Qualität übertragener Sprache in drahtlosen Ad-Hoc-Netzwerken

Tobias Hübschen, Bastian Kaulen, Mohammedul Emin-Yurdakul, Gerhard Schmidt
Digitale Signalverarbeitung und Systemtheorie, 24143 Kiel, Deutschland, Email: thu@tf.uni-kiel.de

Kurzfassung

Die Sprachkommunikation über das Mobilfunknetz ist heutzutage absoluter Standard und zeichnet sich dabei vor allem durch Flexibilität bezüglich des Standortes aus. Es existieren jedoch lokale Anwendungen, für die der etablierte Mobilfunk nicht flexibel genug ist. Gründe hierfür können beispielsweise in der Netzverfügbarkeit, extremen Latenz- und Sicherheitsanforderungen oder der gewünschten Kommunikationsform an sich liegen. Eine Alternative zum klassischen Mobilfunk stellen drahtlose Ad-Hoc-Netzwerke dar, deren Protokolle jeweils für die angestrebte lokale Anwendung modifiziert werden können. In diesem Beitrag werden solche experimentelle Netzwerke in Hinblick auf die Qualität der übertragenen Sprache untersucht. Die Analyse erfolgt dabei in Abhängigkeit verschiedener Netzwerktopologien und Anwendungsfälle.

Einleitung

Zur Übertragung (beliebiger) Daten wird oft auf bestehende Netzwerke Dritter, sogenannter Netzbetreiber, zurückgegriffen. Allerdings ist es insbesondere bei Einsätzen von Sicherheitsbehörden kritisch, wenn die Kommunikationsfähigkeit sich auf Dienstleistungen anderer stützt. Da es sich in diesem Kontext oft um eine lokal eingegrenzte Kommunikation handelt, sind hier eigene Ad-Hoc-Netzwerke eine gute Alternative. Solche Netzwerke bieten: Flexibilität, da sie für die Ansprüche einer bestimmten Nutzergruppe entworfen werden; Verfügbarkeit, da sie unabhängig vom Netzausbau der Netzbetreiber sind und Sicherheit, da sie von anderen großen Netzwerken entkoppelt sind und zusätzliche Sicherheitsprotokolle implementieren können.

Besonders flexibel sind Ad-Hoc-Netzwerke, welche auf spezieller Ad-Hoc-Hardware [1] basieren. Die hier referenzierte Plattform setzt beispielsweise auf eine Datenübertragung in den offenen Frequenzbändern um 2,4 und 5 GHz und implementiert auf klassischen WLAN-Chips ein neues MAC-Protokoll für ein deterministisches Übertragungsverhalten mit hoher Robustheit. Für Ad-Hoc-Netzwerke allgemein besonders wichtig ist dabei der verwendete dezentrale Netzwerkansatz. Dass sich solch eine Plattform prinzipiell auch für Sprachübertragung eignet, ergibt sich aus der niedrigen beworbenen Latenz, der Freiheit zur Übertragung beliebiger Nutzdaten sowie der Möglichkeit zur schnellen Rekonfiguration des Netzwerkes.

Ist die Netzwerk-Plattform gegeben, so ist besonders die Netzwerktopologie entscheidend für die Performanz des Netzwerkes. Performanz fasst hier sowohl Datenrate und Delay als auch Attribute wie Sprachqualität zusammen. Im Folgenden soll allerdings speziell die Sprachqualität im Netzwerk bedingt durch die gewählte Netzwerktopologie untersucht werden. Hierfür werden zunächst zwei unterschiedliche Topologien vorgestellt und deren Eigenschaften skizziert und verglichen. Die sogenannte Ringtopologie wird anschließend bezüglich besonderer Qualitätsaspekte in einer Machbarkeitsstudie untersucht. Hierfür wird zunächst die Datenbasis beschrieben bevor die Ergebnisse vorgestellt und interpretiert werden.

Topologien

Im Folgenden werden zwei Topologien vorgestellt, welche Gruppengespräche ermöglichen sollen. Das heißt, dass jeder von N Teilnehmern im Netzwerk die anderen $N-1$ Teilnehmer hören soll, sprich dass eine Art Telefonkonferenz durchgeführt wird. Da hier kein Server verfügbar ist, der das Zusammenfügen der Signale vornehmen kann, muss die Topologie ermöglichen, dass das Zusammenfügen auf Nutzerseite geschehen kann.

Die Triviallösung ist in Abb. 1 für $N = 5$ aufgezeichnet. Hier sendet jeder Teilnehmer sein Signal separat an jeden anderen Teilnehmer. Eine klassische „Broadcast“-Übertragung wird hier nicht betrachtet, da WLAN-basierte Technologien dies in der Regel nicht zulassen. Bei einer Basisdatenrate B der Sprachsignale ergibt sich ohne Kanalkodierung eine Netzwerkdatenrate von $B{\cdot}(N^2{-}N)$, sodass die Datenrate quadratisch von der Nutzerzahl abhängt. Jeder Teilnehmer muss die $N-1$ empfangenen Signale dekodieren und aufsummieren. Die Rechenlast jedes Teilnehmers ist somit ebenfalls abhängig von der Teilnehmerzahl.

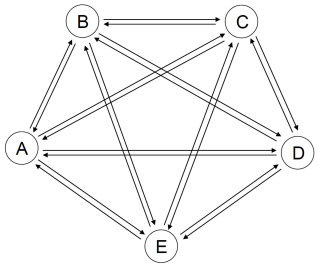

Abbildung 1: Topologie für ein Gruppengespräch, bei der jeder Teilnehmer sein Signal separat an den anderen Teilnehmer sendet.

Alternativ lässt sich die Ringtopologie in Abb. 2 definieren. Hier wird nicht jedes Sprachsignal separat übertragen, sondern es werden Summensignale gebildet, die dann übertragen werden. Abgesehen von den Ringenden hat jeder Teilnehmer zwei Nachbarn, von denen er jeweils ein Signal erhält. Dieses Signal wird dekodiert und

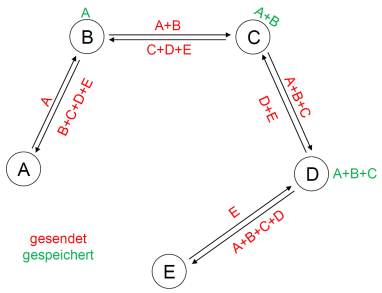

Abbildung 2: Topologie für ein Gruppengespräch, bei der Summensignale im Ring propagiert werden. Der Inhalt gespeicherter (grün) und gesendeter (rot) Signale ist jeweils angemerkt.

das eigene Signal addiert. Nach erneuter Kodierung wird das Signal an den anderen Nachbarn weitergeleitet. Wird so der Ring wie in Abb. 2 einmal im und einmal gegen den Uhrzeigersinn durchlaufen, so verfügt jeder Teilnehmer über alle $N-1$ Signale und kann diese abspielen. Mit diesem Ansatz lässt sich die Datenrate auf $2B \cdot (N-1)$ zu einer linearen Teilnehmerabhängigkeit reduzieren. Anstatt $(N-1)$-mal dekodieren zu müssen, muss so jeder Teilnehmer genau zweimal kodieren und dekodieren. Da die Kommunikation nur mit den jeweiligen Nachbarn stattfindet, kann die Topologie als verkettet Liste interpretiert werden, sodass die gesamte Topologie nicht jedem Teilnehmer bekannt sein muss und gleichzeitig Änderungen zur Laufzeit gut vorgenommen werden können. Dies steht im Gegensatz zu der Topologie in Abb. 1, wo die komplette Topologie jedem Teilnehmer bekannt sein und somit jede Topologieänderung dem Netzwerk bekanntgemacht werden muss. Somit bietet die Ringtopologie einige konkrete Vorteile.

Wird angenommen, dass in beide Topologien die Datenübertragung ohne Fehler erfolgt, so ist allerdings im Netzwerk mit Ringtopologie eine geringere Sprachqualität zu erwarten. Dies ist durch die Bildung und Übertragung der Summensignale zu begründen, da hier einzelne Sprachsignale potenziell mehrfach kodiert werden. Tabelle 1 zeigt wie oft das Signal eines Teilnehmers kodiert wurde bevor es bei einem bestimmten Teilnehmer abgespielt wird. Es ist ersichtlich, dass so im „worst-case“ die Sprache eines Teilnehmers $(N-1)$-mal kodiert wird. So könnte sich ein Qualitätsverlust sowohl anhand eines einzelnen Sprechers, aber auch durch die Qualitätsvariationen zwischen den Sprechern manifestieren. Um die Ausprägung des potenziellen Qualitätsverlustes, und die Performanz der Ringtopologie insgesamt, einschätzen zu können, wird dementsprechend im Folgenden eine Machbarkeitsstudie mit einem Hörtest entworfen und vorgestellt.

Datenbank

Der im Folgenden beschriebene Hörtest wurde durchgeführt, um zwei Qualitätsfragen zu beantworten: Zum einen wie stark der Qualitätsverlust bei einer Mehrfachkodierung ist und zum anderen wie Hörer ein Gespräch bewerten, in dem die Sprache der Sprecher unterschied-

		Signal				
		A	B	C	D	E
Ort	A	x	1	2	3	4
	B	1	x	1	2	3
	C	2	1	x	1	2
	D	3	2	1	x	1
	E	4	3	2	1	x

Tabelle 1: Anzahl der Kodierungen eines Sprachsignals an den verschiedenen Abspielorten. Da ein Teilnehmer sein eigenes Signal nicht abspielt, sind diese Kombinationen mit einem x gekennzeichnet.

lich oft kodiert wurde. Bezüglich der Mehrfachkodierung sind zwar Hörtestergebnisse in den Evaluierungsdokumenten [2] mancher Codecs verfügbar, allerdings wird dort jeweils nur das simple Codec-Tandem, sprich die zweifache Kodierung, betrachtet.

Der Hörtest wurde mit breitbandiger Sprache durchgeführt. Referenzsignale waren die vier deutschen Doppelsätze aus der ITU-T Rec. P.501 [3], welche mit [4] entsprechend bandbegrenzt und auf 16 kHz unterabgetastet wurden. Bei der Auswahl der Codecs wurde der Bedarf für eine moderate Basisbitrate berücksichtigt, sodass letztlich der flexible und bewährte AMR-WB Codec [5] bei seiner Ankerbitrate (12,65 kbit/s), sowie bei 6,60 und 15,85 kbit/s, ausgewählt wurde. Als zusätzlicher Vergleich zu diesen drei Codecs wurde der Opus Codec [6], der hauptsächlich für VoIP verwendet wird, hinzugefügt. Hier wurde die Bitrate zur Vergleichbarkeit auf 16 kbit/s festgelegt, welche somit im empfohlenen Bereich für breitbandige Sprache liegt.

Zur Gewinnung der gestörten Signale wurden die Doppelsätze bei 16 kHz für jeden Codec jeweils ein- bis viermal hintereinander kodiert, sodass hier 64 Signale vorliegen. Die zu bewertenden Gespräche wurden in Anlehnung an Abb. 1 und 2 generiert, wo der Hörer passiv die Position einer der fünf Teilnehmer einnimmt, sodass er vier verschiedenen Sprechern zuhört. Hierfür wurde zur Realisierung von Abb. 1 jeder verfügbare Sprecher einfach kodiert und die vier Signale anschließend ohne Überlappung aneinander gereiht, wobei zunächst alle möglichen Permutationen generiert wurden. Dieselben Permutationen der Sprecheranordnung wurden auch zur Umsetzung von Abb. 2 generiert, wobei hier dann jeder Sprecher unterschiedlich oft (ein- bis viermal) kodiert wurde, sodass der „worst-case“ aus Tabelle 1 umgesetzt wurde. Da die Anzahl der Kodierungen eines Sprechers unabhängig vom Zeitpunkt seines „Turns“ im Gespräch sein soll, wurden bezüglich der Kodierungsanzahl ebenfalls alle möglichen Permutationen generiert. Diese Generierung der Gespräche wurde für jeden Codec durchgeführt, allerdings ohne Codecs im Gespräch zu mischen.

Qualitätsanalyse

Unter Verwendung der generierten Signale wurde ein *Comparative-Mean-Opinion-Score*-Test (CMOS) durchgeführt. Zusätzlich zum in [7] beschriebenen Verfahren

wurde bei jedem präsentierten Signal die Frage (binär) nach Verständlichkeitsproblemen gestellt. Dieses Verfahren ersetzt natürlich keinen Verständlichkeitstest, allerdings wird so bereits eine erste Tendenz zur Verfügung gestellt. Auftretende Verständlichkeitsprobleme würden die spätere Umsetzung der entsprechenden Kondition untersagen.

Der Hörtest wurde mit 19 weitestgehend naiven Teilnehmern (5 weiblich, 14 männlich) im Alter von 24 bis 48 Jahren (Mittelwert 27) durchgeführt, deren Audiogramm nicht auf einen Hörverlust schließen lässt. Bewertet wurden 20 Konditionen. Zur separierten Bewertung des Qualitätsverlustes bei Mehrfachkodierung wurden für jeden Codec jeweils Doppelsätze zusammen präsentiert, wo die Anzahl der Kodierungen sich um eins oder drei unterscheidet, sodass pro Codec vier Konditionen (insgesamt 16) vorliegen. Jedem Hörer wurden diese Konditionen mit einem weiblichen und einem männlichen Sprecher präsentiert, sodass pro Kondition 38 Wertungen verfügbar sind. Die restlichen vier Konditionen stellen für jeden Codec den Vergleich zwischen den Gesprächen dar, wobei jeweils für eine zufällig ausgewählte Sprecherpermutation das Signal mit einfacher und gemischter Kodierungsanzahl präsentiert wurde. Die Kodierungspermutation wurde ebenfalls zufällig gewählt. Pro Codec, und somit Kondition, wurden jeweils zwei Permutationen präsentiert, sodass auch hier 38 Wertungen pro Kondition vorliegen.

Die Ergebnisse (Abb. 3) zeigen zunächst wie erwartet, dass durch jede weitere Kodierung eines Signals ein Qualitätsverlust zu erwarten ist, wobei allerdings an keiner Stelle ein Verständlichkeitsproblem angemerkt wurde. Die Ausprägung des Qualitätsverlustes ist dabei klar codecabhängig. Insbesondere wenn Signale mit einfacher und vierfacher Kodierung zusammen präsentiert werden, wird diese Abhängigkeit deutlich. So ist die Verschlechterung für AMR-WB bei 6,6 kbit/s als extrem einzustufen, wobei sich die Verschlechterung für AMR-WB 15,85 kbit/s in Grenzen hält. Hieraus kann geschlossen werden, dass bei einer sorgfältigen Auswahl des Codecs keine große Qualitätsvariation bemerkt wird, wenn einfach und mehrfach kodierte Signale zusammen präsentiert werden. Diese Beobachtung wird durch die Bewertung der Gespräche untermauert.

Zusammenfassung

In den vorigen Abschnitten wurde die Sprachübertragung in Ad-Hoc-Netzwerken thematisiert, wobei zwei grundsätzliche Netzwerktopologien vorgestellt wurden. Die sogenannte Ringtopologie bietet vielversprechende Vorteile, allerdings unter einer potenziellen Sprachqualitätseinbuße. Um diesen möglichen Qualitätsverlust zu bewerten wurde im Rahmen einer Machbarkeitsstudie ein Hörtest durchgeführt. Hier konnte gezeigt werden, dass sich bei einer sorgfältigen Auswahl des Sprachcodecs der Qualitätsverlust in Grenzen hält und somit der Einsatz der Ringtopologie, motiviert durch seine beschriebenen Vorteile, in Betracht gezogen werden kann. Für eine Umsetzung könnte die Zahl der Kodierungen durch

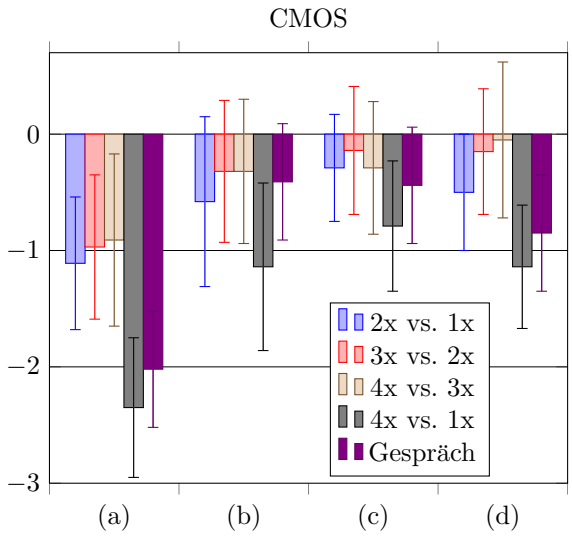

Abbildung 3: Mittelwerte und Standardabweichungen der Ergebnisse des Hörtests für (a) AMR-WB 6,6 kbit/s, (b) AMR-WB 12,65 kbit/s, (c) AMR-WB 15,85 kbit/s, (d) Opus 16 kbit/s.

den Einsatz einer Sprachaktivitätserkennung reduziert werden. Besonders wichtig ist aber zunächst auch die Entwicklung eines robusten Netzwerkprotokolls.

Danksagung

Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft (DFG) unter dem Zeichen SCHM 1507/10-1 finanziert.

Literatur

[1] Echoring Homepage, URL: `https://echoring.com/de/`, Letzter Zugriff: 14.03.2019

[2] 3GPP TR 26.976, Performance characterization of the Adaptive Multi-Rate Wide-Band (AMR-WB) speech codec, 2002

[3] ITU-T Rec. P.501, Test signals for use in telephonometry, 2017

[4] ITU-T Rec. G.191, Software tools for speech and audio coding standardization, 2009

[5] ITU-T Rec. G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), 2003

[6] Opus Interactive Audio Codec, URL: `opus-codec.org`

[7] ITU-T Rec. P.800, Methods for subjective determination of transmission quality, 1996