Audiosignalverarbeitung für Videokonferenzsysteme

Thomas Schlien, Florian Heese, Magnus Schäfer, Christiane Antweiler, Peter Vary

Institut für Nachrichtengeräte und Datenverarbeitung
RWTH Aachen
Muffeter Weg 3a
52074 Aachen
{schlien,heese,schaefer,antweiler,vary}@ind.rwth-aachen.de

Abstract: Durch stetig steigende Datenraten sowohl mobiler als auch leitungsgebundener Internetzugänge haben sich die Rahmenbedingungen für Videokonferenzsysteme deutlich verbessert. Auf dieser Grundlage hat es sich das öffentlich geförderte Gemeinschaftsprojekt[1] "*Connected Visual Reality* (CoVR) – Hochqualitative visuelle Kommunikation in heterogenen Netzwerken" zur Aufgabe gemacht, die Medienqualität durch die Integration neuartiger Algorithmen der Video- und Audiosignalverarbeitung und -übertragung entscheidend zu verbessern.

Zwei Teilaspekte des Systems aus dem Bereich der Audiosignalverarbeitung werden in diesem Beitrag vorgestellt: die künstliche Bandbreitenerweiterung und die Bestimmung der akustischen Sprecheraktivität. Bei der Bandbreitenerweiterung werden Sprecheradaption sowie die Extraktion von aussagekräftigen Signalmerkmalen in gestörter Umgebung behandelt. Die Bestimmung der Sprecheraktivität erfolgt mit einem neuartigen numerisch optimierten Beamforming-Algorithmus, dessen überlegene Leistungsfähigkeit im Vergleich mit dem MVDR-Beamformer durch ein Simulationsbeispiel illustriert wird.

Mit diesen Audiosignalverarbeitungsverfahren ergeben sich neue Möglichkeiten für die Anwendung von Videokonferenzsystemen in unterschiedlichen Umgebungen sowie eine deutliche Verbesserung der wahrgenommenen Kommunikationsqualität, die durch ein entsprechendes Echtzeit-Demonstrationssystem erlebbar gemacht wird.

1 Einleitung

In den letzten Jahren hat die Bedeutung der HD Videokonferenz- und Telepräsenzsysteme deutlich zugenommen. Bislang verhinderten jedoch hohe Anschaffungskosten, eine stark eingeschränkte Qualität oder komplizierte Bedienung eine weitreichende Akzeptanz und Verbreitung dieser Systeme.

Im Gemeinschaftsprojekt *Connected Visual Reality* (CoVR) [CoV] wurde an dem Ziel gearbeitet, die Audio- und Videoqualität sowie die Interoperabilität der Systeme zu ver-

[1]Das Projekt wurde durch das NRW Ziel 2-Programm "Regionale Wettbewerbsfähigkeit und Beschäftigung" 2007-2013 und dem ERDF 'Europäischer Fonds für regionale Entwicklung' unterstützt.

EUROPÄISCHE UNION
Investition in unsere Zukunft
Europäischer Fonds
für regionale Entwicklung

Teilnehmende Projektpartner sind Ericsson GmbH, MainConcept GmbH, part of Rovi, sowie zwei Institute der RWTH Aachen, Institut für Nachrichtentechnik und Institut für Nachrichtengeräte und Datenverarbeitung.

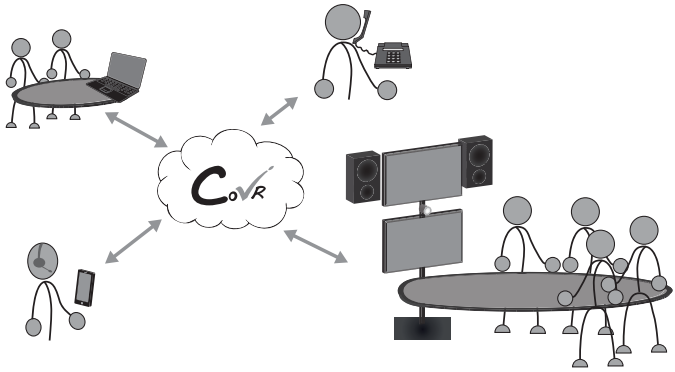


Abbildung 1: CoVR - Videokonferenz

bessern. Die Teilnehmer können sich dabei sowohl hinsichtlich ihres Aufenthaltsorts, ihrer Netzwerkanbindung als auch im Hinblick auf die eingesetzte Hardware signifikant voneinander unterscheiden. Für eine hohe Heterogenität von Systemen gemäß Abb. 1, d.h. vom speziell für die Videokonferenz eingerichteten Raum mit leistungsfähiger Hardware über Videokonferenzsysteme anderer Anbieter und Arbeitsplatzrechner, soll eine bestmögliche Qualität gewährleistet werden.

Zu diesem Zweck sind neue Methoden der Audio- und Videosignalverarbeitung, Kodierung und Übertragung entwickelt und untersucht worden. Im Bereich der Videoverarbeitung arbeiteten die Projektpartner u.a. an der Personendetektion sowie der Entwicklung und Standardisierung von HEVC (ITU-T/MPEG *High Efficiency Video Coding*). Dieser Standard erzielt gegenüber dem bisherigen H.264/AVC Standard bei vergleichbarer Videoqualität eine Reduktion der Datenrate in einer Größenordnung von ca. 50%. Im Bereich Audiosignalverbesserung sind Algorithmen u.a. zur Echokompensation, Störgeräuschreduktion und Enthallung für die mehrkanalige Kommunikation betrachtet worden.

In diesem Beitrag werden zwei weitere Schlüsselaspekte vorgestellt und am Echtzeit-System demonstriert (Kap. 2), das u.a. auf dem Gemeinschaftsstand des Landes NRW auf der CeBIT 2013 in Hannover präsentiert wurde.

Der erste Teilaspekt, die akustische Bandbreitenerweiterung, wird in Kap. 3 eingeführt. Diese erhöht die Sprachverständlichkeit für den Fall, dass sich Teilnehmer aus dem klassischen Fest- und Mobilfunknetz mit einer geringen akustischen Bandbreite in eine HD Videokonferenz einwählen. Durch die Angleichung der Sprachqualität der unterschiedlichen Konferenzteilnehmer kann der subjektiv empfundene Hörkomfort deutlich verbessert werden.

Der zweite Teilaspekt umfasst Arbeiten an einer kombinierten Video- und Audioanalyse in Kap. 4, auf die sich neue Funktionalitäten für die CoVR-Videokonferenz stützen. Personen in einem Konferenzraum können detektiert und räumlich lokalisiert werden. Darüber hinaus wird die unterschiedliche Sprachaktivität der einzelnen Teilnehmer mit Hilfe eines 8-kanaligen Mikrofonarrays und eines Beamformers dynamisch erfasst. Dies erlaubt die virtuelle Platzierung der Teilnehmer in einer gemeinsamen Gesprächssituation sowie das

optische Hervorheben des aktiven Sprechers in einer Szene. Diese intelligente Komposition und Darstellung in der Szene stellt einen innovativen Schritt in Richtung “virtueller Präsenz” dar.

2 Echtzeit-Demonstrationssystem

Um die neuen Funktionalitäten erlebbar zu machen, wurde ein Echtzeit-Videokonferenzsystem entwickelt, dessen Konstruktion die Anforderungen der zu demonstrierenden Teilaspekte berücksichtigt. Das Gesamtsystem besteht im Wesentlichen aus der Netzwerkinfrastruktur, die auf dem *IP Multimedia Subsystem* (IMS)-Standard basiert, sowie diversen, heterogenen Endgeräten. Wie in Abb. 1 dargestellt, können sich sowohl spezialisierte, leistungsfähige HD-Systeme als auch konventionelle PC- oder Telefonteilnehmer in die Videokonferenz einwählen.

Für die Demonstrationen wurde u.a. ein HD-System entworfen, das auf spezielle Hardware und Rauminstallationen verzichtet. Stattdessen wurde ein portabler Demonstrator aus vergleichsweise günstigen, handelsüblichen Hardwarekomponenten (Abb. 2-a) aufgebaut.

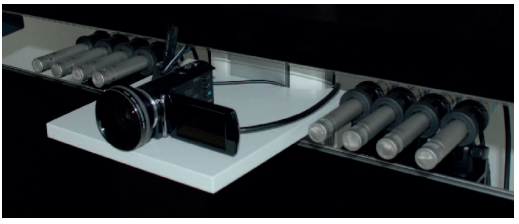
a)  b) 

Abbildung 2: Demonstratoraufbau und Mikrofonarrayinstallation

Der zweite installierte Monitor erlaubt gleichzeitig zur Videokonferenz weitere Funktionalitäten wie z.B. das gemeinsame Bearbeiten von Dokumenten. Die Kamera befindet sich zentral zwischen den beiden Monitoren. Dagegen sind die Lautsprecher zur räumlichen Wiedergabe unter den Randbedingungen eines mobilen Aufbaus möglichst weit außen, rechts und links neben dem oberen Monitor, angeordnet.

Die intelligente Szenenkomposition basiert auf einer kombinierten Video-Audioanalyse zur Quellenlokalisation mit Sprachaktivitätserkennung. Ein Kernelement dieser Analyse ist ein neuartiger Beamforming-Algorithmus. Zu diesem Zweck wurde zwischen den beiden Displays ein 8-kanaliges Mikrofonarray integriert. Die Dimensionierung des Mikrofonarrays wurde mit Blick auf die unterschiedlichen Randbedingungen von konstruktiven

Möglichkeiten und algorithmischen Wünschen optimiert. Hierbei ergab sich eine Anordnung der Mikrofone in zwei Gruppen links und rechts der Kamera gemäß Abb. 2-b als beste Lösung.

Mit diesem HD-System werden Arbeiten zur Bandbreitenerweiterung und zur Szenenkomposition – zwei neue Teilaspekte – präsentiert.

3 Akustische künstliche Bandbreitenerweiterung

Trotz der rasant fortschreitenden technischen Entwicklung in den letzten Jahrzehnten hat sich die Audioqualität im Bereich Telefonie nur wenig verändert. Dies ist durch die Begrenzung der akustischen Bandbreite von 300 Hz - 3,4 kHz begründet, die zu einem Verlust von Natürlichkeit und Verständlichkeit der Sprache führt. Zwar gibt es seit mehr als 10 Jahren standardisierte Breitband-Sprachcodecs, wie z.B. den AMR-WB [ITU03], allerdings setzt sich die breitbandige Sprachübertragung mit Frequenzen zwischen 50 Hz und 7 kHz nur sehr langsam durch. Vor allem Softwareprodukte wie Skype oder Google Hangout bieten hier einen Vorteil, da sie ihre abgeschlossenen Netzstrukturen flexibler ändern können. Um den Übergang von Schmalband- zur Breitbandtelefonie zu erleichtern und voranzutreiben, sind Verfahren zur sog. akustischen Bandbreitenerweiterung entwickelt worden (z.B. [Jax02] und [PA11]). Sie ermöglichen mit Hilfe statistischer Verfahren eine künstliche Erhöhung der akustischen Bandbreite des Signals von 4 kHz auf 8 kHz.

In Videokonferenzsystemen ist meistens eine Option zur Telefoneinwahl gewünscht, damit Teilnehmer, die unterwegs sind oder nicht über einen Videoclient verfügen, ebenfalls an der Konferenz teilnehmen können. Zu diesem Zweck wurde auch in den CoVR-Demonstrator die Einwahl von herkömmlichen Telefonteilnehmern ermöglicht. Da die Audiosignale aller anderen Konferenzteilnehmer mit voller Bandbreite übertragen werden, fällt die Audioqualität der Telefonteilnehmer deutlich ab, so dass ein sehr heterogener Höreindruck entsteht. Die Telefonteilnehmer werden regelrecht als störend wahrgenommen. Der Einsatz der künstlichen Bandbreitenerweiterung wirkt diesem Eindruck entgegen. Durch die künstliche Anhebung der Bandbreite erfolgt eine signifikante Verbesserung der Audioqualität und damit verbunden ein homogener Gesamteindruck.

Der Basisalgorithmus zur Bandbreitenerweiterung wird in Kap. 3.1 vorgestellt, während in den sich anschließenden Abschnitten auf die weiteren algorithmische Arbeiten in den Bereichen der sprecheradaptiven künstlichen Bandbreitenerweiterung (Kap. 3.2), neuer und verbesserter Merkmale für die künstliche Bandbreitenerweiterung (Kap. 3.3) und der künstlichen Bandbreitenerweiterung in gestörter Umgebung (Kap. 3.4) eingegangen wird.

3.1 Grundlagen der künstlichen Bandbreitenerweiterung

Der verwendete Algorithmus für die künstliche Bandbreitenerweiterung basiert auf [JV03] und [HGV12]. Er erweitert das im Telefonnetz übliche Schmalbandsignal (300 Hz bis 3,4 kHz) auf ein breitbandiges Audiosignal, indem Frequenzen des sog. Erweiterungsban-

des zwischen 4 kHz und 7 kHz künstlich hinzugefügt werden. Um dies zu ermöglichen, ist eine Modellannahme nötig, die die Abhängigkeiten zwischen Schmalbandsignal und Erweiterungsband ausnutzt. Verwendet wird das Modell der Spracherzeugung, welches sowohl zur Erweiterung des Anregungssignals als auch der spektralen Einhüllenden eingesetzt wird. Eine Übersicht über den Algorithmus ist in Abb. 3 wiedergegeben.
Für die Erweiterung des Anregungssignals (mittlerer Zweig in Abbildung 3) wird zunächst das Residuum des Schmalbandsignals mittels *Linear Predictive Coding* (LPC) Analyse [VM06] bestimmt, normalisiert und weißes Rauschen addiert.
Die Blöcke zur Schätzung der spektralen Einhüllenden sind im oberen Pfad dargestellt. Für die Erweiterung der spektralen Einhüllenden werden zunächst Merkmale (siehe Kapitel 3.3) aus dem Schmalbandsignal extrahiert. Mit Hilfe dieser Merkmale und einem vorher trainierten Schätzer werden *Autoregressive Modell* (AR) Koeffizienten $\underline{\tilde{a}}$ der spektralen Einhüllenden des Erweiterungsbandes auf Basis eines *Hidden Markov Models* [Rab89] bestimmt.
Aus dem künstlich erzeugtem Anregungssignal $s_{\mathrm{a}}(k)$ und den AR Koeffizienten $\underline{\tilde{a}}$ erzeugt ein Synthesefilter das Erweiterungsbandsignal $\tilde{s}_{\mathrm{hb}}(k)$. Dieses wird zuletzt mit dem Schmalbandsignal $s_{\mathrm{nb}}(k)$ mittels einer *Quadrature Mirror Filter* (QMF)-bank [EG77] zusammengemischt. Das so entstandene Signal $\tilde{s}_{\mathrm{wb}}(k)$ weist zwar eine spektrale Lücke zwischen 3,4 kHz und 4 kHz auf, diese hat aber nur wenig Einfluss auf die wahrgenommene Qualität.

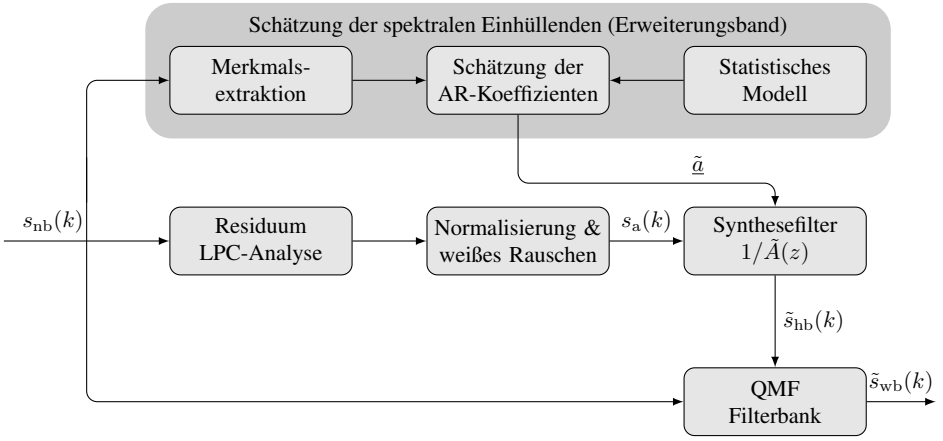


Abbildung 3: Blockschaltbild der künstlichen Bandbreitenerweiterung

3.2 Sprecheradaptive künstliche Bandbreitenerweiterung

Eins der Kernelemente der Bandbreitenerweiterung ist das statistische Modell, das die Verknüpfung von Merkmalen des schmalbandigen Signals mit Parametern des breitbandigen Signals herstellt. Untersuchungen, wie z.B. in [BF08] und [Jax02], haben gezeigt, dass die künstliche Bandbreitenerweiterung bessere Ergebnisse liefert, wenn das Modell

für die nötige Schätzung auf den Sprecher trainiert wird. In einem realistischen Anwendungsszenario ist es jedoch normalerweise nicht möglich das System auf jeden Nutzer zu trainieren. Daher wurde ein Ansatz verfolgt, der es erlaubt, einen Sprecher auf einen definierten Normsprecher abzubilden, um daraufhin eine künstliche Bandbreitenerweiterung anzuwenden. Es sind zwei Methoden möglich: Erstens eine Transformation des Sprechers im Zeit- oder Frequenzbereich und zweitens eine Transformation der Merkmale. Im ersten Fall werden Grundfrequenzen und Formanten eines Sprechers mit Techniken der sog. *Voice Conversion* möglichst genau auf einen vorher definierten Normsprecher abgebildet. Anschließend wird das Signal mittels künstlicher Bandbreitenerweiterung erweitert und wieder rücktransformiert. Im zweiten Fall werden nur die Merkmale (siehe Kapitel 3.3) des Sprechers auf einen Normsprecher abgebildet. Eine Rücktransformation ist in diesem Fall nicht mehr möglich. Dies hat jedoch keinen großen Einfluss, da die resultierende Verzerrung der spektralen Einhüllenden im Erweiterungsband kaum wahrnehmbar ist.

3.3 Merkmalsextraktion und Verbesserungen

Für die künstliche Bandbreitenerweiterung ist es essentiell, aussagekräftige Merkmale aus dem schmalbandigen Signal zu extrahieren, um eine präzise Schätzung der Parameter des breitbandigen Signals zu erlauben. In [Jax02] wird gezeigt, dass sich besonders die *Mel-Frequency Cepstral Coefficients* (MFCC) [DM80] für die Bandbreitenerweiterung eignen, da sie sowohl eine hohe Transinformation zwischen Schmalband- und Erweiterungsband bieten, als auch eine hohe Separabilität der Merkmale im Merkmalsraum besitzen. Zusätzlich zu den MFCCs werden häufig der sog. *Gradient Index* [Pau97] oder die *Zero-Crossing Rate* [AR76] als weitere Merkmale benutzt, die sich besonders gut zur Unterscheidung von stimmhaften und stimmlosen Signalen eignen.

Diese Merkmale können jedoch unter bestimmten Randbedingungen problematisch hinsichtlich ihrer Robustheit sein. Daher wurden neuere Merkmale, die z.B. im Bereich der Spracherkennung eingesetzt werden, untersucht und in das Bandbreitenerweiterungssystem integriert. Hier seien im Speziellen die *Power-Normalized Cepstral Coefficients* (PNCC) [KS12] genannt, die eine verbesserte Variante der MFCCs sind und zusätzlich eine Störgeräuschunterdrückung integriert haben, auf die im nächsten Abschnitt eingegangen wird.

3.4 Künstliche Bandbreitenerweiterung in gestörter Umgebung

Die künstliche Bandbreitenerweiterung in gestörter Umgebung stellt eine besondere Herausforderung dar und wurde bislang nur wenig untersucht. Das Problem besteht darin, dass der Schätzalgorithmus in der Regel in Störfreiheit trainiert wird, aber in einer gestörten Umgebung zum Einsatz kommt. Damit kann nicht mehr erwartet werden, dass die gleiche hohe Signalqualität erreicht wird wie unter ungestörten Bedingungen. So wird u.a. neben dem Nutzsignal auch das Störsignal spektral erweitert. Deshalb stellt sich grundsätzlich

die Frage, ob die Bandbreitenerweiterung im Falle von Hintergrundstörungen überhaupt zu Verbesserungen führen kann. Grenzexperimente zeigen, dass sich Sprachqualität und Verständlichkeit eines gestörten Schmalbandsignals durch das Hinzufügen eines perfekt geschätzten Erweiterungsbandsignals wesentlich verbessern lassen.

Prinzipiell gibt es zwei Ansätze die künstliche Bandbreitenerweiterung in gestörter Umgebung durchzuführen: Zum einen kann das Schmalbandsignal durch einen Standardstörgeräuschalgorithmus verarbeitet und anschließend der künstlichen Bandbreitenerweiterung zugeführt werden. Untersuchungen haben gezeigt, dass in diesem Fall das beste Ergebnis erreicht wird, wenn die angewendete Störgeräuschreduktion besonders aggressiv parametriert wird. Der zweite Ansatz verlagert die Störgeräuschreduktion auf die Merkmalsextraktionsebene. Hier können z.B. die in Kapitel 3.3 erwähnten PNCCs verwendet werden. Diese sind auf der einen Seite bereits auf Sprachsignale optimiert, so dass Störgeräusche ausgeblendet werden und auf der anderen Seite ist zusätzlich noch eine Störgeräuschreduktion integriert, welche die Merkmale direkt entstört.

Mit beiden Ansätzen lassen sich die Ergebnisse der künstlichen Bandbreitenerweiterung in gestörter Umgebung deutlich verbessern. Bislang kann noch keine konkrete Aussage getroffen werden, welcher der beiden Ansätze für die zu betrachtenden unterschiedlichen Randbedingungen die besten Ergebnisse liefert. Die Forschungsarbeiten in diesem Bereich dauern an.

4 Beamforming

Im Rahmen des CoVR-Videokonferenzsystems bildet die intelligente Szenenkomposition der jeweils aktivsten Sprecher einen Arbeitsschwerpunkt. Hierfür werden Analyseergebnisse aus den unterschiedlichen Signalbereichen Audio und Video miteinander verknüpft und zur räumlichen Lokalisation und Sprachaktivitätsbestimmung der einzelnen Konferenzteilnehmer eingesetzt. Mittels Videoanalyse können zuverlässig die Positionen der einzelnen Konferenzteilnehmer (innerhalb eines Raumes) bestimmt werden. Der Audiosignalverarbeitung kommt in diesem Zusammenhang die Aufgabe zu, eine quantitative Aussage über die Aktivität der beteiligten Sprecher zu liefern. Da die Sprachsignale der unterschiedlichen Sprecher aus verschiedenen Richtungen eintreffen, lässt sich hierzu diese räumliche Filterung einsetzen, die die Signale der Sprecher voneinander trennen kann. Grundlage für eine räumliche Filterung ist ein Mikrofonarray, eine entsprechende Anordnung mehrerer Mikrofone im Raum, kombiniert mit einer sich anschließenden Signalverarbeitungsstufe, dem sog. *Beamforming*.

Beamforming ist der Oberbegriff für eine Vielzahl von Mikrofonarray-Verarbeitungstechniken. Das Beamforming führt in der Regel eine Verstärkung von Signalen aus bestimmten räumlichen Regionen durch, während Signale aus anderen Richtungen gedämpft werden. Dies ermöglicht z.B. die Trennung eines Zielsignals von Störsignalen oder – wie in CoVR vorgesehen – die Trennung von Sprachsignalen konkurrierender Sprecher, die sich in einem Raum befinden.

Für diesen Anwendungsfall ergeben sich zwei interessante technische Aspekte, die es erforderlich machen, einen speziell angepassten Beamforming-Algorithmus zu entwickeln: Üblicherweise ist ein Beamformer derart ausgelegt, dass eine gewünschte Richtcharakteristik nur für schmalbandige Signale oder im Grenzfall sogar nur für eine bestimmte Frequenz vorliegt. Da die menschliche Sprache jedoch einen vergleichsweise breiten Frequenzbereich (ca. 100 Hz bis 12 kHz) umfasst, können mit diesen Techniken nur sehr eingeschränkte Ergebnisse erzielt werden. Weiterhin wird in der Literatur häufig angenommen, dass sich die Quelle im sog. Fernfeld befindet, also im Idealfall hinreichend weit vom Mikrofonarray entfernt, so dass die beim Array einfallende Welle als homogene, ebene Welle angenommen werden kann. Dies erlaubt entsprechende Modellannahmen und Vereinfachungen. In vielen praktischen Anwendungen, u.a. bei Videokonferenzsystemen, trifft diese Annahme jedoch nicht zu, da sich der Sprecher meist im sog. Nahfeld befindet, also in einer Größenordnung von einigen wenigen Wellenlängen vom Array entfernt. Sämtliche Berechnungen sind folglich auf Basis einer kugelförmigen Welle durchzuführen. Durch die Anwendung des Beamformers für breitbandige Signale wird die Problematik noch komplexer, da sich in bestimmten Szenarien die Quelle für tiefe Frequenzen im Fernfeld und zeitgleich für hohe Frequenzen im Nahfeld befinden kann.

Im Rahmen von CoVR wurden daher neue numerische Optimierungsverfahren entwickelt [SHWV12, HSV$^+$12, HSWV13], mit denen sich ein Beamformer unter Berücksichtigung eines zuvor definierten Empfangsverhaltens im Nahfeld, d.h. unter Berücksichtigung verschiedener Entfernungen und Winkel, optimieren lässt. Weiterhin wird berücksichtigt, dass es sich bei dem Eingangssignal um ein breitbandiges Signal (z.B. ein Sprachsignal) handelt.

4.1 Beamformer Struktur

Der entwickelte Beamformer gehört zur Klasse der *Filter-and-Sum* Beamformer. Ein vereinfachtes Blockschaltbild ist in Abb. 4 dargestellt. Um eine möglichst gleichmäßige Richtcharakteristik unabhängig von der Betriebsfrequenz zu erhalten, werden die M Mikrofonsignale $u_m(k)$ mit Hilfe einer Filterbank in N ungleichförmige Teilbänder aufgeteilt. Dadurch erhöht sich die Anzahl der Freiheitsgrade bei der Bestimmung der Filterkoeffizienten. Im Anschluss werden die einzelnen Teilbänder mittels individueller *Filter-and-Sum* Einheiten (*Finite Impulse Response* (FIR) Filter), repräsentiert durch die Impulsantworten $\mathbf{h}_n^m$ mit $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$, prozessiert und anschließend zu $v(k)$ überlagert. Dabei entspricht n dem Teilbandindex und m dem Mikrofonindex. Die Abtastwerte der Mikrofonsignale werden durch eine analog-digital Wandlung mit einer Abtastrate von $f_s = 48\,\text{kHz}$ gewonnen, wobei k dem diskreten Zeitindex entspricht.

Eine Punktschallquelle $s(k)$ befinde sich am Raumpunkt $\mathbf{p}$ auf einem geeignet gewählten räumlichen Abtastgitter (z.B. ein zweidimensionales kartesisches Koordinatensystem $\mathbf{p} = (x, y)^T$). Zusammen mit der Impulsantwort $h_{\mathbf{p}m}(k)$ zwischen dem Punkt $\mathbf{p}$ und dem m-ten Mikrofon lässt sich das Mikrofonsignal $u_m(k)$ zu

$$u_m(k) = h_{\mathbf{p}m}(k) * s(k) \tag{1}$$

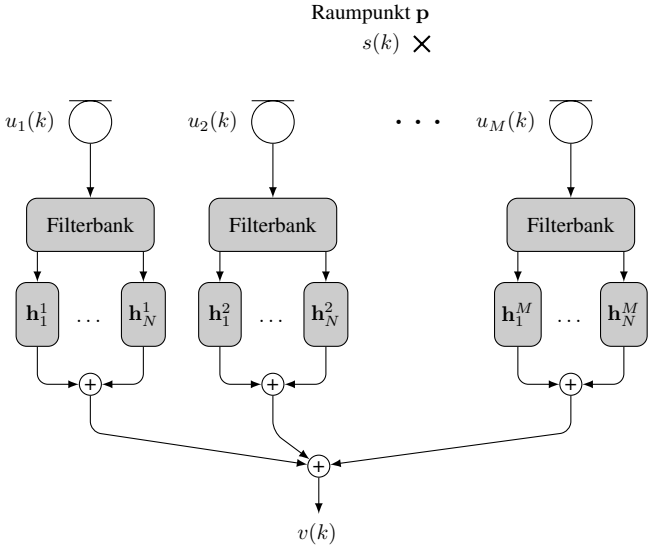


Abbildung 4: Modifizierter *Filter-and-Sum* Beamformer mit M Mikrofonen und N ungleichförmigen Teilbändern

bestimmen. Das Ausgangssignal des Beamformers $v(k)$ ist von der Quellenposition $\mathbf{p}$ abhängig und kann wie folgt berechnet werden

$$v(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * \left(h_n^{\mathrm{FB}}(k) * u_m(k) \right), \tag{2}$$

dabei entspricht $h_n^{\mathrm{FB}}(k)$ der Filterbank und $h_n^m(k)$ den FIR Teilband-Filtern mit einer Länge von L.

4.2 Numerische Optimierung

Die Bestimmung von geeigneten Impulsantworten $\mathbf{h}_n^m$ zum Design des *Filter-and-Sum* Beamformers stellt in diesem Zusammenhang eine zentrale Aufgabe dar. Für die Berechnung der Filterkoeffizienten wurde ein iterativer numerischer Optimierungsalgorithmus entwickelt, dessen Struktur in Abb. 5 dargestellt ist. Grundlage des Optimierungsverfahrens ist eine Simulation der akustischen Umgebung des Mikrofonarrays. Durch die Kombination der akustischen Umgebung mit den Filterkoeffizienten wird die Richtcharakteristik, d.h. die räumliche Verteilung von Bereichen der Verstärkung bzw. Dämpfung vor dem Mikrofonarray, bestimmt.

Dazu wird zunächst für jeden Raumpunkt $\mathbf{p}$ auf dem räumlichen Abtastgitter das dazuge-

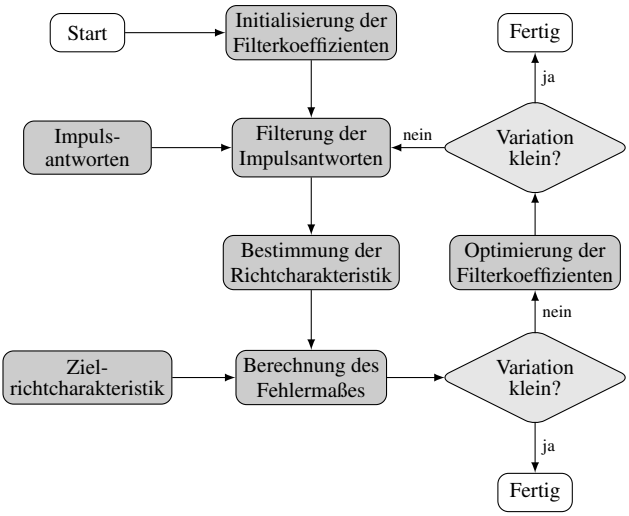


Abbildung 5: Ablauf des numerischen Optimierungsverfahrens

hörige gefilterte Ausgangssignal $v(k)$ des Beamformers berechnet gemäß

$$v(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k) * s(k). \tag{3}$$

Das resultierende Gesamtfilter $g_{\mathbf{p}}(k)$ ergibt sich somit zu

$$g_{\mathbf{p}}(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k) \tag{4}$$

und die Fouriertransformierte des Gesamtfilters $g_{\mathbf{p}}(k)$ gemäß

$$G_{\mathbf{p}}(f) = \mathcal{F}\left\{g_{\mathbf{p}}(k)\right\} . \tag{5}$$

Damit lässt sich die Richtcharakteristik $S_{\mathbf{p}}(f)$ in dB für jede Frequenz f an jedem Raumpunkt $\mathbf{p}$ in der Umgebung des Mikrofonarrays mit

$$S_{\mathbf{p}}(f) = 20 \cdot \log_{10} |G_{\mathbf{p}}(f)| \tag{6}$$

angeben. Diese berechnete Richtcharakteristik wird mit einer vorgegebenen Zielrichtcharakteristik $\hat{S}_{\mathbf{p}}(f)$ verglichen. Die Zielrichtcharakteristik besteht aus einer (örtlichen) Verteilung von Gebieten mit definierter Verstärkung $\mathbb{P}_{\mathrm{high}}$ (Zielpegel S_{high}) bzw. Dämpfung $\mathbb{P}_{\mathrm{low}}$ (Zielpegel S_{low}). Prinzipiell lässt sich diese Verteilung individuell für unterschiedliche Frequenzen erstellen, allerdings erscheint eine frequenzunabhängige Verteilung in vielen Szenarien vorteilhaft

$$\hat{S}_{\mathbf{p}}(f) = \hat{S}_{\mathbf{p}} = \begin{cases} S_{\mathrm{high}} & \text{für } \mathbf{p} \in \mathbb{P}_{\mathrm{high}} \\ S_{\mathrm{low}} & \text{für } \mathbf{p} \in \mathbb{P}_{\mathrm{low}} . \end{cases} \tag{7}$$

Die endgültige Wahl der Zielrichtcharakteristik hängt von der Anwendung des Beamformers ab. Für die Sprecheraktivitätserkennung in CoVR beispielsweise wird die Zielrichtcharakteristik durch die Anzahl und Position möglicher Sprecher im Raum vorgegeben – eine Information, die von der Videoanalyse zur Verfügung gestellt wird.

Zwischen den beiden Richtcharakteristiken wird ein quadratisches Fehlermaß Δ_S bestimmt, das für alle Punkte, an denen $\hat{S}_{\mathbf{p}}(f)$ definiert ist (siehe Gl. 7), und über alle Frequenzen f_i ($i \in \{i_{\min}, \ldots, i_{\max}\}$) die Leveldifferenzen aufsummiert

$$\Delta_S(n) = \sum_{i=i_{\min}}^{i_{\max}} \sum_{\mathbf{p}\in(\mathbb{P}_{\text{high}}\cup\mathbb{P}_{\text{low}})} \hat{S}_{\mathbf{p}}(f_i) - S_{\mathbf{p}}(f_i), \tag{8}$$

dabei bezeichnen $f_{i_{min}}$ und $f_{i_{max}}$ die untere bzw. obere Grenzfrequenz des entsprechenden Teilbands n.

In Abhängigkeit dieses Fehlermaßes werden die optimalen Filterkoeffizienten für jedes Teilband n im Sinne des *minimalen mittleren quadratischen Fehlers* (MMSE) bestimmt:

$$\left[\mathbf{h}_n^1, \ldots, \mathbf{h}_n^M\right]_{\text{opt}} = \arg\min_{\mathbf{h}} \Delta_S(n)^2 \,. \tag{9}$$

Die Optimierung erfolgt mittels des *iterative interior-point* Algorithmus [BGN00] unter der Nebenbedingung, dass die Filterkoeffizienten in einem Wertebereich zwischen -1 und 1 liegen. Zur Stabilisierung des Algorithmus wird geprüft, inwieweit sich die Filterkoeffizienten überhaupt noch verändern. Ist die Änderung hinreichend klein, wird der Algorithmus abgebrochen. Die Optimierung der Filterkoeffizienten kann sowohl auf der Basis generierter als auch auf gemessenen Impulsantworten stattfinden.

4.3 Leistungsvergleich

Die Leistungsfähigkeit des neuen Verfahrens soll anhand eines Vergleichs mit einem etablierten Verfahren zur Bestimmung der Filterkoeffizienten, dem *Minimum Variance Distortionless Response* (MVDR)-Beamformer [VM06] verdeutlicht werden. Dazu sind in den folgenden Abbildungen die Richtcharakteristiken der beiden Verfahren für ein identisches Setup bei zwei unterschiedlichen Frequenzen (500 Hz und 2000 Hz) dargestellt. Die Simulationen wurden zur besseren Vergleichbarkeit beider Algorithmen unter Freifeldbedingungen durchgeführt.

Für die numerische Optimierung der Filterkoeffizienten kann eine gewünschte Richtcharakteristik im Nahfeld des Mikrofonarrays vorgegeben werden. Beide Mikrofonsysteme wurden so entworfen, dass sie Quellen auf der linken Seite ($-0.5\,\text{m} \leq x < 0\,\text{m} \wedge 0.2\,\text{m} < y \leq 0.8\,\text{m}$) verstärken, wohingegen Quellen auf der rechten Seite ($0\,\text{m} < x \leq 0.5\,\text{m} \wedge 0.2\,\text{m} < y \leq 0.8\,\text{m}$) gedämpft werden. In Abb. 6 und 7 ist dies durch die weißumrandeten Boxen markiert. Die Auflösung des räumlichen Abtastgitters beträgt in beiden Raumdimensionen (x, y) 0.01 m, was in jeweils 3000 Raumpunkten für die Gebiete $\mathbb{P}_{\text{high}}$ und $\mathbb{P}_{\text{low}}$ resultiert. Für beide Systeme wurde eine Pegeldifferenz von 40 dB zwischen den verstärkten $\mathbb{P}_{\text{high}}$ und bedämpften $\mathbb{P}_{\text{low}}$ Gebieten gewählt.

Das Mikrofonarray besteht aus $M = 8$ Sensoren, die gemäß Abb. 2-b in Abständen von [3, 3, 3, 30, 3, 3, 3] cm angeordnet sind. Die sog. räumliche Aliasingfrequenz, oberhalb der es zu Mehrdeutigkeiten hinsichtlich der Richtung kommt, liegt bei diesem Aufbau bei ca. 5600 Hz. Somit lässt sich das Verhalten oberhalb dieser Frequenz nicht mehr eindeutig durch die Ansteuerung kontrollieren. Für das neue System wird eine ungleichmäßige Filterbank [Löl11] verwendet, die aus $N = 6$ Teilbändern besteht. Der Frequenzbereich der einzelnen Teilbänder ist in Tab. 1 zu sehen. Der Einfachheit halber sind die Teilband-Filter als FIR Filter realisiert worden. Die Impulsantwortlänge der *Filter-and-Sum* Einheiten $\mathbf{h}_n^m$ wurde zu $L = 8$ bestimmt. Die Filterlänge des MVDR Beamformers wurde pro Mikrofon

Band	Frequenzbereich [Hz]		Band	Frequenzbereich [Hz]	
1	1	268	4	1549	2614
2	268	839	5	2614	4731
3	839	1549	6	4731	12049

Tabelle 1: Filterbank Teilbänder

und für den gesamten Frequenzbereich zu 96 bestimmt. Damit ist sie doppelt so lang wie die effektive Filterlänge des neuen Systems ($N \cdot L = 48$). Untersuchungen in [Dör98] haben ergeben, dass für die maximale Suszeptibilität K_0 ein Wert zwischen 2 und 5 optimal ist, so dass für die folgenden Simulationen $K_0 = 3$ gewählt wurde.

Der Vergleich beider Systeme erfolgt für zwei unterschiedliche Frequenzen:

- $f_i = 500\,\mathrm{Hz}$ als ein Beispiel für eine tiefe Frequenz, bei der das Mikrofonarray betrieben werden kann und
- $f_i = 2000\,\mathrm{Hz}$ als eine Frequenz, die sich in der Mitte des Operationsbereichs des Mikrofonarrays befindet.

In Abb. 6 sind die Richtcharakteristiken für $500\,\mathrm{Hz}$ und $2000\,\mathrm{Hz}$ des MVDR-Systems dargestellt. Bei $500\,\mathrm{Hz}$ erzielt der MVDR-Beamformer lediglich eine äußerst geringe Richtwirkung. Für die Konfiguration bei $2000\,\mathrm{Hz}$ ist eine gewisse Pegeldifferenz zwischen $\mathbb{P}_{\mathrm{high}}$ und $\mathbb{P}_{\mathrm{low}}$ zu erkennen. Der MVDR-Beamformer weist in diesem Fall jedoch ein sehr inhomogenes Verhalten im Sperrbereich auf.

Die Richtcharakteristiken des neuen Systems (siehe Abb. 7) zeigen ein deutlich verbessertes Empfangsverhalten. Für beide Betriebsfrequenzen stellen sich signifikante und homogene Pegeldifferenzen zwischen den Gebieten $\mathbb{P}_{\mathrm{high}}$ und $\mathbb{P}_{\mathrm{low}}$ ein. Insbesondere wird die vorgegebene Richtcharakteristik auch in den Randbereichen gut angenähert.

Im direkten Vergleich mit dem in der Praxis weit verbreiteten MVDR-Beamformer zeigt sich, dass das neue numerische Optimierungsverfahrens in einem breiten Frequenzbereich eine zuverlässige Trennung von (in diesem Fall zwei) aktiven Sprechern erlaubt. Das komplette Beamforming-System wurde in den Demonstrator integriert und kann unter Echtzeitbedingungen eingesetzt werden, um den Grad der Sprachaktivität der einzelnen Sprecher zu quantifizieren. Diese Information dient als Grundlage für eine intelligente Szenenkomposition des CoVR Videokonferenzsystems.

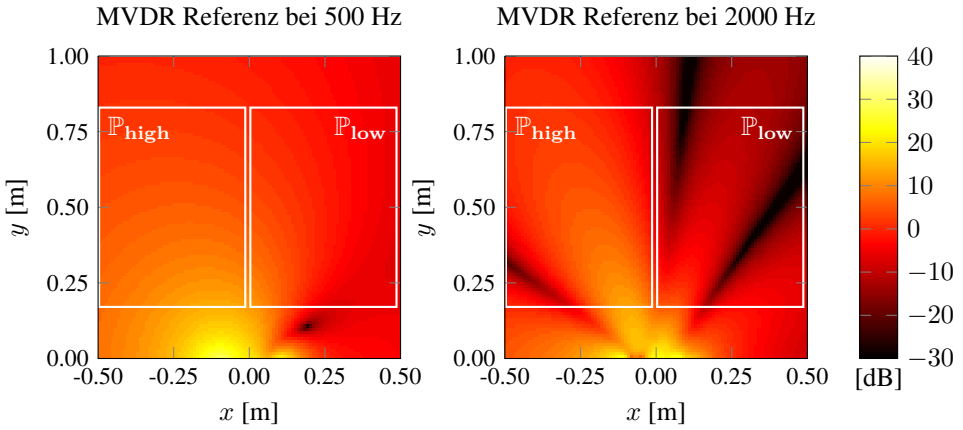


Abbildung 6: Richtcharakteristik des klassischen MVDR-Beamformers

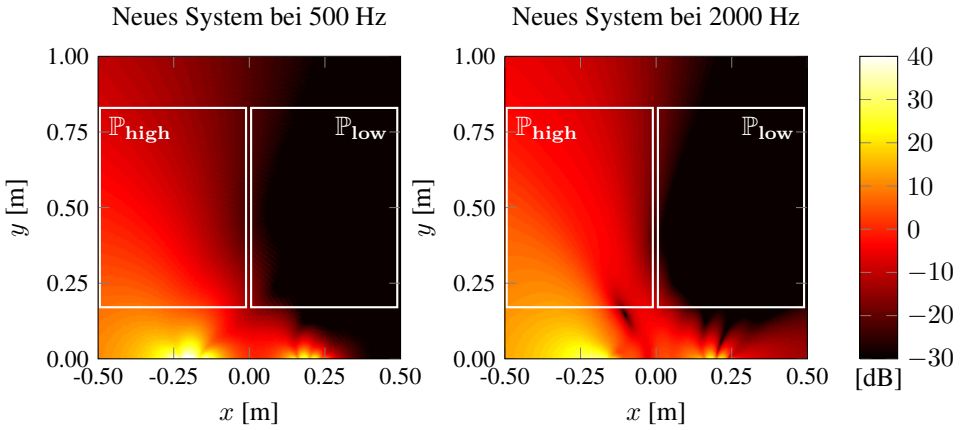


Abbildung 7: Richtcharakteristik mit dem neuartigen Optimierungsverfahren

5 Zusammenfassung

Im Rahmen des Forschungsprojekts CoVR sind Algorithmen zur Audiosignalverarbeitung in Videokonferenzsystemen untersucht und in einen Demonstrator integriert worden. Kernthemen in diesem Kontext sind Verfahren zur Echokompensation, Störgeräuschreduktion, Enthallung und die mehrkanalige Signalverarbeitung. In diesem Beitrag wurden mit der akustischen Bandbreitenerweiterung und der Sprecheraktivitätserkennung zwei neue Teilaspekte vorgestellt, die die Implementierung neuer Funktionalitäten in den HD-Demonstrator erlauben.

Zunächst wurde die grundlegende Struktur der Bandbreitenerweiterung eingeführt. Unterschiedliche Sprechercharakteristiken, Hintergrundstörungen und eine heterogene Kommunikationsinfrastruktur stellen die besonderen Herausforderungen beim Einsatz der Bandbreitenerweiterung dar. Zur Lösung dieser Probleme wurden unterschiedliche Ansätze vorgestellt. Hierbei fanden Konzepte zur Sprecheradaptivität und Varianten der Merkmalsextraktion – inbesondere in gestörter Umgebung – Berücksichtigung.

Die Integration der Bandbreitenerweiterung in dem CoVR-Demonstrator hat das enorme Potential des Verfahrens aufgezeigt. Durch den Einsatz der Bandbreitenerweiterung lässt sich der subjektiv empfundene Gesamt-Höreindruck signifikant verbessern. Die Arbeiten an diesem Themenkomplex sind derzeit noch nicht abgeschlossen. Um die Bandbreitenerweiterung in der Praxis für die unterschiedlichen Randbedingungen zu optimieren, sind weiterführende Untersuchungen geplant.

Im zweiten Teil des Beitrags wurde ein neuartiger, numerisch optimierter Beamforming-Algorithmus vorgestellt. Dieser bestimmt in einer gemeinsamen Video- und Audiosignalanalyse die Sprachaktivität der einzelnen Konferenzteilnehmer. Die vier aktivsten Sprecher werden anschließend empfangsseitig, künstlich in einer Szene kombiniert. Durch eine Gegenüberstellung mit einem klassischen MVDR-Beamformer konnten die Vorteile des neuen Algorithmus deutlich gemacht werden.

Neben der Audiosignalverarbeitung sind von den Partnern in CoVR ganz unterschiedliche Themen bearbeitet worden. Die Heterogenität der Netzwerkstruktur, die Interoperabilität diverser Systeme, die Qualität und der Einsatz neuer Funktionalitäten standen hierbei im Mittelpunkt der Untersuchungen. Teile dieser Arbeiten sind gemeinsam in einem Echtzeit-Prototypen umgesetzt. Auf diese Weise ist ein Videokonferenzsystem entstanden, mit dem die Funktionsfähigkeit und die Realisierbarkeit der bisher entwickelten Algorithmen demonstriert werden kann. Darüber hinaus bietet dieses System eine Entwicklungsplattform für zukünftige Arbeiten.

Literatur

[AR76] B.S. Atal und L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):201–212, 1976.

[BF08] P. Bauer und T. Fingscheidt. An HMM-based artificial bandwidth extension evaluated by cross-language training and test. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 4589–4592, April 2008.

[BGN00] R. H. Byrd, J. C. Gilbert und J. Nocedal. A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming. *Mathematical Programming*, 89(1):149–185, 2000.

[CoV] Connected Visual Reality (CoVR) - Hochqualitative visuelle Kommunikation in heterogenen Netzwerken, 2013, Gemeinschaftprojekt der Unternehmen Ericsson GmbH, MainConcept GmbH, part of Rovi, und zweier Institute der RWTH Aachen: dem Institut für Nachrichtentechnik und dem Institut für Nachrichtengeräte und Datenverarbeitung, http://www.covr.rwth-aachen.de.

[DM80] S. Davis und P. Mermelstein. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[Dör98] Matthias Dörbecker. *Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gestörter Sprachsignale am Beispiel elektronischer Hörhilfen*. Doktorarbeit, Rheinisch-Westfälische Technische Hochschule Aachen, Institut für Nachrichtengeräte und Datenverarbeitung, Muffeter Weg 3, 52072 Aachen, Juli 1998.

[EG77] D. Esteban und C. Galand. Application of quadrature mirror filters to split band voice coding schemes. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Jgg. 2, Seiten 191–195, 1977.

[HGV12] Florian Heese, Bernd Geiser und Peter Vary. Intelligibility Assessment of a System for Artifical Bandwidth Extension of Telephone Speech. In *Proceedings of German Annual Conference on Acoustics (DAGA)*, Seiten 905–906. DEGA, Marz 2012.

[HSV+12] Florian Heese, Magnus Schäfer, Peter Vary, Elior Hadad, Shmulik Markovich Golan und Sharon Gannot. Comparison of Supervised and Semi-supervised Beamformers Using Real Audio Recordings. In *Proceedings of IEEE 27-th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*. IEEE, November 2012.

[HSWV13] Florian Heese, Magnus Schäfer, Jona Wernerus und Peter Vary. Numerical Near Field Optimization of a Non-Uniform Sub-band Filter-and-Sum Beamformer. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Mai 2013.

[ITU03] ITU-T Recommendation G.722.2. Wideband Coding of Speech at around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB), Juli 2003.

[Jax02] Peter Jax. *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*. Dissertation, IND, RWTH Aachen, 2002.

[JV03] Peter Jax und Peter Vary. On Artificial Bandwidth Extension of Telephone Speech. *Signal Processing*, 83(8):1707–1719, August 2003.

[KS12] Chanwoo Kim und R.M. Stern. Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 4101–4104, 2012.

[Löl11] Heinrich W. Löllmann. *Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application*. Dissertation, IND, RWTH Aachen, November 2011.

[PA11] H. Pulakka und P. Alku. Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2170–2183, 2011.

[Pau97] Jürgen Paulus. *Codierung breitbandiger Sprachsignale bei niedriger Datenrate*. Dissertation, IND, RWTH Aachen, 1997.

[Rab89] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[SHWV12] Magnus Schäfer, Florian Heese, Jona Wernerus und Peter Vary. Numerical Near Field Optimization of Weighted Delay-and-Sum Microphone Arrays. In *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*. IWAENC, September 2012.

[VM06] Peter Vary und Rainer Martin. *Digital Speech Transmission - Enhancement, Coding and Error Concealment*. John Wiley & Sons, 2006.