Onlinefähige kombinierte blinde Quellentrennung und Enthallung von Sprachmixturen durch RLS-Optimierung

Timo Schuster, Stefan Feldes
Institut für Digitale Signalverarbeitung, Hochschule Mannheim, 68163 Mannheim, Deutschland
Email: timo.schuster4@hotmail.de, s.feldes@hs-mannheim.de

Einleitung

Menschen mit Hörbeeinträchtigungen haben es insbesondere in geräuscherfüllter Umgebung schwer eine gewünschte Audioquelle, meist eine sprechende Person, aus dem Gemisch von Störgeräuschen und anderen konkurrierenden Sprechern (Sprachmixturen) herauszuhören. Abhilfe können hier Hörgeräte schaffen, die sich in jedem der beiden Ohren befinden und einen drahtlosen Datenlink zueinander besitzen. Dies ermöglicht den Einsatz binauraler Signalverarbeitungsalgorithmen, welche die Signale von linkem und rechtem Ohr erzeugen mit dem Ziel, die Sprachverständlichkeit zu erhöhen. Zur Trennung des erwünschten Sprachsignals von den unerwünschten Signalen, wurden bereits verschiedene Methoden der Blind Source Separation (BSS) auf die Verarbeitung der Mikrofonsignale der beiden Hörkapseln angewendet. Ebenso wurden dabei Varianten entwickelt, die gemäß der hier erforderlichen Echtzeitfähigkeit online adaptieren [1][2] . Die Leistungsfähigkeit der BSS sinkt jedoch erheblich, wenn die akustischen Signale mit starken Nachhallanteilen an den Hörgeräten ankommen. Zur Reduktion des Nachhalls und damit zur Verbesserung der BSS bieten sich Verfahren der blinden Enthallung (Blind Dereverberation (BD)) an. Eine etablierte Methode ist die Multi Channel Linear Prediction (MCLP), wobei die ungewollten Nachhallkomponenten aus den Mikrofonsignalen prädiziert und von den aktuellen Mikrofonsignalen subtrahiert werden. Aktuelle Varianten der MLCP mit schneller Adaption für dynamische Sprechersituationen wurden in [3] vorgeschlagen. Die RLS-basierte Variante weist auch die für den Echtzeitbetrieb nötige Online-Fähigkeit auf. Jedoch wird dabei angenommen, dass nur eine Quelle aktiv ist. Sind mehrere Quellen aktiv, wie es in einem realen Szenario meist der Fall ist, leidet die Performance der Enthallung. Vorteilhaft ist es, so wie in [4] vorgeschlagen, BSS und BD kombiniert auszuführen: So profitiert die BSS von enthallten und die BD von getrennten Signalen. Der dort vorgestellte Algorithmus läuft im Batch-Betrieb. Für den Einsatz in Hörgeräten ist jedoch ein onlinefähiger Algorithmus erforderlich, damit der Hörgeräteträger verzögerungsfrei am Gespräch teilhaben kann. In diesem Beitrag wird daher ein RLS-basierter Algorithmus entwickelt, der eine solche kombinierte Quellentrennung und Enthallung im Online-Betrieb erlaubt.

Signalmodell

Abbildung 1 zeigt ein kombiniertes Modell der Signalerzeugung, das den Vorgang der Mischung mehrerer Sprechersignale und die anschließende Verhallung im Raum verbindet. Hierbei sei $s_i(n, f)$ das i-te reine Sprachsignal und $x_m(n, f)$ das m-te Mikrofonsignal in der Kurzzeit-Fourier-Transformation-Domäne, wobei n und f der Zeitframe-Index und Frequenzindex sind. Im Folgenden wird der Frequenzindex f fallen gelassen, da das Signalmodell in jeder Frequenzkomponente unabhängig gelten soll. Im herkömmlichen Signalmodell der BSS (siehe Abbildung 2) in der Kurzzeit-Fourier-Transformation-Ebene erfolgt die Mischung der N reinen Sprachsignale $\mathbf{s}(n) = \left[s_1(n), \ldots, s_N(n)\right]^T$ zu den gemischten Signalen $\hat{\mathbf{x}}(n)$ mit einer Mischungsmatrix $\mathbf{H}(n)$. Die aktuel-

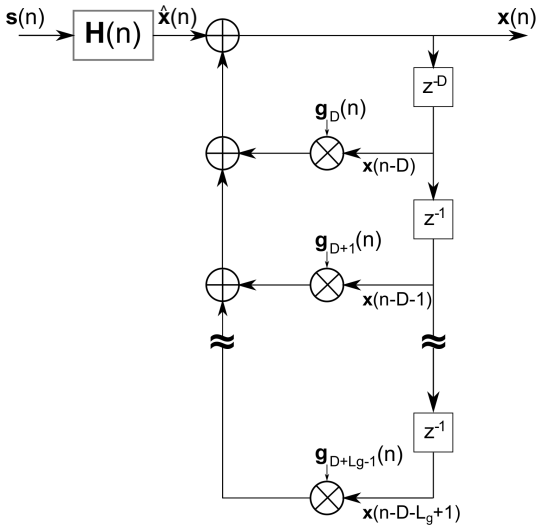

Abbildung 1: Kombiniertes Signalerzeugungsmodell

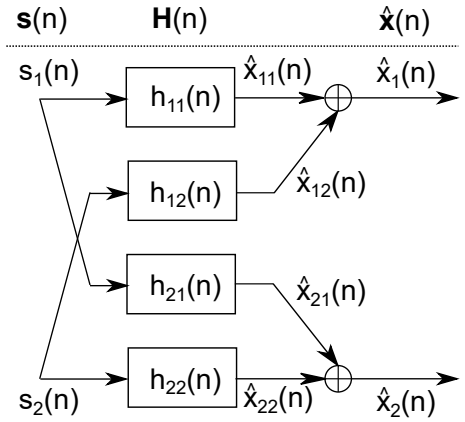

Abbildung 2: Herkömmliches Signalmischungsmodell für $M = 2$ bei BSS

len gemischten Signale sind, in diesem Modell, nur von den aktuellen Signalen des Frames und einer aktuellen Mischungsmatrix abhängig. Die Framelänge ist (in der Regel) jedoch sehr viel kleiner als die Länge des Nachhalls, mit welchem die Sprachsignale an den Mikrofonen ankommen. Somit sind die aktuellen Mikrofonsignale auch abhängig von Signalen vergangener Frames. Um den Nachhall miteinzubeziehen, wird das Signalmischungsmodell durch ein MCLP-Modell zum kombinierten Signalerzeugungsmodell aus Abbildung 1 erweitert. Es sei $\mathbf{x}(n) = \left[x_1(n), \ldots, x_M(n)\right]^T$ der Vektor der M Mikrofonsignale, der zu

$$\mathbf{x}(n) = \mathbf{d}(n) + \mathbf{u}(n) \tag{1}$$

aufgeteilt wird, wobei $\mathbf{d}(n)$ die direkte Signalkomponente und die frühen Reflektionen enthält, während $\mathbf{u}(n)$ die unerwünschten späten Reflektionen darstellt. $\mathbf{u}(n)$ kann, gemäß des MCLP-Modells, durch die Summe aus gefilterten und verzögerten Mikrofonsignalen prädiziert werden [3]:

$$\mathbf{u}(n) = \mathbf{X}_D(n)\mathbf{g}(n), \tag{2}$$

wobei

$$\underset{(M\times M^2L_g)}{\mathbf{X}_D(n)} = \begin{bmatrix} \mathbf{x}_{n-D}^T & \mathbf{0} & \cdots & \mathbf{x}_{n-D-L_g+1}^T & \mathbf{0} \\ \ddots & & & \ddots & \\ \mathbf{0} & \mathbf{x}_{n-D}^T & \mathbf{0} & \cdots & \mathbf{x}_{n-D-L_g+1}^T \end{bmatrix} \tag{3}$$

mit $\mathbf{x}_t = \mathbf{x}(t)$ ein Signalpuffer ist, welcher vergangene Werte aller Mikrofonsignale enthält. Die Verzögerung D sorgt dabei dafür, dass die Kurzzeitkorrelation und die frühen Reflektionen der Sprachsignale erhalten bleiben. Die Regressionslänge L_g gibt die Länge der späten Reflektionen an. Dabei wird der noch zu bestimmende Regressionsvektor

$$\begin{aligned} \underset{(M^2L_g\times 1)}{\mathbf{g}} = \big[&\mathbf{g}_{1,D}^T, \quad \ldots, \quad \mathbf{g}_{M,D}^T, \quad \ldots, \\ &\mathbf{g}_{1,D+L_g-1}^T, \quad \ldots, \quad \mathbf{g}_{M,D+L_g-1}^T\big] \end{aligned} \tag{4}$$

definiert, wobei $\mathbf{g}_{j,T}$ der Koeffizientenvektor für das j-te Mikrofonsignal der T-ten Verzögerung ist.

Kombinierte Optimierung im Online-Betrieb

In der kombinierten Optimierung werden die Mikrofonsignale zunächst enthallt und anschließend erfolgt eine BSS. Abbildung 3 zeigt dieses Schema. Im Batch-Betrieb lässt sich der Regressionsvektor $\mathbf{g}$ durch die Lösung einer Yule-Walker-Gleichung [5] bestimmen:

$$\begin{aligned} \mathbf{g} &= \left[\sum_{n=1}^{n_{tot}} \mathbf{X}_D^H(n)\hat{\mathbf{\Sigma}}^{-1}(n)\mathbf{X}_D(n)\right]^{-1} \\ &\quad \left[\sum_{n=1}^{n_{tot}} \mathbf{X}_D^H(n)\hat{\mathbf{\Sigma}}^{-1}(n)\mathbf{x}(n)\right] \\ &= \mathbf{Q}^{-1}\mathbf{R}, \end{aligned} \tag{5}$$

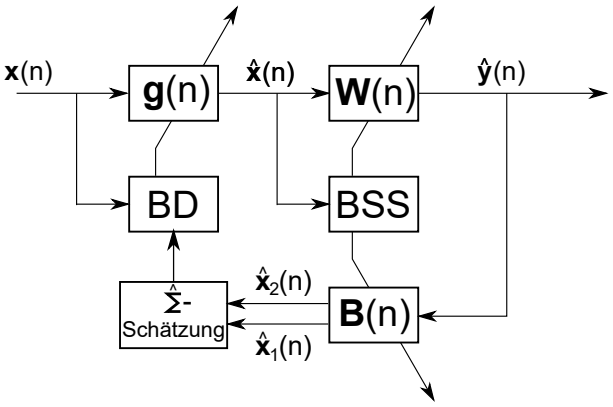

Abbildung 3: Kombinierte Optimierung für $M = 2$

wobei n_{tot} für die Gesamtzahl der Frames und $\hat{\mathbf{\Sigma}}(n)$ für die Kovarianzmatrix des n-ten Frames steht. Durch diese Batch-Optimierung wird ein, über alle Frames konstanter, Regressionsvektor ermittelt. Um den Regressionsvektor $\mathbf{g}(n)$ für den n-ten Frame adaptiv zu bestimmen und so zu einem Online-Algorithmus zu gelangen, wird die Gewichtung vergangener Frames durch den Vergessensfaktor $0 \leq \gamma \leq 1$ exponentiell reduziert:

$$\underset{(M^2L_g\times M^2L_g)}{\mathbf{Q}(n)} = \sum_{\tilde{n}=1}^{n} \gamma^{n-\tilde{n}}\mathbf{X}_D^H(\tilde{n})\hat{\mathbf{\Sigma}}^{-1}(\tilde{n})\mathbf{X}_D(\tilde{n}) \tag{6}$$

$$\underset{(M^2L_g\times 1)}{\mathbf{R}(n)} = \sum_{\tilde{n}=1}^{n} \gamma^{n-\tilde{n}}\mathbf{X}_D^H(\tilde{n})\hat{\mathbf{\Sigma}}^{-1}(\tilde{n})\mathbf{x}(\tilde{n}) \tag{7}$$

Unter zur Hilfenahme des Matrix-Inversions-Lemmas [6] folgt der RLS-Algorithmus zu

$$\begin{aligned} \mathbf{K}(n) = \mathbf{Q}^{-1}(n-1)&\mathbf{X}_D^H(n) \\ &[\gamma\hat{\mathbf{\Sigma}}(n) + \mathbf{X}_D(n)\mathbf{Q}^{-1}(n-1)\mathbf{X}_D^H(n)]^{-1} \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{g}(n) = \mathbf{g}(n-1)&+ \\ &\mathbf{K}(n)[\mathbf{x}(n) - \mathbf{X}_D(n)\mathbf{g}(n-1)] \end{aligned} \tag{9}$$

$$\mathbf{Q}^{-1}(n) = \gamma^{-1}[\mathbf{I} - \mathbf{K}(n)\mathbf{X}_D(n)]\mathbf{Q}^{-1}(n-1) \tag{10}$$

$$\mathbf{u}(n) = \mathbf{X}_D(n)\mathbf{g}(n) \tag{11}$$

$$\hat{\mathbf{x}}(n) = \mathbf{x}(n) - \mathbf{u}(n). \tag{12}$$

Die Kovarianzmatrix $\hat{\mathbf{\Sigma}}(n)$ lässt sich für $M = 2$ durch

$$\hat{\mathbf{\Sigma}}(n) = \begin{bmatrix} |\hat{x}_{11}(n)|^2 + |\hat{x}_{12}(n)|^2 & \beta \\ \beta^* & |\hat{x}_{21}(n)|^2 + |\hat{x}_{22}(n)|^2 \end{bmatrix} \tag{13}$$

$$\beta = \hat{x}_{11}(n)\hat{x}_{21}^*(n) + \hat{x}_{12}(n)\hat{x}_{22}^*(n) \tag{14}$$

bestimmen, wobei $\hat{\mathbf{x}}_{m,i}$ die getrennten und mit $h_{m,i}$ skalierten Quellsignale sind. Jedoch sind nur die Mikrofonsignale $\mathbf{x}(n)$ bekannt. Um die gesuchten Signale aus $\mathbf{x}(n)$ zu schätzen, wird die BSS angewendet. Hierzu werden der Regressionsvektor $\mathbf{g}(n-1)$ und die Entmischungsmatrix $\mathbf{W}(n-1)$ des vorangehenden Frames genutzt, um das aktuelle Mikrofonsignal $\mathbf{x}(n)$ zu den enthallten und getrennten Signalen

$$\hat{\mathbf{y}}(n) = \mathbf{W}(n)\hat{\mathbf{x}}(n) \tag{15}$$

zu überführen. Diese werden mit $\mathbf{B}(n) = \mathbf{W}^{-1}(n)$ skaliert, um $\hat{\mathbf{x}}_{m,i}$ zu schätzen. Erzielt die BSS eine optimale Trennung, dann ist auch diese Schätzung der Kovarianzmatrix optimal. Das BSS-Problem lässt sich durch Einsatz der Independent Vector Analysis lösen [1][2], jedoch kann auch ein beliebiger anderer, onlinefähiger BSS-Algorithmus verwendet werden. Die Ausgangssignale $\hat{\mathbf{y}}(n)$ sind monaural. Um binaurale Signale zu erzeugen und damit auch gleichzeitig das Skalierungsproblem zu lösen, werden diese anschließend hörgerecht skaliert [7].

Experimente

Im experimentellen Teil der Arbeit wird der hier vorgestellte Algorithmus (BSS + BD kombiniert) auf simulativ erzeugte Testsignale angewendet und dessen Performance mit der BSS ohne Enthallung und mit einer sequentiellen Kombination aus BSS mit Enthallung (BSS + BD sequentiell) verglichen. Als BSS-Algorithmus wird jeweils [2] verwendet. Die jeweiligen Einstellungen sind Tabelle 1 zu entnehmen. Die reinen Sprachsignale wur-

Parameter	**Einstellung**
Anzahl Mikrofone	$M = 2$
Abtastfrequenz	$F_s = 16\text{kHz}$
STFT-Länge	$F = 1024$
Fensterfunktion	Hanning − Fenster
Fenstershift	$J = F/4 = 256$
Regressionslänge	$L_g = 15$
Delay	$D = 2$
Vergessensfaktor (BD)	$\gamma = 0.995$
Vergessensfaktor (BSS)	$\alpha = 0.96$

Tabelle 1: Parametereinstellungen für die durchgeführten Experimente

den dem Sprachkorpus SI100 des bayerischen Archivs für Sprachsignale (BAS) [8] entnommen. Dieser besteht aus gesprochenen Aussagen verschiedener Länge von rund 100 männlichen und weiblichen Sprechern. Daraus wurden jeweils zwei Sprecher und je ein 30 Sekunden langer Ausschnitt zufällig ausgewählt. Diese wurden anschließend mit Raumimpulsantworten aus der Datenbank in [9] gefiltert, um Sprecher in einer realen Raumumgebung zu simulieren. Die Experimente wurden für Raumimpulsantworten mit den Nachhallzeiten $\text{RT}_{60} = 160\text{ms}$ und $\text{RT}_{60} = 610\text{ms}$ durchgeführt. Die beiden Sprecher wurden zwei Meter entfernt zum Mikrofonarray, welches aus zwei Mikrofonen mit Abstand 16cm besteht, im Winkel von $-45°$ bzw. $45°$ (siehe Abbildung 4) simuliert. Zur objektiven Performancemessung der Signaltrennung wurde das Signal to Interference Ratio (SIR) [10] benutzt. Hierzu wird das verarbeitete Signal in Frames von einer Sekunde Länge aufgeteilt und das SIR so über der Zeit bestimmt. Das SIR aller Ausgangssignale wird anschließend gemittelt. Um die Performance der Enthallung zu messen, wird das Direct to Reverberant Energy Ratio (DRR), welches in [11] beschrieben ist, verwendet. Wie in Abbildung 5 zu sehen ist, erzielt die kombinierte Optimierung, sowohl bei schwachem ($\text{RT}_{60} = 160\text{ms}$) und

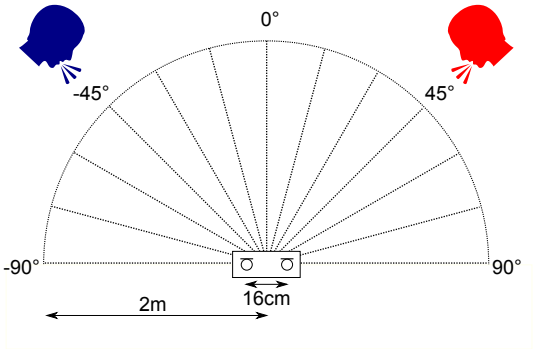

Abbildung 4: Aufbau des Experiments

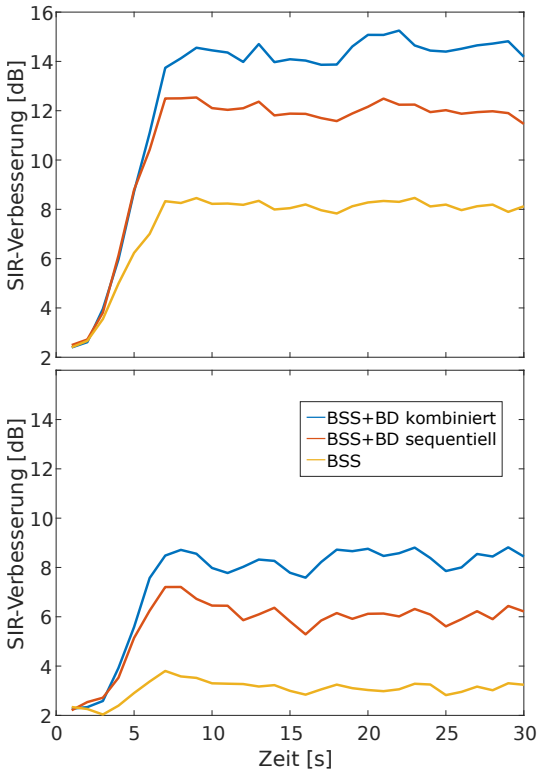

Abbildung 5: Mittlere SIR-Verbesserung für $\text{RT}_{60} = 160\text{ms}$ (oben) und $\text{RT}_{60} = 610\text{ms}$

starkem ($\text{RT}_{60} = 610\text{ms}$) Nachhall im Mittel eine Verbesserung von 6dB im Vergleich zur BSS ohne Enthallung und eine Verbesserung von 2dB verglichen mit einer sequentiellen Optimierung. Die Konvergenz tritt bei allen Algorithmen nach ca. 7 Sekunden ein. Abbildung 6 zeigt die mittlere DRR-Verbesserung. Bei beiden Nachhallzeiten zeigt die kombinierte Optimierung eine verbesserte Enthallung. Die BSS und MCLP zeigen eine Synergie zueinander und durch die Kombination wird die Performance beider Algorithmen deutlich verbessert. In einem zweiten Experiment wird die Position beider Sprecher abrupt geändert: Von $-45°$ nach $75°$ bzw. $45°$ nach $-15°$. Nach einiger Zeit stellen sich alle drei Algorithmen auf die neue Audioszenerie ein und erreichen gleiche SIR- bzw. DRR-Verbesserungen wie vor dem Positionswechsel, wie in Abbildung 7 zu sehen ist.

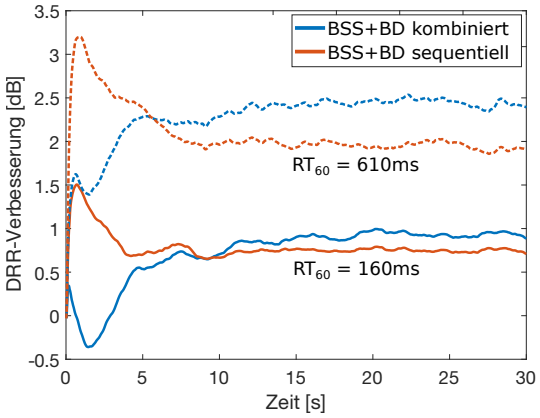

Abbildung 6: Mittlere DRR-Verbesserung für RT_{60} = 160ms und RT_{60} = 610ms

Fazit

Es wurde ein Algorithmus entwickelt, der blind eine Signaltrennung und Enthallung in kombinierter Optimierung online vornimmt. Hierdurch lässt sich die Performance herkömmlicher BSS-Algorithmen verbessern, die insbesondere unter stark verhallten Signalen beeinträchtigt ist. Durch die kombinierte Optimierung lässt sich ein zusätzlicher SIR-Gewinn von 2dB im Vergleich zur sequentiellen Optimierung erzielen. Durch die Online-Optimierung des Algorithmus funktioniert dieser auch für bewegte Sprecher. Weitere Arbeit ist notwendig, um hörbare Artefakte zu vermeiden, die durch Überschätzen der Nachhallkomponente bei der Enthallung entstehen.

Literatur

[1] Kim, T.: Real-Time Independent Vector Analysis for Convolutive Blind Source Separation. In: IEEE Transactions on Circuits and Systems I: Regular Papers 57 (2010), July, Nr. 7, S. 1431–1438.

[2] Taniguchi, T.; Ono, N.; Kawamura, A.; Sagayama, S.: An auxiliary-function approach to online independent vector analysis for real-time blind source separation. In: 2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014, S. 107–111

[3] Jukić , A. ; Waterschoot , T. van ; Doclo , S.: Adaptive Speech Dereverberation Using Constrained Sparse Multichannel Linear Prediction. In: IEEE Signal Proces- sing Letters 24 (2017), Jan, Nr. 1, S. 101–105

[4] Yoshioka, T.; Nakatani, T.; Miyoshi M.; Okuno, H. G.: Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization. In: IEEE Transactions on Audio, Speech, and Language Processing 19 (2011), Jan, Nr. 1, S. 69–84

[5] Jukić , A. ; Doclo , S.: Speech dereverberation using weighted prediction error with Laplacian model of the desired signal. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014. – ISSN 1520–6149, S. 5172–5176

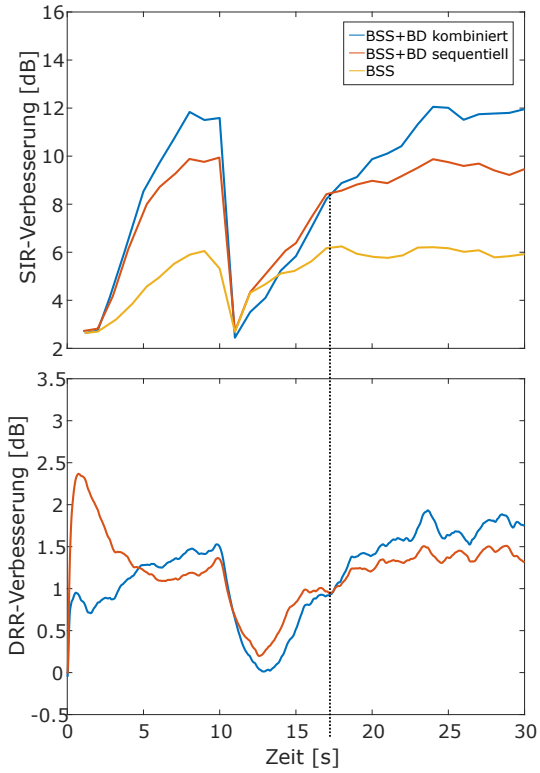

Abbildung 7: Mittlere SIR- bzw. DRR-Verbesserung für RT_{60} = 360ms bei abruptem Ändern beider Sprecherpositionen

[6] Fill, J. A.; Fishkind, D. E.: The Moore-Penrose Generalized Inverse for Sums of Matrices. In: ISAM J. Matrix Anal. Appl. 21 (1999), Oct, Nr. 2, 629-635.

[7] Marin-Hurtado , Jorge I. ; Anderson , David V.: Preservation of Localization Cues in BSS-Based Noise Reduction: Application in Binaural Hearing Aids, InTech, 2012

[8] Schiel , F. ; Draxler , Ch. ; Tillmann , H. G.: The Bavarian Archive For Speech Signals: Resources For The Speech Community. In: Proceedings of Eurospeech '97. Rhodes, 1997, S. 1687–1690

[9] Hadad , E. ; Heese , F. ; Vary , P. ; Gannot , S.: Multichannel audio database in various acoustic environments. In: 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), 2014, S. 313–317

[10] Vincent , E. ; Gribonval , R. ; Fevotte , C.: Performance measurement in blind audio source separation. In: IEEE Transactions on Audio, Speech, and Language Processing (2006), July, Nr. 7, S. 1462-1469

[11] Jeub , M. ; Nelke , C. ; Beaugeant , C. ; Vary , P.: Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals. In: 2011 19th European Signal Processing Conference, 2011. – ISSN 2076–1465, S. 1347–1351