Hör- und Konversationstests zur Untersuchung der Vorteile räumlicher Audiokonferenzen

Alexander Raake, Claudia Schlegel, Matthias Geier,Jens Ahrens
Deutsche Telekom Laboratories, TU Berlin, Email: alexander.raake@telekom.de

Einleitung

Breitbandige statt schmalbandige Telefonsprache und eine räumliche, z.B. binaurale statt monaurale Wiedergabe der Teilnehmer einer Audiokonferenz führen zu einer verbesserten Quellentrennung, Sprachverständlichkeit und Zuordnung des Gesagten zu den einzelnen Gesprächspartnern (s. [1]; [2]). Um die Gültigkeit der meist in Hörversuchen gemessenen Vorteile auch für den Fall einer tatsächlichen Konversation zwischen mehreren Gesprächspartnern zu überprüfen, haben wir realistische Konversationsszenarien für strukturierte, aber weitgehend freie Konversationen mit drei Gesprächspartnern entwickelt [3]. Der vorliegende Beitrag fasst die Ergebnisse eines entsprechenden Konversationstests zusammen und beschreibt einen Hörtest, welcher Teilaspekte des Konversationstests im Detail beleuchtet. Im Falle der verwendeten Binauralsynthese war in den Konversationsversuchen kein Head-Tracking zur Nachführung der binauralen Raumimpulsantworten eingesetzt worden. Im Hörtest wurden daher als Einstelloptionen die verwendete Audiobandbreite (NB: Schmalband, 300-3400 Hz; WB: Breitband, 50-7000 Hz; FB: Vollband, 50-16000 Hz) und die Wiedergabeart (diotisch; räumlich mit Headtracking; räumlich ohne Headtracking) variiert.

Konversationstest

In einer vorherigen Arbeit wurden Konversationsszenarien für drei Teilnehmer erstellt [3], angelehnt an die von [4] entwickelten Konversationsszenarien für zwei Konversationspartner. Den Gesprächspartnern liegen komplementäre Informationen vor, die die groben Gesprächsinhalte festlegen. Für trotzdem möglichst freie Gespräche sind die Szenarien nicht ausformuliert sondern enthalten die Informationen in tabellarischer Form. Zwei Sätze von je 12 Szenarien wurden entwickelt: Satz 1 mit Geschäfts-Themen wie die Absprache eines Besprechungstermins und Satz 2 mit privaten Themen wie die Organisation einer Geburtstagsfeier. Wesentliche Ziele bei der Entwicklung waren u.a. Natürlichkeit und ausgeglichene Gesprächsdauer zwischen Szenarien.

In einem Konversationstest wurden die Geschäfts-Szenarien mit acht Gruppen von drei Gesprächspartnern evaluiert. Die Evaluierung wurde kombiniert mit einem konkreten Messziel, dem Vergleich der Gesprächsqualität bei diotischer und bei räumlicher Darbietung. Die unterschiedlichen Versuchsbedingungen sind in Tabelle 1 dargestellt. Die $3 \cdot 8 = 24$ Versuchspersonen waren Mitarbeiter der Deutschen Telekom Laboratories und allesamt regelmäßige Benutzer von Telekonferenzsystemen (12 weiblich, 12 männlich, Durchschnittsalter 34,4 Jahre).

#	Bandbreite	*TELR* [dB]	*T* [ms]	Wiedergabe
1	NB	65	0	diotisch
2	NB	65	0	räumlich
3	WB	65	0	diotisch
4	WB	65	0	räumlich
5	FB	65	0	diotisch
6	FB	65	0	räumlich
7	NB	35	100	diotisch
8	NB	35	100	räumlich
9	FB	35	100	diotisch
10	FB	35	100	räumlich

Tabelle 1: Versuchsbedingungen des Konversationstests. *TELR* ≡ Talker Echo Loudness Rating, d.h. Echo-Abschwächung. *T* ≡ mittlere Einweg-Verzögerung. Für die Abkürzungen der Bandbreiten s. Abschnitt .

Zur räumlichen Wiedergabe per Kopfhörer wurden feste Paare von in einem akustisch trockenen Studioraum aufgezeichneten binauralen Raumimpulsantworten verwendet, mit Darbietung der beiden Gesprächspartner bei $\pm 30°$. Für die Wiedergabe kamen offene Headsets des Typs Sennheiser HMD 410-6 zum Einsatz. Szenarien und Versuchsbedingungen wurden mittels eines Griechisch-Lateinischen Quadrates so kombiniert, dass Paare aus Szenario und Versuchsbedingung nicht mehrfach auftraten. Die Versuchspersonen wurden gebeten nach jedem Gespräch die Gesamtqualität der Audiokonferenz auf einer kontinuierlichen sieben-Punkte Skala zu bewerten. Für Details siehe [3].

Die Gespräche dauerten zwischen 5:50 und 7:20 min, mit einem Mittelwert von 6:25 min. Eine zweifaktorielle Varianzanalyse mit *Szenario* und *Gruppe* als feste Faktoren und Dauer als abhängige Variable zeigt einen größeren Einfluss durch *Gruppe* (*Szenario*: $F = 2.6,\ p < 0.05$; *Gruppe*: $F = 5.1,\ p < 0.005$). Eine Varianzanalyse der Versuchspersonenurteile mit einem gemischten linearen Modell mit Messwiederholung [5] ergab einen hochsignifikanten Qualitätseinfluss durch die Verbindung ($F = 14.3, p < 0.001$). Eine gleichartige aber zweifaktorielle Varianzanalyse der echofreien Verbindungen mit den Faktoren *Bandbreite* und *Wiedergabe* ergibt einen signifikanten und etwa gleich großen Einfluss der beiden Faktoren (*Bandbreite*: $F = 7.5, p < 0.005$; *Wiedergabe*: $F = 5.6, p < 0.05$).

Hörtest

Die Testgespräche wurden mittels der Geschäfts-Szenarien von drei männlichen geübten Sprechern in

getrennten Räumen eingesprochen (wesentliche Kriterien: Ähnliche Stimmcharakteristika, Vergleichbare Gesprächsdauern, keine Atemgeräusche, Ernsthaftigkeit). Der Hörtest unterschied sich in wesentlichen Punkten vom Konversationstest: (1) Testmethode — Hör- statt Konversationstest; (2) Anzahl der Sprecher für eine Versuchsperson — 2 im Konversationstest, 3 im Hörtest; (3) Fragestellungen am Ende des Gesprächs — Qualität im Konversationstest, Qualität (MOS), Erkennbarkeit der Gesprächspartner (REC), wahrgenommene Verständlichkeit (INT), notwendige Aufmerksamkeit um Gesprächen zu folgen (ATT), Nutzen der räumlichen Aufteilung (USP) im Hörtest. Zudem wurde auf Echoverbindungen verzichtet, um die Aufmerksamkeit nicht von den interessierenden Merkmalen durch Bandbreite und Wiedergabe zu lenken. Drei der insgesamt 9 Verbindungen wurden mit dynamischer Binauralsynthese realisiert (siehe Tabelle 2; Binauralsynthese mittels [6], mit Polhemus FASTRAK zum Headtracking).

#	Bandbreite	Wiedergabe	Headtracking
1	NB	diotisch	-
2	WB	diotisch	-
3	FB	diotisch	-
4	NB	räumlich	-
5	WB	räumlich	-
6	FB	räumlich	-
7	NB	räumlich	ja
8	WB	räumlich	ja
9	FB	räumlich	ja

Tabelle 2: Hörversuchsbedingungen.

Aufgrund der in [1] berichteten Vorteile einer räumlichen Darbietung für die Merkfähigkeit der Teilnehmer wurden zwischen der Bewertung der Gesamtqualität MOS und den weiteren Qualitätsaspekten (REC, INT, ATT, USP) ein zweiteiliger Erinnerungstest durchgeführt. Im ungestützten Teil sollten die Probanden nach der Qualitätsbeurteilung so viele Aussagen jedes Sprechers aufschreiben wie möglich. Im gestützten Teil wurden $3 \cdot 8 = 24$ zuvor annotierte Äußerungen vorgegeben die dem jeweiligen Sprecher zugeordnet werden sollten. Ausgewertete Maße sind hier die mittlere Anzahl der im ungestützten Fall korrekt erinnerten Aussagen $FREm$, sowie die mittleren Anzahlen der im gestützten Fall richtig, falsch und nicht zugeordneten Aussagen ($CORm$, $FALm$, $NASm$).

Die 24 Versuchspersonen die am Hörtest teilnahmen wurden diesmal aus der Studentenschaft der TU Berlin rekrutiert, da eine erneute Teilnahme erfahrener Konferenznutzer als unzumutbar erachtet wurde (13 weibl., 11 männl., Durchschnittsalter 26,6 Jahre).

Die Auflösung der beiden Erinnerungstests ist sehr gering, jedoch konnte in einer Varianzanalyse mit gemischtem linearen Modell und Messwiederholung z.B. für $FREm$ ein signifikanter Einfluss der Versuchsbedingung nachgewiesen werden ($F = 2.7, p < 0.05$). Von den Beurteilungen lassen die Erkennbarkeit REC und der Nutzen der räumlichen Aufteilung USP die beste Differenzierung einzelner Verbindungen zu (MOS: $F = 9.7$, $p < 0.001$; REC: $F = 11.2$, $p < 0.001$; INT: $F = 8.4$, $p < 0.001$; ATT: $F = 7.4$, $p < 0.001$; USP: $F = 23.9$, $p < 0.001$).

Untersucht man den Einfluss der Bandbreite und Wiedergabe auf die Qualitätsurteile (MOS), zeigt sich ein größerer Einfluss der Wiedergabe als der Bandbreite, im Gegensatz zum Konversationsversuch (Bandbreite: $F = 8.468$, $p < 0.001$; Wiedergabe: $F = 30.426$, $p < 0.001$). Bei ausschließlicher Betrachtung der Fälle mit räumlicher Wiedergabe ($\#4 - 9$) konnte kein signifikanter Einfluss durch Headtracking auf die Messgrößen gefunden werden.

Zusammenfassung

Die Konversationsszenarien erlauben es, vergleichbare Konferenzen nachzubilden und so die Konferenzqualität anwendungsnah zu messen. Im Konversationstest mit hoher kognitiver Einbindung der Versuchspersonen lässt sich wie im Hörtest ein Vorteil durch räumliche Schallwiedergabe messen. Wie erwartet steigt der Nutzen der räumlichen Wiedergabe mit der Anzahl der Teilnehmer an und dominiert bei drei Gesprächspartnern bereits den Einfluss durch eine größere Sprachbandbreite. Ein Einfluss durch Headtracking konnte nicht gezeigt werden.

Die grundsätzlichen Anzahlen korrekt erinnerter oder zugeordneter Aussagen fällt wesentlich geringer aus als z.B. beim Hörtest von [1]. Aufgrund der dort vier statt drei verwendeten Sprecher ist anzunehmen, dass dieser Effekt wie auch der Vorteil räumlicher Wiedergabe versus Bandbreite mit zunehmender Anzahl von Sprechern stärker hervortritt. In zukünftigen Arbeiten wird der Einfluss durch die Anzahl von Sprechern konkret untersucht.

Literatur

[1] Baldis, J.: Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: Proc. CHI (2001).

[2] Raake, A., Spors, S., Ahrens, J., Ajmera, J.: Concept and evaluation of a downward-compatible system for spatial teleconferencing using automatic speaker clustering. In: Proc. INTERSPEECH (2007).

[3] Raake, A., Schlegel, C.: Auditory assessment of conversational speech quality of traditional and spatialized teleconferences. In: Proc. 8th ITG Conference Speech Communication (2008).

[4] Möller, S.: Assessment and prediction of speech quality in telecommunications. Kluwer Academic Publishers, USA–Boston (2000).

[5] Quené, H., van den Bergh, H.: On multi-level modeling of data from repeated measures designs: A tutorial. Speech Communication (2004) 43(1-2), 103-121.

[6] Geier, M., Ahrens, J., Spors, S.: The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In: Proc. 124th AES Convention (2008).