HD-Voice-3D: Herausforderungen und Lösungen bei der Audiosignalverarbeitung

Matthias Rüngeler*, Hauke Krüger*, Gottfried Behler+, Peter Vary*

*Institut für Nachrichtengeräte und Datenverarbeitung
RWTH Aachen, Muffeter Weg 3a, 52074 Aachen
{ruengeler,krueger,vary}@ind.rwth-aachen.de

+Institut für Technische Akustik
RWTH Aachen, Neustraße 50, 52066 Aachen
gkb@akustik.rwth-aachen.de

Abstract: Der Übergang von Schmalband-Sprache zu Breitband-Sprache mit höherer Qualität - auch bezeichnet als *HD-Voice* - in öffentlichen Telefonnetzen war und ist (immer noch) ein steiniger Weg: Telekommunikationsanbieter fürchten zusätzliche Investitions- und Betriebskosten durch Einführung neuer Technologien mehr, als dass sie einen Vorteil in höherer Kommunikationsqualität im Kundensinne und damit Kundenzufriedenheit suchen.
Erst die Einführung von neuartigen Voice-over-Internet-Protocol (VoIP) Anwendungen im Zuge der weiten Verbreitung hochratiger Internetanschlüsse, in denen fast ausschliesslich *HD-Voice*-Technologie zum Einsatz kommt, scheint hier gerade einen Durchbruch zu schaffen, der zu einem Umdenken führt.
Aber was kommt nach *HD-Voice*? Wir denken, dass es *HD-Voice-3D* ist - auch bezeichnet als *Binaurale Telefonie* -, die den nächsten evolutionären Schritt darstellen wird. Dabei werden durch die Übertragung von Binauralsignalen anstatt monauraler Signale nicht nur die Inhalte von Sprache an sich, sondern auch die Atmosphäre und die akustische Umgebung realitätsnah vom einen zum anderen Ende transportiert. Das Resultat ist das Gefühl, mit den Ohren des Kommunikationspartners zu hören - so als wäre man wirklich vor Ort. Der Vorteil bei *HD-Voice-3D* gegenüber einer Stereoübertragung liegt darin, dass nicht nur die Unterscheidung zwischen Signalen von links und rechts, sondern auch von oben, unten, hinten und vorne ermöglicht wird. Der Grund liegt in der binauralen Aufnahmetechnik, die durch Abschattungs-, Beugungs- und andere Filtereffekte an Korpus, Kopf und Ohren eine natürliche räumliche Klangwahrnehmung erreicht.

In Bezug auf die Sprach/Audiosignalverarbeitung muss bei der Binauralen Telefonie jedoch im Vergleich zur Signalverarbeitung bei der monauralen Telefonie einiges beachtet werden, um die sogenannten „Binauralen Cues“, also bestimmte Eigenschaften des binauralen Audiosignals die zu einem realistischen räumlichen Eindruck bei der Perzeption der übertragenen Signale führen, nicht zu zerstören.
In diesem Paper werden typische Funktionalitäten eines binauralen VoIP-Terminals vorgestellt, ihre Auswirkung auf Binauralsignale diskutiert und erste Lösungsansätze präsentiert. Weiter wird die Thematik mittels eines Echzeit-Demonstrators vertieft, der auf dem WASP-Event „HD-Voice-3D zum Anfassen“ vorgestellt werden soll.

1 Einleitung

Fast jeder telefoniert - sei es über einen drahtgebundenen Fernsprecher oder unter Verwendung eines hochmodernen Smartphones mittels Funktechnologie. Kaum ein Endkunde ist sich jedoch bewusst, dass trotz all der Innovationen im Bereich der mobilen Endgeräte und der Signalverarbeitung im Allgemeinen, die Sprachqualität bei der Telefonie immer noch den gleichen Stand wie vor 20 Jahren hat: Der Kunde sendet und empfängt ein monaurales Audiosignal mit einer Audio-Bandbegrenzung, die dazu führt, dass Frequenzen unterhalb von 300 Hz sowie oberhalb von 3400 Hz nicht enthalten sind. Diese Technologie wird im Folgenden auch als *konventionelle Telefonie* bezeichnet.
Als „Fortschritt“ in diesem Bereich konnte über viele Jahre lediglich die Reduktion der Kosten für Betreiber und Nutzer angeführt werden [FKM12], nicht jedoch eine verbesserte Kommunikationsqualität im Sinne des Kunden. Im Gegenteil, mancher Telekommunikationsanbieter reduziert die Betriebskosten sogar noch zu Lasten der Gesprächsqualität, z.B., indem Sprachsignale vermehrt in stark komprimierter Form im internen Netz des Anbieters parallel mit vielen anderen Gesprächen gemeinsam von einem Ort zum anderen übertragen werden.
In den letzten Jahren ist es jedoch durch die Einführung der Internet basierten Telefonie (engl.: *Voice-over-Internet-Protocol*, [TW07]) teilweise zu einem Umdenken gekommen: Anbieter wie *Skype* oder *Google* bieten Kommunikationslösungen, die neuartige Sprachcodecs verwenden, die sogenannte *HD-Voice*-Qualität [3GP01, ITU02] ermöglichen: Die so über das *Internet-Protocol* (IP) übertragenen Sprachsignale decken einen Frequenzbereich von 50 Hz bis 7 kHz oder sogar mehr ab (sogenannte *Wideband-Telefonie* oder *Superwideband-Telefonie* [3GP05]). In der Folge hat sich die Sprachqualität signifikant erhöht, und die Akzeptanz bei den Kunden ist hervorragend.
Aber was kommt nach *HD-Voice*? Ein mögliches Zukunftsszenario ist *HD-Voice-3D*, im Folgenden auch als *Binaurale Telefonie* bezeichnet. Dabei liegt der Unterschied zur *konventionellen Telefonie* und zu *HD-Voice* in der Verwendung eines binauralen anstatt eines monauralen Audiosignals, welches über neuartige binaurale Endgeräte aufgenommen und abgespielt wird. Mit entsprechenden Endgeräten kann erreicht werden, dass genau die Signale übertragen werden, die der Kommunikationspartner auf der anderen Seite hört, so dass der nahe Sprecher quasi mit den „Ohren des fernen Sprechers“ hört. Es werden neben den reinen Inhalten der aufgenommenen Sprache auch der Raumeindruck sowie die akustische Umgebung mit übertragen. Der Raumeindruck entsteht im Unterschied zu einer Stereoübertragung nicht nur durch Laufzeit- und leichte Dämpfungsunterschiede zwischen beiden Kanälen, sondern auch durch Abschattungs-, Beugungs- und andere Filtereffekte an Korpus, Kopf und Ohren, die es dem Höhrer ermöglichen, nicht nur zwischen Signalen von links und rechts, sondern auch von oben, unten, hinten und vorne zu unterscheiden.
Tests mit ersten prototypischen Ausführungen der Binauralen Telefonie [RKSV12] haben gezeigt, dass neben einer erhöhten Verständlichkeit bei der Kommunikation mit gleichzeitig sprechenden Kommunikationspartnern eine ganz neue Dimension der Kommunikation erreicht werden kann, die einem Sprecher das Gefühl vermittelt als sei er - zumindest akustisch - wirklich vor Ort.
So wie *HD-Voice-3D* in eine neue Dimension der Kommunikation vorstößt, erfordet diese Technologie eine ganze Reihe von neuen Algorithmen zur digitalen Audiosignalverarbeitung, die auf die besonderen Bedürfnisse bei der binauralen Wahrnehmung von Schal-

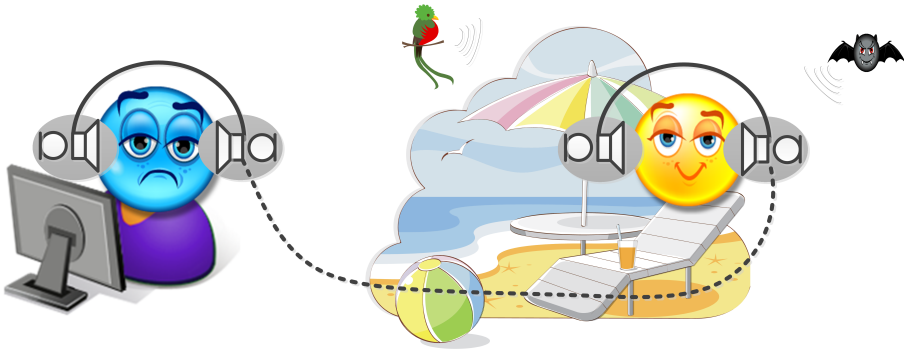

Abbildung 1: **Telefonieren mit Übertragung der natürlichen akustischen Atmosphäre mittels *HD-Voice-3D***: Eine Person mit mobilem Anschluss ist verbunden mit einer Person an einem festen Ort und lässt diese teilhaben an der akustischen Umgebung durch Verwendung von *HD-Voice-3D*-Endgeräten.

lereignissen Rücksicht nehmen müssen. Insbesondere müssen die in binauralen Signalen enthaltenen sogenannten „Binauralen Cues“, also die inherenten Merkmale der Relationen zwischen linkem und rechtem Kanal, die die Wahrnehmung von 3D-Audio erst ermöglichen, in der komplexen Signalverarbeitungskette einer Ende-zu-Ende VoIP Kommunikation geschützt und unverändert übertragen werden. Dies erfordert ein Umdenken im Design von Teilfunktionen wie z.B. der Echokompensation, der Störreduktion, der Signalkomprimierung oder der Parametrierung eines Adaptiven Jitterbuffers. Ansätze für neuartige Algorithmen von Teilfunktionalitäten bzgl. binauraler Signale wurden in der Vergangenheit z.B. im Zusammenhang mit Hörgeräten entwickelt und untersucht z.B. [Lot04, Jeu12], in anderen Bereichen wie z.B. dem Adaptiven Jitterbuffer wird technologisches Neuland betreten.

Um die neuartigen Herausforderungen exemplarisch zu demonstrieren und eine Diskussion anzustoßen, soll auf dem WASP-Workshop 2013 ein Echtzeit-Prototyp vorgeführt werden, mit dem mittels *HD-Voice-3D* Technologie telefoniert werden kann.

Grundlage für diesen Demonstrator ist das sogenannte *Software-Defined-Terminal* (SDT), eine Entwicklungsplattform, die erstmals in [KSRV12] vorgestellt wurde und ein generisches Softphone realisiert. Besonderheit dieses Terminals ist, dass beliebige Komponenten der Signalverarbeitungskette zur Laufzeit geladen und anschliessend in Echtzeit verwendet werden können. Basierend auf dieser Plattform können eine Vielzahl von Anwendungen, von der *konventionellen (bandbegrenzten) Telefonie* bis hin zur *HD-Voice-3D* Telefonie, durch einfache Neukonfiguration umgesetzt werden. In der Echtzeitdemonstration werden insbesondere die Vorteile der binauralen Übertragung und die Auswirkungen des Verlusts der binauralen Cues bei der *konventionellen Telefonie* und der Verwendung von *HD-Voice* sowie erste neuartige Ansätze, z.B. eines binauralen Adaptiven Jitterbuffers, vorgeführt.

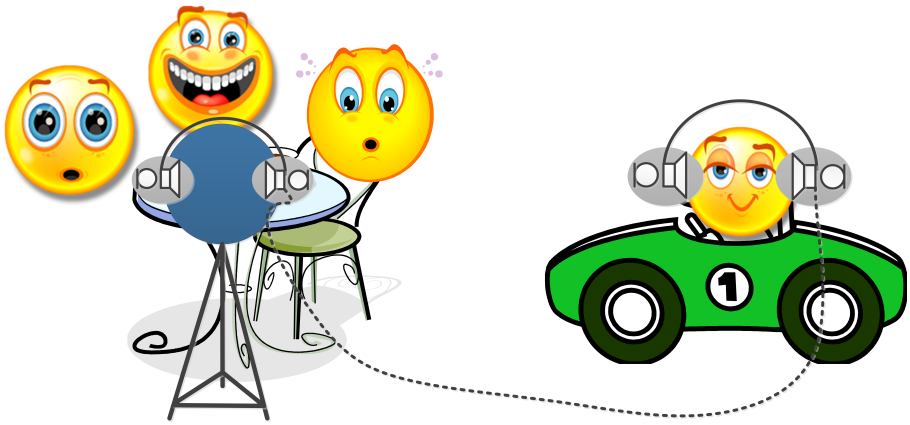

Abbildung 2: **Teilnahme an einer Besprechung mit mehreren Teilnehmern**: Eine Person mit mobilem Anschluss befindet sich z.B. in einem Auto und ist verbunden mit mehreren Personen in einer Konferenzsituation. Durch die Verwendung eines *HD-Voice-3D*-Endgeräts kann der mobile Teilnehmer der Konferenzsituation mühelos folgen.

2 *HD-Voice-3D* Kommunikation

2.1 Nutzungsszenarien für *HD-Voice-3D*

Ziel bei der *HD-Voice-3D* Kommunikation ist die Übertragung sowohl von sprachlichen Inhalten als auch der akustischen Atmosphäre, in der sich der Kommunikationspartner befindet. Typische Benutzer-Szenarien, die demonstrieren, wo die entscheidenden Unterschiede zwischen *HD-Voice-3D* und *konventioneller Telefonie* liegen, werden im Folgenden kurz skizziert:

- **Telefonieren mit Übertragung der natürlichen akustischen Atmosphäre mittels *HD-Voice-3D***
  Dieses Anwendungsszenario ist in Abbildung 1 dargestellt. Dabei ist ein Partner mittels mobilem Endgerät unterwegs und verbindet sich mit einem Kommunikationspartner mit ortsfesten Anschluss. Er trägt ein binaurales Endgerät, so dass der Teilnehmer mit dem ortsfesten Anschluss quasi mit den Ohren des mobilen Teilnehmers hört. Durch die Verwendung von *HD-Voice-3D* bekommt der angerufene Kommunikationspartner somit den Eindruck als sei er selbst vor Ort und kann auch die räumlichen Feinheiten der akustischen Umgebung wahrnehmen (Vögel, Fledermaus, Strandgeräusche).

- **Teilnahme an einer Besprechung mit mehreren Teilnehmern mittels *HD-Voice-3D***
  Dieses Szenario ist in Abbildung 2 dargestellt. Dabei wird hier *HD-Voice-3D*-Technologie verwendet, um eine Einzelperson in eine Besprechung mit mehreren Teilnehmern einzubinden: Das binaurale Endgerät trägt entweder einer der Teilnehmer der Besprechung, oder es wird ein Aufnahmemedium verwendet, das die Charak-

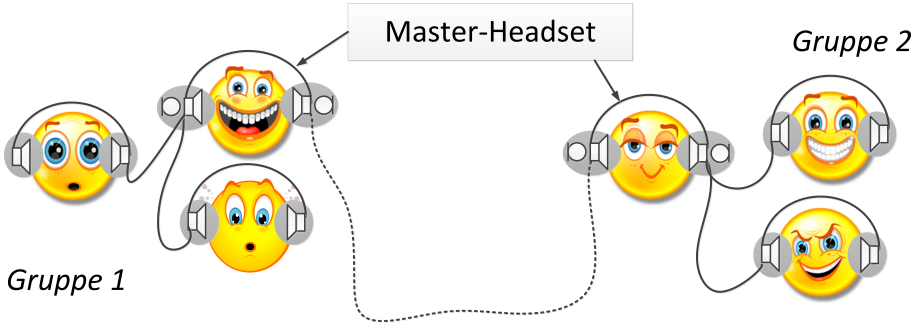

Abbildung 3: **Gruppenkommunikation**: Gruppen von Personen befinden sich an unterschiedlichen Orten und führen ein Gespräch mittels *HD-Voice-3D*-Technologie.

teristiska eines natürlichen Kopfes nachbildet, wie etwa der in Abbildung 2 dargestellte Kunstkopf. Die Einzelperson trägt ebenfalls ein *HD-Voice-3D*-Endgerät und erfährt dadurch nicht nur wie bei der herkömmlichen Telefonie die sprachlichen Inhalte, die die einzelnen Teilnehmer von sich geben: Anhand der Richtungsinformation, die in dem binauralen Signal enthalten sind, können Beiträge deutlich besser einzelnen Personen zugeordnet werden. Sofern mehrere Sprecher gleichzeitig sprechen, kann durch Verwendung von *HD-Voice-3D* auch eine signifikante Erhöhung der Verständlichkeit erreicht werden.

- ***HD-Voice-3D* basierte Gruppenkommunikation**
  Eine Gruppenkommunikation mittels *HD-Voice-3D* ist in Abbildung 3 dargestellt. Dabei sind auf beiden Seiten der Verbindung mehrere Personen eingebunden. Ein Teilnehmer auf jeder Seite hat ein "Master-Headset", das mit Lautsprechern und Mikrophonen versehen ist. Alle anderen Teilnehmer tragen normale Headsets ohne Mikrophone. Die Teilnehmer können sich innerhalb der Gruppe der jeweiligen Seite direkt miteinander unterhalten. Durch den *HD-Voice-3D* Link zur jeweils anderen Gruppe jedoch kann auch eine Kommunikation zwischen Teilnehmern der jeweils anderen Gruppe stattfinden. Dadurch, dass die Richtungsinformation mit übertragen wird, sind auch mehrere Teildiskussion unter einzelnen Teilnehmern zeitgleich möglich, so, als würden sich die Teilnehmer beider Gruppen in einem Raum befinden.

2.2 Übertragungstechnik für *HD-Voice-3D*

Es hat sich in der Vergangenheit gezeigt, dass eine Verbreitung von einer neuen Technologie wie *HD-Voice-3D* gegenwärtig nur über Übertragungskanäle möglich ist, bei denen der Anwender bzw. der Betreiber des Services eine Ende-zu-Ende-Kontrolle besitzen. Die Telefonnetze von Telekommunikationsanbietern kommen aufgrund mangelnder Kooperation nicht in Frage. Durch die weite Verbreitung von Smartphones mit ausreichender Rechenleistung und der nahezu flächendeckende mobile hochratige Zugang zum Internet in den meisten urbanen Ballungsgebieten - insbesondere in LTE-Netzen [3GP11] - stellt dies jedoch keine Einschränkung mehr dar. Hieraus resultiert jedoch, dass es sich nur um

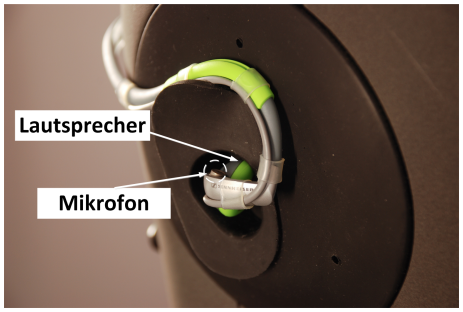

Abbildung 4: Prototypisches Endgerät für die Aufnahme und Wiedergabe binauraler Signale.

paketbasierte Übertragung handeln kann. Letzteres wiederum erfordert u.a. technische Lösungen zur Berücksichtigung von Paketverlusten und -verzögerungen.

2.3 Endgeräte für *HD-Voice-3D*

Für *HD-Voice-3D* werden neuartige Endgeräte benötigt. Anstatt ein einzelnes Mikrophon dicht an dem Mund des Sprechers zu platzieren (wie z.B. bei der *konventionellen Telefonie*), werden **zwei** Mikrophone benötigt, die jeweils in der Nähe des natürlichen Ortes, der der Schallaufnahme beim Menschen dient, platziert werden. Eine mögliche Realisierung ist in Abbildung 4 am Beispiel eines Kunstkopfes demonstriert. Die Mikrophonkapseln sind in der Nähe des Eingangs des Ohrkanals befestigt. Um einen Einsatz entsprechend dem Szenario **Gruppenkommunikation** zu gewährleisten, ist darauf zu achten, dass trotz des Kopfhöhrers Schall der lokal anfällt, direkt das Trommelfell erreicht. Wäre dies nicht der Fall, wäre eine lokale Kommunikation innerhalb einer Gruppe im Szenario **Gruppenkommunikation** nicht möglich. Bei der Verwendung von geschlossenen Kopfhörern oder Ohrkanal-Hörern müssten hierzu die lokalen Mikrophonsignale direkt an den Kopfhöhrer weitergeleitet werden. Hierbei ist zu beachten, dass die Zeitverzögerung zwischen Mikrophon und Kopfhöhrer sehr gering sein muss, damit eine Überlagerung von Direktschall und Kopfhöhrersignal nicht zu Verzerrungen und Artefakten führen. Eine technisch einfacher zu realisierende Variante ist die Verwendung von offenen Kopfhöhrern, wie in Abbildung 4 gezeigt, bei denen das lokale Signal direkt das Trommelfell erreicht.
Besondere Berücksichtigung bedarf die Tatsache, dass gegenwärtig verfügbare Smartphones nur einen monauralen Eingangskanal besitzen. An ausgewählte Smartphones kann jedoch unter bestimmten Randbedingungen eine Stereo-USB-Soundkarte angeschlossen werden.

3 Signalverarbeitungskette bei *HD-Voice-3D*

Prinzipiell entsprechen die funktionalen Blöcke der Signalverarbeitung bei *HD-Voice-3D* denen bei einem konventionellen VoIP-Terminal. Ein Überblick ist schematisch in Abbildung 5 gegeben. Besonderes Unterscheidungsmerkmal ist natürlich zunächst die Tatsache,

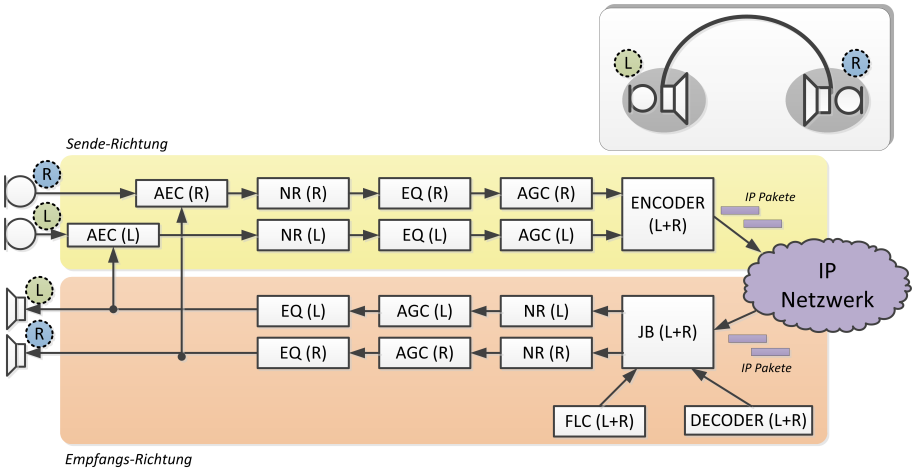

Abbildung 5: Signalverarbeitungskette eines *HD-Voice-3D*-Terminals.

dass ein *HD-Voice-3D*-Terminal jeweils zwei Mikrophone und zwei Lautsprecher besitzt, die in der Abbildung dem binauralen Endgerät jeweils auf der rechten oder der linken Seite zugeordnet sind.
Integraler Bestandteil jedes *HD-Voice-3D*-Terminals ist der doppelt ausgeführte akustische Echokompensator (**AEC**). Dieser ist der Tatsache geschuldet, dass der verwendete *HD-Voice-3D*-Kopfhörer in der Regel nicht geschlossen sein kann und das Mikrophon sich sehr dicht am Lautsprecher befindet, um insbesondere das Szenario der Gruppenkommunikation realisieren zu können. Es kommt zu einer starken Rückkopplung des Signals des fernen Sprechers, dem entgegengewirkt werden muss, da sonst eine Kommunikation nicht ungestört möglich ist. Eine Rückkopplung des **linken** Lautsprechersignals zum **rechten** Mikrophon und anders herum wurde bei prototypischen Tests mit dem verwendeten Headset nicht beobachtet: Messungen haben ergeben, dass hier eine natürliche Dämpfung von ca. 30 - 40 dB mit dem Prototypen-Headset ausreicht. Diese ist dadurch bedingt, dass der Lautsprecher eine gewisse Einstecktiefe in den Ohrkanal aufweist und sich auf der jeweils abgeschatteten Seite des Kopfes befindet. Somit handelt es sich nur um einen doppelt ausgeführten einkanaligen Echokompensator und nicht um ein mehrkanaliges Echokompensationsproblem.
In *Sende-Richtung* folgen den Echokompensatoren ein doppelt ausgeführter Störreduktionsblock (**NR**), ein Equalizer (**EQ**) und eine automatische Lautstärkekontrolle (**AGC**). Das Ausgangssignal des **AGC**-Blocks wird schliesslich dem **Encoder** des Quellcodierers zugeführt. Es muss ein besonderer Quellcodierer eingesetzt werden, um das aus zwei Kanälen bestehende binaurale Signal in einen Strom von IP Paketen mit moderater Datenrate zu wandeln, die mittels IP Link über das IP Netzwerk an den Partner übertragen werden.
In *Empfangs-Richtung* erreicht das binaurale Signal das Terminal mittels IP Link in Form von aufeinander folgenden Paketen. Zunächst werden diese Pakete dem Adaptiven Jitterbuffer (**JB**) zugeführt, der gleichzeitig auch die Kontrolle über den **Decoder** der Quellcodierung hat. Im Adaptiven Jitterbuffer werden Netzwerk-Verzögerungen (sogenannte

Netzwerk-Jitter) ausgeglichen. Möglicherweise verlorenene Pakete werden mittels Frame-Loss-Concealment (**FLC**) derart „verdeckt“, dass Lücken im Signal unhörbar werden (soweit dies möglich ist). Das Ausgangssignal des Jitterbuffers wird einer optionalen Signalverarbeitungskette zugeführt, die aus Störreduktion (**NR**) und automatischer Lautstärkekontrolle (**AGC**) besteht. Diese Funktionalitäten sind im Optimalfall bereits auf der Seite des verbundenen Terminals des fernen Sprechers realisiert, können aber dennoch sinnvoll sein, wenn entweder das andere verbundene Terminal des fernen Sprechers nicht die erwünschte Audioqualität liefert oder um im Fall eines schlechten Quellcodierers, z.B. Codierartefakte zu beseitigen.
Der sich in *Empfangs-Richtung* anschliessende Equalizer (**EQ**) dient dazu, den verwendeten Kopfhörer des nahen Sprechers zu entzerren und möglicherweise persönliche Vorzüge des Endgeräteträgers abzubilden. Eine individuelle Entzerrung ist an dieser Stelle besonders wichtig, da die Qualität des erreichten räumlichen Eindruck bei der Wiedergabe von binauralen Signalen stark von der Anpassung des Wiedergabesystems an den Träger abhängt. Das Signal wird anschliessend an den Echokompensatorblock weitergeleitet und schliesslich über die Lautsprecher abgespielt.

4 Besonderheiten der Audiosignalverarbeitung bei *HD-Voice-3D*

4.1 Binaurale Cues

Im Vergleich zur Signalverarbeitung bei der *konventionellen (monauralen) Telefonie* muss bei *HD-Voice-3D* darauf geachtet werden, dass die sogenannten *binauralen Cues* erhalten bleiben. Dabei handelt es sich um Eigenschaften der Relation der beiden Teilsignale eines zweikanaligen Binauralsignals, die in der Regel als sogenannte *Interaural Time Differences* (ITD) und *Interaural Level Differences* (ILD) ausgedrückt werden [Bla83]. Die ITD-Merkmale beeinflussen die räumliche Zuordnung von Schallereignissen bei tiefen Frequenzen und basieren auf einer zeitlichen Relation des Eintreffens der Wellenfronten zunächst an dem einen und dann an dem anderen Ohr. Man spricht in diesem Zusammenhang häufig auch von Phasenverschiebungen zwischen den beiden binauralen Teilsignalen. Bezüglich dieser Phasenverschiebungen ist die menschliche Wahrnehmung sehr empfindlich, eine Verschiebung bereits um Bruchteile einer Millisekunde haben stark hörbare Auswirkungen. Dies wird schnell klar, wenn man berücksichtigt, dass ein Schallereignis bei einer Schallgeschwindigkeit von $c = 340$ m/s sich in gerade einmal ungefähr 0.5 ms vom rechten zum linken Ohr (Entfernung ca. 18 cm) ausbreitet.
Im Gegensatz dazu haben die ILD-Merkmale eine starke Bedeutung bei höheren Frequenzen und gehen zurück auf die Abschattungseffekte des Kopfes bei Eintreffen eines Signals aus einer bestimmten Richtung und damit einer reduzierten Amplitude auf der dem Schallereignis abgewandten Seite des Kopfes. Für ITD aber auch für ILD entstehen Mehrdeutigkeiten, die dadurch entstehen, dass z.B. für Quellen, die auf einer Kreisbahn um die Ohrachse herum bewegt werden die Werte für ITD und ILD unverändert bleiben. Auf diesen sogenannten „Cones of Confusion“ kommt es vermehrt zu Fehlortungen z.B. zwischen vorne und hinten aber auch bezüglich der Quellenhöhe. Zur Verbesserung der Ortung werden daher weitere Cues benötigt, die aus der Beugung des Schalls an Schulter, Kopf und Ohrmuschel entstehen und in Form von charakteristischen Anhebungen und Absenkungen bestimmter Frequenzbänder [Bla83] zu einer klanglichen Veränderung beitragen.

4.2 Quellcodierung bei *HD-Voice-3D*

Die Empfindlichkeit der menschlichen Wahrnehmung gegenüber Phasenverschiebungen in binauralen Signalen muss bei der Quellcodierung mit berücksichtigt werden. VoIP Anwendungen tendieren dazu, verschiedene Signaltypen in unabhängigen Streams zu versenden und auf Empfangsseite wieder zu synchronisieren. Dies macht durchaus Sinn bei der Kombination von Video und Audio, da diese in der Regel von verschiedenen physikalischen Einheiten mit unabhängigen Aufnahmetakten (Clocks) aufgezeichnet werden. Eine Synchronisation ist hier jedoch auch relativ unkritisch, so dass für Film und Fernsehen ein maximaler Versatz zwischen Audio und Video in der Größenordnung von $15 - 45$ ms empfohlen wird [C$^+$03].
Dieses Vorgehen ist bei den beiden Kanälen einer *HD-Voice-3D*-Übertragung nicht denkbar. Insbesondere sollten die Signale beider Kanäle mit physikalischen Einheiten die einen gemeinsamen Aufnahmetakt besitzen, aufgezeichnet werden. Eine nachträgliche Synchronisation ist sonst nicht oder nur mit großem Aufwand mit einer Genauigkeit erreichbar, die den Randbedingungen der ITD-Merkmale bzgl. Phasendifferenz entspricht.
Prinzipiell ist die Übertragung der codierten binären Daten zweier unabhängiger identischer monauraler Quellcodierer in einem Datenpaket (mit doppelter Länge) der einfachste Ansatz, sofern linker und rechter Kanal sample- und rahmensynchron zugeführt werden. Dies führt jedoch im Vergleich zum einkanaligen *HD-Voice*-System zu einer Verdopplung der Datenrate. Abhilfe verschafft hier der Einsatz von Quellcodierverfahren, die Redundanzen zwischen den Kanälen ausnutzen [Fuc93, KV08a]. Viele der heute weit verbreiteten Verfahren zum sogenannten *Joint-Stereo* (z.B. [BF08, HBL94]) sind jedoch mit Vorsicht zu geniessen, da hier häufig die Phasenbeziehungen zwischen linkem und rechtem Kanal vernachlässigt und ledigliche die Amplitudenunterschiede berücksichtigt werden, was zu einem eingeschränkten räumlichen Eindruck eines Schallereignisses führt.
Insgesamt muss die Qualität der Quellcodierung bei *HD-Voice-3D* höher sein als bei monauralen Verfahren, da Codierungsartefakte als unabhängige Rauschprozesse modelliert mit beiden Ohren besser wahrgenommen werden können als mit einem Ohr, insbesondere wenn die ILD und ITD-Merkmale der Störung und des Nutzsignals nicht übereinstimmen. Dieser Effekt wird durch die *Binaural Masking Level Differences* beschrieben [MM03].

4.3 Adaptiver Jitterbuffer

VoIP-Übertragungen basieren in der Regel auf dem sogenannten *User Datagram Protocol* (UDP). Aufgrund der weiten Verbreitung von einschränkenden Firewalls werden VoIP Anwendungen aber auch häufig über das *Transmission Control Protocol* (TCP) betrieben. In beiden Fällen kann es passieren, dass Pakete signifikant verspätet beim Empfänger ankommen (auch *Netzwerk-Jitter* genannt). Im Falle von UDP kann es auch sein, dass Pakete gar nicht beim Empfänger eintreffen (auch als *Frameloss* bezeichnet). Sofern ein Paket nicht rechtzeitig eintrifft, kommt es auf Seiten des Empfängers zu Lücken bei der Wiedergabe des Audiosignals.
Um Paket-Verzögerungen auszugleichen, bietet es sich an, einen gewissen Vorrat von Abtastwerten am Empfänger vorzuhalten. Diesem Zweck dient der sogenannte Jitterbuffer (JB). Der Vorhalt von Abtastwerten darf jedoch auch nicht zu groß sein, da sonst eine hohe Ende-zu-Ende-Verzögerung bei der Kommunikation auftritt.

Die beobachteten Verzögerungen auf dem Netzwerk sind in der Regel sehr stark zeitvariabel. Ein alltägliches Beispiel für stark variierende Netzwerk-Jitter stellt jeder haushaltsübliche WLAN Router dar, der die Übertragung immer wieder für mehrere Hundert Millisekunden aussetzt, wenn z.B. eine Mikrowelle verwendet wird und diese im gleichen Frequenzband Störungen produziert oder z.B. ein Mobiltelefon eine Bluetooth-Verbindung im gleichen Frequenzband aufbaut. Ein guter Jitterbuffer adaptiert sich an die gegenwärtig vorliegende Netzwerkjitter-Charakteristik, welche somit ebenfalls überwacht werden muss. Ein solcher Jitterbuffer wird auch als *Adaptiver Jitterbuffer* bezeichnet.
Bei der Anpassung des Adaptiven Jitterbuffers an sich ändernde Übertragungszeiten wird die Füllhöhe des Jitterbuffers modifiziert. Dies geschieht in der Regel durch signalmodifizierende Verfahren wie z.B. dem Waveform Similarity Overlap-Add (WSOLA) [VR93], einem Verfahren aus der Familie der Phase-Vocoder [Dol86] oder ähnlichen Ansätzen. Grundprinzip all dieser Verfahren ist es, ein Audiosignal schneller oder langsamer als ursprünglich abzuspielen, ohne dass Audioartefakte entstehen (engl.: Time-Stretching). Für binaurale Signale kann hier jedoch nicht beliebig vorgegangen werden: Beim WSOLA zum Beispiel wird das mit anderer Geschwindigkeit abgespielte Signal durch neuerliches Zusammenfügen von Signalsegmenten erzeugt, die unterschiedlichen zeitlichen Ursprüngen zugeordnet waren. Es kann dabei zu Phasenverschiebungen kommen, die zwar für sich genommen nicht hörbar sind, aber als Manipulation der ITD-Merkmale im Zusammenhang mit *HD-Voice-3D* fatale Auswirkung haben können. Das gleiche Problem betrifft auch den Phasevocoder und viele andere verwandte Technologien, die auf Phasenmanipulation der Audiosignale basieren.
Ein möglicher Ansatz, der die ITD-Merkmale nicht beeinflusst, ist ein adaptiver Resampler. Basis für diesen Ansatz ist ein flexibler Resampler, dessen Ausgangsabtastfrequenz zur Laufzeit auch in kleinen Schritten in beiden Kanälen auf identische Art und Weise variiert werden kann. Ein solcher adaptiver Resampler ist z.B. in [PEV10] vorgestellt. Einziger Nachteil bei diesem Ansatz ist eine leichte Variation der Tonhöhe bei Änderung der Abspielgeschwindigkeit (Leiern), was jedoch bei rein auf Sprache ausgerichteten Anwendungen nicht weiter auffällt.

4.4 Frameloss-Concealment

Beim Frameloss-Concealment (FLC) werden Signallücken verdeckt, die trotz Adaptiven Jitterbuffers durch den Verlust von Übertragungspaketen auftreten können. Dabei wird in den meisten Fällen basierend auf der Analyse des Signals vor der Lücke aus der Vergangenheit ein künstliches Signal erzeugt, das an den Grenzen der Lücke angepasst wird, um einen kontinuierlichen Verlauf zu erreichen. Auch hier kann es schnell zu Problemen bei der binauralen Wahrnehmung kommen, wenn in beiden Kanälen unterschiedliche Signale künstlich erzeugt werden. Als Konsequenz werden im Raum fluktuierende Phantomsignale hörbar, die als sehr unangenehm wahrgenommen werden können.

4.5 Automatische Lautstärkekontrolle

Bei der automatischen Lautstärkekontrolle (AGC) werden Signalpegel angepasst, so dass die über den Kanal zu übertragenden Signale weder zu laut noch zu leise sind. Dies verbessert in der Regel die wahrgenommene Gesamtqualität, da ein gut ausgesteuerter Quellco-

dierer weniger Artefakte produziert, als ein schlecht ausgesteuerter Quellcodierer. Darüber hinaus ist natürlich ein gut ausgesteuertes Sprachsignal in der Regel auch besser zu verstehen als ein schlecht ausgesteuertes.
Die ILD-Merkmale eines binauralen Signals basieren auf Pegelunterschieden der beiden Kanäle des binauralen Signals. Sofern unabhängig voneinander operierende AGCs zum Einsatz kommen, werden diese wichtigen Pegelunterschiede zerstört. Der Einsatz zweier herkömmlicher AGCs ist also für *HD-Voice-3D* nicht möglich, stattdessen müssen die beiden funktionalen Blöcke gekoppelt werden.

4.6 Equalizer

Bei der binauralen Audiowiedergabe werden in den meisten Fällen die Ohrsignale nicht mit den eigenen Ohren, sondern mit einem Kunstkopf aufgenommen, dessen Ohren den physikalischen Abmessungen von „Durchschnittsohren“ entsprechen. Hierbei kommt es zwangsläufig zu mehr oder weniger großen Abweichungen bei ITD, ILD und den die Klangfarbe verändernden Beugungserscheinungen. Die Lokalisationsleistung beim Hörer nimmt hierdurch ab. Es konnte gezeigt werden, dass die Individualisierung der sogenannten HRTF (Head Related Transfer Functions) zu einer deutlichen Verbesserung des räumlichen Eindrucks führt. Allerdings hört bei *HD-Voice-3D* der nahe Sprecher immer mit den „fremden“ Ohren des fernen Sprechers. Auf Grund der möglichen Vielzahl von Endgeräten unterschiedlicher Hersteller, die frei untereinander kombinierbar sein sollten, muss daher sichergestellt werden, dass für das übertragene binaurale Signal eine wohldefinierte, neutrale und für alle Sendeseiten identische Entzerrung gewährleistet ist. Dies erlaubt für die Wiedergabeseite dann auch eine individuelle Entzerrung des Hörers auf die tragende Person ohne dabei auf die aufnehmende Seite Rücksicht zu nehmen.
In Sende-Richtung des binauralen Terminals nach Abbildung 5 sollte daher ein Entzerrer dafür eingesetzt werden, die Mikrophoncharakteristika auszugleichen. Diese Standardisierung des Sendesignals hinsichtlich Bandbreite, Frequenzgang und Pegel wäre unabhängig von der Wiedergabeseite möglich und würde durch die Bauart (Mikrofonanordnung etc.) bedingte Fehler korrigieren.
Wiedergabeseitig wird ein Equalizer benötigt, der die Übertragungsfunktion des Kopfhörers an einem Norm-Ohr (z.B. ITU-T Standard) oder falls möglich an dem realen Ohr des Trägers so korrigiert, dass die senderseitig angebotenen Signale möglichst fehlerfrei zu Gehör gebracht werden. In Empfangs-Richtung kann zusätzlich ein Equalizer, der über die Entzerrung des Endgeräts im Sinne der beschriebenen individualisierten binauralen Wiedergabe hinausgeht eingesetzt werden, um die Vorlieben des Hörers abzubilden. Defizite des Nutzers beim Hören z.B. im Falle älterer Benutzer, sollten ebenfalls an dieser Stelle berücksichtigt werden, um die Verständlichkeit zu verbessern. Auch hier gilt es jedoch, durch die Betonung bestimmter Frequenzbereiche die notwendigen Cues der binauralen Signale nicht zu zerstören.

4.7 Störreduktion

Ziel von *HD-Voice-3D* ist die Übertragung sowohl der sprachlichen Inhalte als auch der akustischen Umgebung inklusive der Umgebungsgeräusche. In diesem Sinne scheint eine Störreduktion zunächst nicht sinnvoll. Stellt man sich jedoch vor, dass bei dem Anwen-

dungsszenario der **Besprechung mit mehreren Teilnehmern** dauerhaft Lüfterhintergundgeräusche vorhanden sind, ist es naheliegend, zumindest stationäre Störungen herauszufiltern.
Die Ansätze zur Bestimmung der Hintergrundgeräusche bei der einkanaligen Störreduktion basieren in der Regel auf der Ausnutzung der Eigenschaft von Sprache, dass immer wieder Sprachpausen auftreten in denen Hintergrundgeräusche analysiert werden [VM06]. Bei *HD-Voice-3D* muss eine genauere Klassifikation der Aufnahmesituation stattfinden, um „gewünschte“ Hintergrundgeräusche von „ungewünschten“ unterscheiden zu können. Zum Vorteil ist hier jedoch das Vorhandensein von zwei Mikrophonen, mit denen nicht nur Schallereignisse selbst, sondern auch deren Positionen im Raum bestimmt werden können. Zum Beispiel kann als Zusatzinformation auch die Position der Mikrophone, die in der Regel variabel ist, durch Sensoren (z.B. Beschleunigungssensoren, Kompass) bestimmt und ausgewertet werden.
Prinzipiell basiert eine Störreduktion auf der frequenzselektiven Manipulation der Amplitude der gestörten Audiosignale derart, dass das aufgenommene Signal in den Frequenzbereichen bedämpft wird, in denen sich überwiegend Störungen befinden. Hingegen werden diejenigen Frequenzanteile ungefiltert durchgelassen, in denen Sprachanteile vorhanden sind. Im Sinne der Beibehaltung der ILD-Merkmale bei *HD-Voice-3D* können unabhängig operierende herkömmliche Ansätze zur Störreduktion schnell zu einer Verfälschung oder Zerstörung der räumlichen Wahrnehmung führen. Aus diesem Grunde muss ein Ansatz ähnlich den z.B. in [Lot04] oder [Jeu12] vorgestellten Ansätze verfolgt werden, die für binaurale Hörgeräte entwickelt wurden.

4.8 Akustische Echokompensation

Bei der akustischen Echokompensation wird in der Regel ein Ansatz verfolgt, der aus einem **Echocanceller**- und einem **Postfilterteil** besteht. Der **Echocancellerteil** basiert in der Regel auf der Anwendung eines adaptiven Filters, das den physikalischen Echopfad nachbilden soll und dann dazu dient, das aufgenommene Echosignal durch Subtraktion eines künstlich erzeugten geschätzten Echosignals auszulöschen [Hay00]. Durch die Physik der Akustik kommt es hier aber immer zu einer gewissen Unsicherheit bei der Nachbildung des Echopfades, die dazu führt, dass doch noch ein Restecho vorhanden ist [Enz06]. Der Beseitigung dieses Restechos dient das **Postfilter**, das durch frequenzselektive Dämpfung verschiedener Signalanteile charakterisiert und der Störreduktion in seiner Wirkungsweise prinzipiell ähnlich ist.
Bei der binauralen Telefonie ist durch die starke Kopplung zwischen Mikrophon und Lautsprecher, die wiederum ihrer lokalen Nähe geschuldet ist, mit einem starken Echosignal zu rechnen. Dieses kann durch einen guten Echokompensator in der Regel um einige dB gedämpft, jedoch nicht komplett beseitigt werden. Als Konsequenz verbleiben mehr oder weniger zufällige und vom Klang her unangenehme Signalfragmente im Eingangssignal, die - wie Phantomsignale - so erscheinen, als kämen sie fluktuierend aus beliebigen Richtungen. Ein Postfilter ist hier von großer Bedeutung, hat jedoch wiederum einen kontraproduktiven Einfluss auf die binauralen Cues des aufgenommenen Signals. An dieser Stelle besteht aus unserer Sicht im Hinblick auf *HD-Voice-3D* in Zukunft der größte Bedarf für neuartige Algorithmik.

5 Demonstrator

Zur Demonstration von *HD-Voice-3D* wird auf dem WASP-Workshop 2013 ein Echtzeitdemonstrator vorgeführt, mit dem binaural von einem Ort des Veranstaltungsgeländes zu einem anderen telefoniert werden kann. Je nach den Gegebenheiten vor Ort wird hierzu entweder eine LAN- der eine WLAN-Verbindung zum Einsatz kommen. Dargestellt werden soll das Szenario der ***HD-Voice-3D* Gruppenkommunikation**.
Die in dem Demonstrator eingesetzten *HD-Voice-3D*-Terminals basieren dabei auf dem *Software-defined-Terminal* (SDT), das auf der IWAENC 2012 vorgestellt wurde [KSRV12] und wiederum auf dem RTProc-System basiert [KV08b]. Das SDT realisiert unterschiedliche Verfahren zur Umsetzung der Teilfunktionalitäten der dargestellten Signalverarbeitungskette, unter anderem:

- Verschiedene Verfahren der Quellcodierung, z.B. der Adaptive Multiraten (AMR) Codec für Schmalband- und Breitbandsignale ([Rec96] und [3GP01, ITU02]). Dabei weichen die verwendeten Umsetzungen jedoch in soweit vom Standard ab, dass sie aus zwei unabhängigen Codec-Einheiten bestehen, die parallel (sample- und rahmensynchron) betreiben werden, deren Bitstrom jedoch in gemeinsamen Paketen übertragen werden. Darüber hinaus steht auch der Opus Codec zur Verfügung [VVT12], der eine ganze Reihe von Qualitätsanforderungen, Bandbreiten und sowohl Mono als auch Stereo (binaural) unterstützt.
- Eine TCP basierte (proprietäre) Verbindung zwischen den beiden Kommunkationsendpunkten.
- Ein Adaptiver Jitterbuffer, der basierend auf dem Verfahren des Resamplings realisiert wurde, damit die binauralen Merkmale von einem zum anderen Ende ohne Verluste transportiert werden können.
- Ein einfaches Frameloss-Concealment, das unangenehme Artefakte bei Rahmenverlusten verdeckt.
- Verfahren der akustischen Echokompensation und der Störreduktion, die in Hinblick auf binaurale Signale entwickelt wurden, sowie auch andere Ansätze, bei denen in beschriebener Art und Weise die binauralen Cues angegriffen werden.

Um die Auswirkungen verschiedener Parametereinstellungen und Algorithmen zu demonstrieren, können Änderungen der Einstellungen zur Laufzeit vorgenommen werden. Ein kleiner Vorgeschmack zur Leistungsfähigkeit der Binauralen Telefonie ist unter www.binaural-telephony.com oder [bin] verfügbar.

6 Zusammenfassung

In diesem Beitrag wird gezeigt, dass der nächste große Schritt in der Evolution der Sprachkommunikation *HD-Voice-3D* ist. Die Grundzüge der auch *Binaurale Telefonie* genannten Technologie und die wesentlichen Unterschiede zu vorhandenen Kommunikationstechniken werden vorgestellt. Des weiteren werden verschiedene Nutzungsszenarien aufzeigt,

die an vielen Stellen des privaten und geschäftlichen Alltags Anwendung finden: Telefonieren mit Übertragung der natürlichen akustischen Atmosphäre, die Teilnahme an einer Besprechung mit mehreren Teilnehmern oder die Gruppenkommunikation.

Die *Binaurale Telefonie* hält einige neue Herausforderungen in der Audiosignalverarbeitung bereit, die fast alle Teilfunktionalitäten eines *HD-Voice-3D*-Endgeräts betreffen. Neben den Aufnahme- und Wiedergabesystemen gilt es in der Audioverarbeitung und Audioübertragung den räumlichen Eindruck, die sogenannten „Binauralen Cues“ des Signals zu erhalten. Im Bereich der Quellcodierung, des Frameloss-Concealment, der Automatische Lautstärkekontrolle, des Equalizers, der Störreduktion und der akustischen Echokompensation sind Anpassungen der einkanaligen Algorithmen notwendig. Insbesondere im Bereich des Adaptiver Jitterbuffers, der Schwankungen der Übertragungsdauer und Übertragungsfehler einzelner Pakete ausgleichen soll, wird technologisches Neuland betreten.

Auf dem WASP-Event „HD-Voice-3D zum Anfassen“ wird ein Echtzeitdemonstrator vorgestellt, der eine *HD-Voice-3D* Kommunikation ermöglicht und den Einfluss der verschiedenen Parameter jeder Komponente der Signalverarbeitung im laufenden Betrieb hör- und erlebbar macht.

Literatur

[3GP01] 3GPP TS 26.190. Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions. 3GPP Technical Specification Group Radio Access Network, 2001.

[3GP05] 3GPP TS 26.290. Audio codec processing functions – Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions. 3GPP Technical Specification Group Radio Access Network, 2005.

[3GP11] 3GPP TS 36.212. Evolved Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding. 3GPP Technical Specification Group Radio Access Network, April 2011.

[BF08] Jeroen Breebaart und Christof Faller. *Spatial audio processing: MPEG Surround and other applications*. Wiley-Interscience, 2008.

[bin] Binaural Telephony, 2013, http://www.ind.rwth-aachen.de/de/forschung/speechaudiocommunication/binaural-telephony.

[Bla83] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. The MIT press, Cambridge, Massachusetts, 1983.

[C+03] Advanced Television Systems Committee et al. ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations. *IS-191*, 26, 2003.

[Dol86] Mark Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

[Enz06] Gerald Enzner. *A Model-Based Optimum Filtering Approach to Acoustic Echo Control: Theory and Practice*. Dissertation, IND, RWTH Aachen, April 2006.

[FKM12] Frank Felden, Thomas Krüger und Eirini Markoula. How IT Is Driving Business Value at European Telcos. *bcg.perspectives*, 2012.

[Fuc93] H. Fuchs. Improving joint stereo audio coding by adaptive inter-channel prediction. In *Applications of Signal Processing to Audio and Acoustics, 1993. Final Program and Paper Summaries., 1993 IEEE Workshop on*, Seiten 39–42, 1993.

[Hay00] Simon Haykin. Adaptive filter theory, 1996, 2000.

[HBL94] Jürgen Herre, Karlheinz Brandenburg und D. Lederer. Intensity Stereo Coding. In *Audio Engineering Society Convention 96*, 2 1994.

[ITU02] ITU-T G722.2. Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB) , 2002.

[Jeu12] Marco Jeub. *Joint Dereverberation and Noise Reduction for Binaural Hearing Aids and Mobile Phones*. Dissertation, IND, RWTH Aachen, August 2012.

[KSRV12] Hauke Krüger, Thomas Schlien, Matthias Rüngeler und Peter Vary. The Software Defined Terminal. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*. RWTH Aachen University, September 2012. Demonstrator Session.

[KV08a] Hauke Krüger und Peter Vary. A New Approach for Low-Delay Joint-Stereo Coding. In *ITG-Fachtagung Sprachkommunikation*. VDE Verlag GmbH, Oktober 2008.

[KV08b] Hauke Krüger und Peter Vary. RTPROC: A System for Rapid Real-Time Prototyping in Audio Signal Processing. In *Proceedings of IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*, Seiten 311–314. IEEE, Oktober 2008.

[Lot04] Thomas Lotter. *Single and Multimicrophone Speech Enhancement for Hearing Aids*. Dissertation, IND, RWTH Aachen, 2004.

[MM03] Brian CJ Moore und Brian C Moore. *An introduction to the psychology of hearing*, Jgg. 4. Academic press San Diego, 2003.

[PEV10] Matthias Pawig, Gerald Enzner und Peter Vary. Adaptive Sampling Rate Correction for Acoustic Echo Control in Voice-Over-IP. *IEEE Transactions on Signal Processing*, 58(1):189 – 199, Januar 2010.

[Rec96] ETSI Rec. GSM 06.60 Digital Cellular Telecommunications System; Enhanced Full Rate (EFR) Speech Transcoding. *Digital Cellular Telecommunication System (Phase 2+)*, 1996.

[RKSV12] Matthias Rüngeler, Hauke Krüger, Thomas Schlien und Peter Vary. Spatial Audio Conferencing using Binaural HD Voice. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*. RWTH Aachen University, September 2012. Demonstrator Session.

[TW07] Ulrich Trick und Frank Weber. *SIP, TCP/IP und Telekommunikationsnetze: Next Generation Networks und VoIP-konkret*. Oldenbourg Verlag, 2007.

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

[VR93] W. Verhelst und M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, Jgg. 2, Seiten 554–557 vol.2, 1993.

[VVT12] JM. Valin, K. Vos und T. Terriberry. Definition of the Opus Audio Codec. RFC 6716, September 2012.