass es hier eine enge Verbindung zu den linearen Prädiktionstechniken gibt, welche uns Richtlinien für die Zahl der zu schützenden zeitlichen Verschiebungen liefern [3].

(ii) Nichtstationarität. Das Sprachproduktionssystem und die Raumakustik unterscheiden sich auch in ihrer Zeitvarianz (Abb. 2) nach [5]. Im Gegensatz zur Raumakustik, welche während des Adaptionsprozesses als (nahezu) konstant angenommen wird, ist das Sprachsignal nur kurzzeitstationär [12], was durch ein zeitvariantes Sprachproduktionsmodell modelliert wird. Typischerweise werden für die Dauer des Stationaritätsintervalls etwa 20ms angenommen [12]. Wir stellen deshalb die Blocklänge N, und vorzugsweise in der Praxis auch die Blockverschiebung N_L, im Optimierungskriterium (6) mit den Modellparameterschätzungen $\boldsymbol{\mathcal{Q}}_s^{(r)}(i)$ und in den entsprechenden Koeffizientenupdates (9) auf die angenommene Dauer des Stationaritätsintervalls ein.

(iii) Nichtgaußheit. Sprache ist ein wohlbekanntes Beispiel für supergaußsche Signale. Aufgrund der Faltungssumme, welche in unserer Anwendung die Filterung durch die Raumakustik beschreibt, tendieren die Wahrscheinlichkeitsdichten der aufgenommenen Mikrophonsignale etwas näher zu einer Gauß-Charakteristik. Daher besteht eine weitere Strategie darin, die Nichtgaußheit der Ausgangssignale des Entmischsystems zu maximieren (soweit es durch das MIMO FIR-Filter möglich ist), z.B. [13, 14, 15]. Diese Strategie wird z.B. mittels der *Kurtosis* $\hat{\kappa}_{4,y_q} = \hat{E}\left\{y_q^4\right\} - 3\hat{\sigma}_{y_q}^4$ durchgeführt, welche ein weitverbreitetes Distanzmaß der Nichtgaußheit darstellt ($\hat{\kappa}_{4,y_q} > 0$ deutet auf eine supergaußsche Dichte hin und $\hat{\kappa}_{4,y_q} < 0$ auf eine subgaußsche Dichte).

Das im folgenden Abschnitt vorgestellte multivariate stochastische Sprachsignalmodell berücksichtigt alle diese Eigenschaften.

4 Ein multivariates Signalmodell für TRINICON-basierte Enthallung und Separation

Zwei unterschiedliche Entwicklungen von Funktionen werden gewöhnlich genutzt, um eine parametrisierte Darstellung von Wahrscheinlichkeitsdichten zu erhalten, welche nur leicht von der Gaußdichte abweichen (oft als *nahezu Gaußsche Dichten* bezeichnet): die Edgeworth- und die Gram-Charlier-Entwicklungen, z.B. [1]. Diese führen zu sehr ähnlichen Näherungen, weshalb wir hier nur die Gram-Charlier-Entwicklung betrachten. Diese Entwicklungen basieren auf den sog. Tschebyscheff-Hermite-Polynomen $P_{\mathrm{H},n}(\cdot)$. Ein Vorteil dieser Darstellung ist, dass die entsprechenden Entwicklungskoeffizienten direkt mit bekannten stochastischen Größen basierend auf Kumulanten höherer Ordnung in Verbindung gebracht werden können, wie z.B. der im vorhergehenden Abschnitt erwähnten Kurtosis.

Um basierend auf dieser Darstellung allgemeine Koeffizientenupdateregeln zu erhalten, betrachten wir eine multivariate Verallgemeinerung der Gram-Charlier-Entwicklung. Wie in [3] gezeigt, kann diese ausgedrückt werden als

$$\begin{aligned}\hat{p}_{y_q,D}(\mathbf{y}_q(j)) &= \frac{1}{\sqrt{(2\pi)^D \det \mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}(i)}} \mathrm{e}^{-\frac{1}{2}\mathbf{y}_q^{\mathrm{T}}(j)\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i)\mathbf{y}_q(j)} \\ &\quad \cdot \sum_{n_1=0}^{\infty} \cdots \sum_{n_D=0}^{\infty} a_{n_1\cdots n_D,p} \prod_{d=1}^{D} P_{\mathrm{H},n_d}\left(\left[\mathbf{L}_q^{-1}(i)\mathbf{y}_q(j)\right]_d\right)\end{aligned}$$

mit den Enwicklungskoeffizienten

$$a_{n_1\cdots n_D,q} = \hat{E}\left\{\prod_{d=1}^{D} \frac{1}{n_d!} P_{\mathrm{H},n_d}\left(\left[\mathbf{L}_q^{-1}(i)\mathbf{y}_q(j)\right]_d\right)\right\},$$

wobei $\mathbf{L}_q$ durch die Choleskyzerlegung $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q} = \mathbf{L}_q^{\mathrm{T}}\mathbf{L}_q$ erhalten wird. (Man beachte, dass $\sqrt{\mathbf{y}_q^{\mathrm{T}}\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}\mathbf{y}_q} = \|\mathbf{L}_q^{-1}\mathbf{y}_q\|_2$.)

In diesem Beitrag betrachten wir weiterhin einen wichtigen Spezialfall dieses allgemeinen multivariaten Modells, welches besonders nützlich für die Sprachverarbeitung ist. In diesem Fall wird die inverse Kovarianzmatrix $\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q} = (\mathbf{L}_q^{\mathrm{T}}\mathbf{L}_q)^{-1}$ zunächst faktorisiert als [11]

$$\mathbf{R}^{-1}_{\mathbf{y}_q\mathbf{y}_q}(i) = \mathbf{A}_q(i)\mathbf{\Sigma}^{-1}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(i)\mathbf{A}_q^{\mathrm{T}}(i), \tag{12}$$

wobei $\mathbf{A}_q(i)$ eine untere Einheitsdreiecksmatrix der Dimension $D \times D$ (d.h., ihre Elemente auf der Hauptdiagonalen sind gleich 1) ist und $\mathbf{\Sigma}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(i)$ eine Diagonalmatrix bezeichnet [11]. Die untere Einheitsdreiecksmatrix $\mathbf{A}_q(i)$ kann interpretiert werden als eine (zeitvariante) Faltungsmatrix eines Whiteningfilters. Aus berechnungstechnischen Gründen ist es deshalb praktisch, das Signal y_q als einen autoregressiven (AR) Prozess mit zeitvarianten AR-Koeffizienten $a_{q,k}(n)$ und Restsignal $\tilde{y}_q(n)$ zu modellieren. Formal wird die obengenannte Ausnutzung der Nichtweissheit zur Unterscheidung zwischen der zeitlichen Färbung der Quellen und des Mischsystems erreicht durch eine Entkopplung der Ordung des AR-Prozesses, d.h., der Prädiktionsordung $0 \leq n_A \leq D-1$, von der Dimension D der Korrelationsmatrix $\mathbf{R}_{\mathbf{y}_q\mathbf{y}_q}$, d.h.,

$$y_q(n) = -\sum_{k=1}^{n_A} a_{q,k}(n) y_q(n-k) + \tilde{y}_q(n). \tag{13}$$

Die Matrizen $\mathbf{A}_q$ und $\mathbf{\Sigma}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}$ erhält man dann durch die D Spaltenvektoren $[1, a_{q,1}(n), a_{q,2}(n), \cdots, a_{q,n_A}(n), 0, \cdots, 0]^{\mathrm{T}}$, $[0, 1, a_{q,1}(n-1), \cdots, a_{q,n_A-1}(n-1), a_{q,n_A}(n-1), \cdots, 0]^{\mathrm{T}}$, usw., und

$$\mathbf{\Sigma}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q} = \mathrm{Diag}\left\{\hat{\sigma}^2_{\tilde{y}_q}(n), \ldots, \hat{\sigma}^2_{\tilde{y}_q}(n-D+1)\right\}. \tag{14}$$

Nun kann das multivariate stochastische Signalmodell umgeschrieben werden, indem die *Vorfiltermatrix* $\mathbf{A}_q$ in die Datenterme hineingezogen wird, d.h.,

$$\tilde{\mathbf{y}}_q := \mathbf{A}_q^{\mathrm{T}}\mathbf{y}_q = [\tilde{y}_q(n), \tilde{y}_q(n-1), \ldots, \tilde{y}_q(n-D+1)]^{\mathrm{T}}. \tag{15}$$

Weiterhin, indem wir die weissgemachten Elemente des Vektors $\tilde{\mathbf{y}}_q$ als i.i.d. annehmen (was in der Praxis eine weitverbreitete Annahme in der AR-Modellierung ist), so dass die Entwicklungskoeffizienten $a_{n_1 \cdots n_D, q}$ faktorisiert werden, erhalten wir mit $\mathbf{L}_q(i) = \mathrm{Diag}\left\{\frac{1}{\hat{\sigma}_{\tilde{y}_q}(j)}, \ldots, \frac{1}{\hat{\sigma}_{\tilde{y}_q}(j-D+1)}\right\} \mathbf{A}_q^{\mathrm{T}}(i)$ und (15) eine kompaktere Modelldarstellung. Die entsprechende Näherung der vierten Ordnung einer mittelwertfreien und nahezu Gaußschen Dichte beinhaltet unmittelbar die bekannten Größen *Schiefe* (skewness) und *Kurtosis*, wobei in unserem Kontext, wie oben erwähnt, letztere die wichtigste statistische Größe höherer Ordnung ist. Allgemein weisen Sprachsignale supergaußsche Dichten auf, deren Kumulanten dritter Ordnung im Vergleich zu ihren Kumulanten vierter Ordung vernachlässigbar sind. Daher erhalten wir, indem wir zusätzlich zur Statistik zweiter Ordnung nur den Term vierter Ordnung berücksichtigen,

$$\hat{p}_{y_q,D}(\mathbf{y}_q(j)) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\ \hat{\sigma}^2_{\tilde{y}_q}(j-d+1)}} \mathrm{e}^{-\frac{\tilde{y}_q^2(j-d+1)}{2\hat{\sigma}^2_{\tilde{y}_q}(j-d+1)}} \left(1 + \frac{\hat{\kappa}_{4,\tilde{y}_q}}{4!\sigma^4_{\tilde{y}_q}(j-d+1)}\ P_{\mathrm{H},n_d}\left(\frac{\tilde{y}_q(j-d+1)}{\hat{\sigma}_{\tilde{y}_q}(j-d+1)}\right)\right).$$

Durch Ausnutzung der Nahezu-Gaußheit mittels der Näherung $\log(1+\epsilon) \approx \epsilon$ in der logarithmierten Darstellung dieser Dichte in (9b), und unter Einbeziehung von $P_{\mathrm{H},4}\left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right) = \left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right)^4 - 6\left(\frac{\tilde{y}_q}{\hat{\sigma}_{\tilde{y}_q}}\right)^2 + 3$ erhalten wir nach einer einfachen Berechnung die entsprechende TRINICON-Koeffizeintenupdateregel basierend auf (9). Eine effiziente Realisierung, welche nach wie vor alle drei grundlegenden Signaleigenschaften (i)-(iii) ausnutzt, wie oben diskutiert, erhalten wir mittels der sog. *Korrelationsmethode*, d.h., durch Annahme einer globalen Nichtstationarität der Quellensignale, aber einer Kurzzeitstationarität innerhalb jedes Blocks, wie aus der linearen Prädiktion bekannt [12]. Nach Nutzung der expliziten Formulierung des generischen Sylvesterconstraints entsprechend [2] führen diese Schritte schließlich zu der MIMO-Koeffizientenupdateregel [3]

$$\begin{aligned}
\check{\mathbf{w}}^{\ell}_{pq}(m) = \check{\mathbf{w}}^{\ell-1}_{pq}(m) - \frac{\mu}{N}\sum_{i=0}^{\infty}\beta'(i,m)\Bigg[&\frac{\sum_{j=iN'_L}^{iN'_L+N-1}\check{\tilde{\mathbf{x}}}^{(q)}_p(j)\tilde{y}_q(j)}{2\hat{\sigma}^2_{\tilde{y}_{q,i}}} \\
&- \left(\frac{\sum_{j=iN'_L}^{iN'_L+N-1}\tilde{y}^4_q(j)}{3\hat{\sigma}^4_{\tilde{y}_{q,i}}} - 1\right)\cdot\left(\frac{\sum_{j=iN'_L}^{iN'_L+N-1}\check{\tilde{\mathbf{x}}}^{(q)}_p(j)\tilde{y}^3_q(j)}{\hat{\sigma}^4_{\tilde{y}_{q,i}}} - \frac{\sum_{j=iN'_L}^{iN'_L+N-1}\check{\tilde{\mathbf{x}}}^{(q)}_p(j)\tilde{y}_q(j)\sum_{j=iN_L}^{iN_L+N-1}\tilde{y}^4_q(j)}{\hat{\sigma}^6_{\tilde{y}_{q,i}}}\right)\Bigg] \\
&+\mu\sum_{i=0}^{\infty}\beta(i,m)\left[\mathcal{SC}\left\{\left(\left(\mathbf{W}^{\ell-1}(m)\right)^{\mathrm{T}}\right)^{+}\right\}\right]_{pq},
\end{aligned} \tag{17}$$

$p = 1, \ldots, P,\ \ q = 1, \ldots, P$. Analog zur Definition (15) bezeichnet das Symbol $\check{\tilde{\mathbf{x}}}^{(q)}_p$ einen Spaltenvektor von gefilterten Sensorsignalen x_p entsprechend der oben eingeführten Vorfiltermatrix $\mathbf{A}_q$.

In anderen Worten, diese Updateregel kann als sog. *Filtered-x-Algorithmus* interpretiert werden, da sowohl der Eingangssignalvektor (d.h. die Mikrophonsignale), als auch die Ausgangssignale als gefilterte Versionen im Update erscheinen. Daraus erhalten wir unmittelbar Abb. 3. Während $\mathbf{W}$ idealerweise die das raumakustische Mischsystem $\mathbf{H}$ invertiert, invertiert das lineare Prädiktionsfilter (bzw. der Satz von linearen Prädiktionsfiltern) $\mathbf{A}$ aus dem stochastischen Quellenmodell idealerweise das/die Sprachproduktionssystem(e) der Quelle(n). Die Koeffizienten $\mathbf{W}$ und $\mathbf{A}$ werden in alternierender Weise geschätzt, wie die Schätzung der anderen stochastischen Modellparameter. Es ist zu beachten, dass (in Übereinstimmung mit dem bekannten Filtered-x-Konzept) der gefilterte Eingangsvektor $\check{\tilde{\mathbf{x}}}^{(q)}_p$ mittels der Filterkoeffizienten aus der linearen Prädiktionsanalyse (LP) der *Ausgangssignale* y_p berechnet werden, d.h., die Koeffizienten der Ausgangsanalysefilter werden in die Eingangstransformationsfilter kopiert.

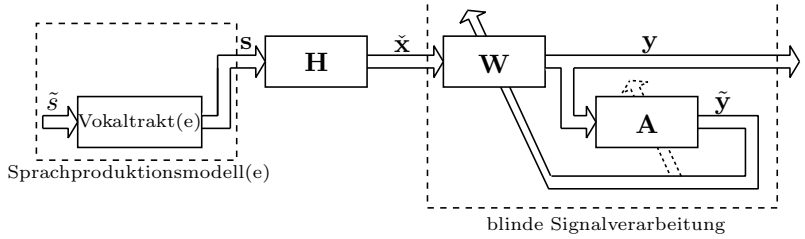

Abbildung 3 - Inversion des/der Sprachproduktionsmodells/e innerhalb der blinden Signalverarbeitung und Interpretation als Filtered-x-Adaption.

5 Experimente

Die Experimente wurden mit $Q = 2$ Sprachsignalen durchgeführt (je ein männl. u. weibl. Sprecher, Abtastrate 16kHz), welche mit gemessenen Impulsantworten eines realen Raums mit einer Nachhallzeit $T_{60} \approx 700$ms gefaltet wurden. Ein lineares Mikrophonarry mit vier Elementen ($P = 4$) und einem Abstand zwischen den Mikrophonen von jeweils 16cm wurde verwendet. Die Einfallswinkel der Sprachsignale betrugen $\pm 24^o$ relativ zur Normalenebene der Arrayachse, und der Abstand zwischen den Sprechern und dem Zentrum des Mikrophonarrays war 165cm.

Als ein Gütemaß für die Evaluierung der Enthallungsergebnisse verwenden wir hier das *signal-to-reverberation ratio* (SRR), welches das Leistungsverhältnis zwischen dem Direktschall und dem Beitrag durch den Nachhall misst. Für Sprachsignale werden hier die ersten 50ms nach dem Hauptpeak der Impulsantworten ebenfalls zum Direktschall hinzugezählt (kritische Verzögerungszeit n_{50}, welche zur Sprachverständlichkeit beiträgt). Im Hinblick auf eine etwas höhere Korrelation zur empfundenen Qualität bei Hörversuchen verwenden wir in unseren Experimenten die Verbesserung des sog. *segmental SRR*. Im Falle von mehreren aktiven Quellensignalen wird die Evaluierung der zusätzlichen Fähigkeit der MCBPD zur Quellentrennung mittels der Verbesserung des sog. (segmental) *signal-to-interference ratio* (SIR) an den Ausgängen analog zum (segmental) SRR vorgenommen.

Unsere Simulationen basieren auf dem Koeffizientenupdate (17) mittels der Korrelationsmethode. Wir wählten $L = 3000$, the Blocklänge $N = N'_L = 320$ entsprechend eines Stationaritätsintervalls von 20ms, und $n_A = 32$.

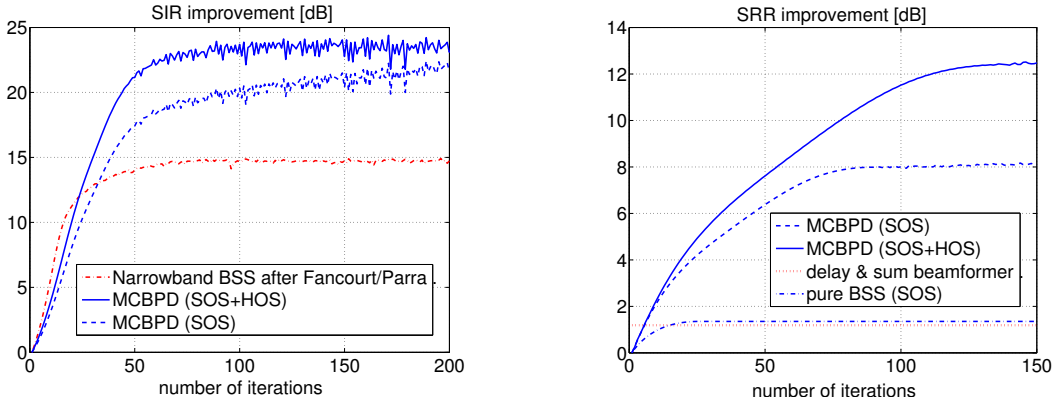

Abbildung 4 - SIR- und SRR-Verbesserung duch die MIMO-basierte MCBPD.

Abbildung 4 zeigt die SIR-Verbesserung (d.h., die Quellentrennung an den Ausgängen) und die SRR-Verbesserung für Offline-Adaption, d.h., $\beta(i,m) = \beta(i)$ in (6) (und daher $\beta'(i,m) = \beta'(i)$ in (17)) entspricht einer Rechteckfensterfunktion über die gesamte verfügbare Signallänge, und die äussere Summe in (6) and (17) wird zu einer Summation der Beiträge aller Blöcke mit einheitlicher Gewichtung. Die SIR- und SRR-Kurven wurden gemittelt aus den Beiträgen der beiden Quellen. Es ist zu erkennen, dass die rein auf Statistik zweiter Ordung basierende Optimierung (SOS, nur der erste Term in den eckigen Klammern in (17) berücksichtigt) eine relativ schnelle initiale Konvergenz aufweist. Weitere Betrachtungen [3] zeigen, dass der rein auf der Kurtosis basierende Ansatz (nur der zweite Term in den eckigen Klammern in (17)) schliesslich eine höhere SRR-Verbesserung liefern kann, welche jedoch auf Kosten einer langsameren initialen Konvergenz geht. Durch Ausnutzung aller verfügbarer statistischer Signaleigenschaften (SOS+HOS, beide Terme in den eckigen Klammern in (17) wurden genutzt) vereint das TRINICON-Framework die Vorteile der ersten beiden Ansätze. Diese Synergien äussern sich sowohl in der Separations- als auch in der Enthallungsleistung. Als eine Referenz haben wir auch die SIR-Konvergenzkurve des populären BSS-Algorithms nach Fancourt und Parra [8] hinzugefügt, welcher ausschliesslich auf SOS und einem schmalbandigem Ansatz basiert. Die Referenzkurve für einen reinen Separationsalgorithmus [7], basierend auf SOS (als Spezialfall von (17) mit $n_A = L - 1$, $N = L$ und der Verwendung nur des ersten Terms in den eckigen Klammern) im SRR-Plot und der Vergleich mit einem konventionellen Delay-and-Sum-Beamformer bestätigen die hohe Effektivität der in diesem Beitrag vorgestellten MCBPD-Erweiterung.

6 Zusammenfassung

Basierend auf dem TRINICON-Framework für breitbandige adaptive MIMO-Filterung haben wir in diesem Beitrag eine Klasse von neuen Algorithmen zur blinden Enthallung und Quellentrennung vorgestellt.

Aufgrund des speziell für Sprachsignale entworfenen stochastischen Quellenmodells können die Nichtweissheit, the Nichtstationarität und die Nichtgaußheit effektiv ausgenutzt werden, was zu einer hohen Separations- und Enthallungsleistung ohne die gefürchteten Whitening-Artefakte bisheriger Verfahren führt.

Literatur

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley & Sons, Inc., New York, 2001.

[2] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON-based blind system identification with application to multiple-source localization and separation," in S. Makino, T.-W. Lee, and S. Sawada (eds.), *Blind Speech Separation*, Springer, Berlin, pp. 101-147, Sept. 2007.

[3] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in P.A. Naylor and N.D. Gaubitch (eds.), *Speech Dereverberation*, Springer, London, to appear in 2010.

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol 36, no. 2, pp. 145-152, Feb. 1988.

[5] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Montreal, Canada, vol. 3, pp. 889-892, May 2004.

[6] S. Amari et al.,"Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. IEEE Int. Workshop Signal Processing Advances in Wireless Communications*, pp. 101-107, 1997.

[7] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in Y. Huang and J. Benesty (eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer Academic Publishers, Boston, pp. 255-293, Feb. 2004.

[8] C.L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of nonstationary signals," in *Proc. Int. Workshop Neural Networks Signal Processing (NNSP)*, 2001, pp. 303-312.

[9] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley & Sons, New York, 1991.

[10] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, no. 6, pp.1260–1277, 2006.

[11] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[12] J.D. Markel and A.H.Gray, *Linear Prediction of Speech*, Springer, Berlin, 3rd edition, 1976.

[13] R.A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol 16, pp. 21-35, 1978.

[14] R.H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*, Ph.D. dissertation, Univ. of Southern California, Los Angeles, CA, May 1996.

[15] B.W. Gillespie, H.S. Malvar, and D.A.F. Florêncio,"Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE ICASSP*, Salt Lake City, UT, USA, May 2001.