ITU-T Workshop

From Speech to Audio:
bandwidth extension, binaural perception

Lannion, France, 10-12 September 2008

Conversational speech quality of spatialized audio conferences

Alexander Raake and Claudia Schlegel

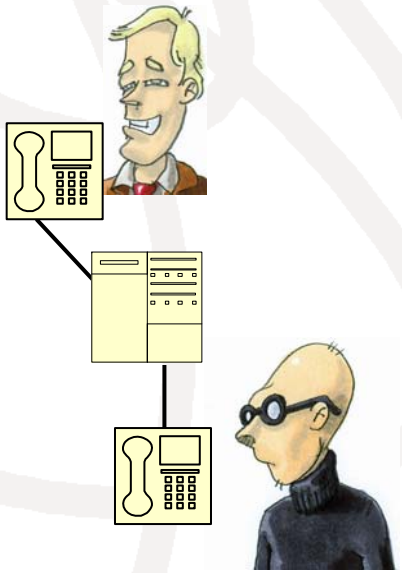
Quality & Usability Lab
Deutsche Telekom Laboratories
Berlin Institute of Technology
alexander.raake@telekom.de

Deutsche Telekom Laboratories

berlin

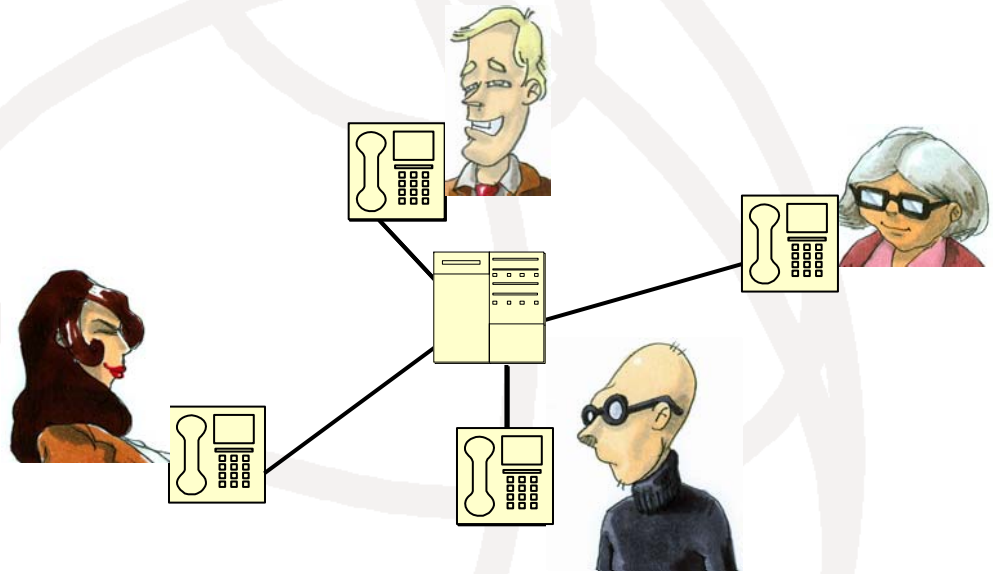
Spatialized audio conference

Telephony



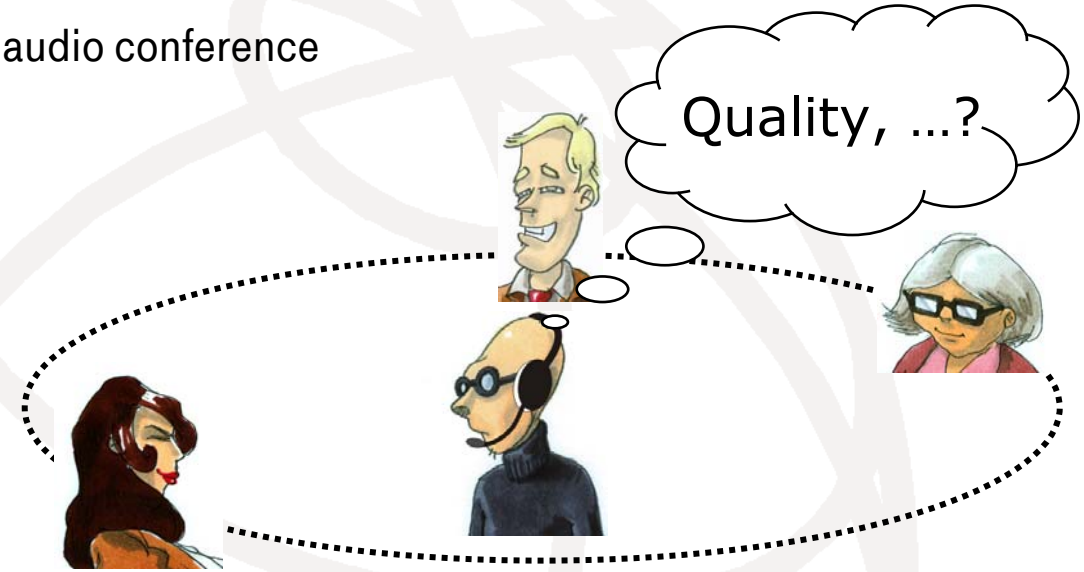
Spatialized audio conference

Classical Teleconference

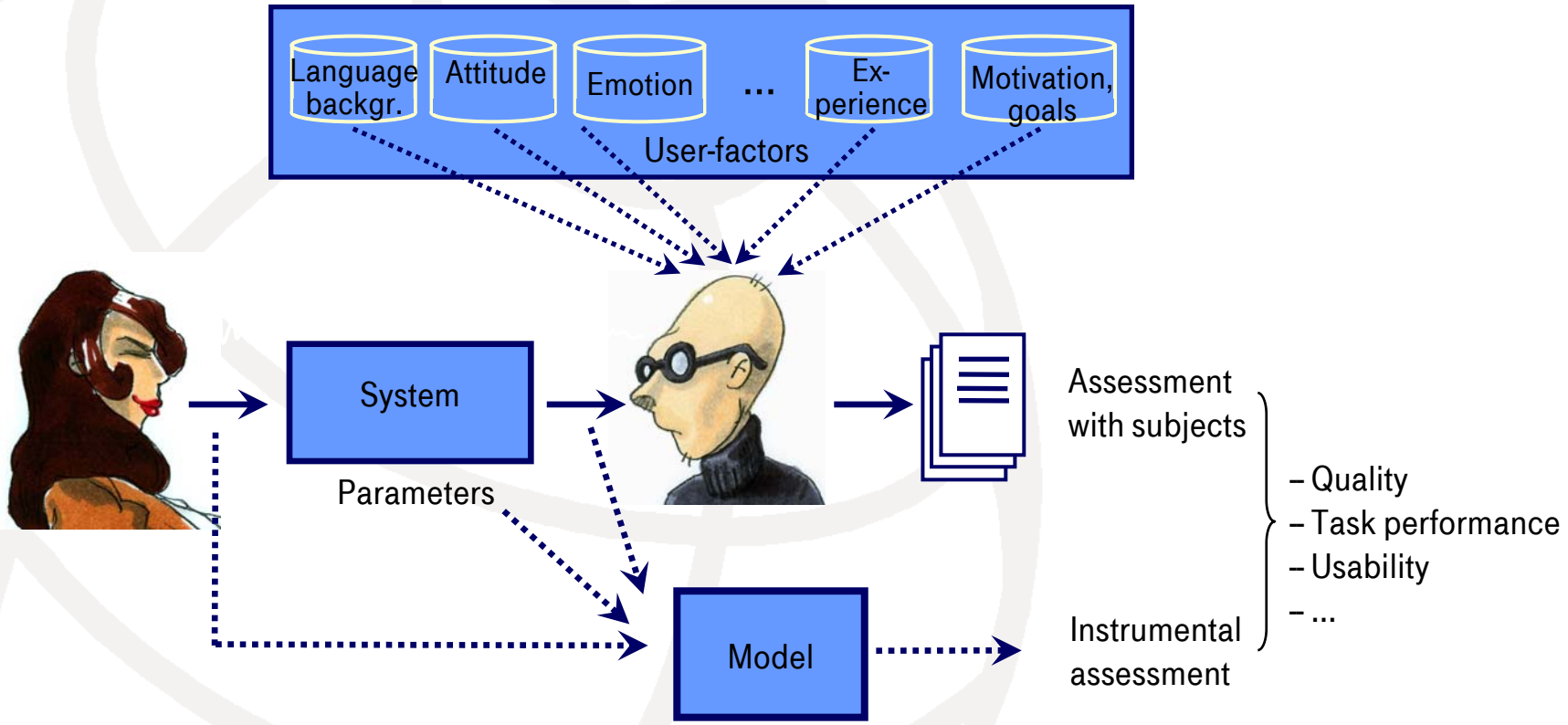


Spatialized audio conference

Spatialized audio conference



Quality assessment



Overview

- Introduction
- **Aspects of 3D conferencing & user perception**
  - Intelligibility
  - Usability & task-performance
  - Quality
- Listening test
- Conversation tests
- Conclusion

Intelligibility

- SRT: Speech reception threshold
  - SNR that yields 50% word intelligibility per sentence
- Comparison of different configurations: ΔSRT

Factor	ΔSRT (improvement) [dB]
Spectral differences	-2 → 2
Fluctuations	6 → 10
Voice similarity	-9 → -3
Spatial separation	**0 → 11**
Reverberation	-9 → 0
Coding	-5 → 0

(Bronkhorst, 2000; Raake & Katz, 2007)

Advantage of spatial separation
→ Cocktail Party Effect (Cherry, 1953)

Usability & performance

Further advantages of spatial audio

- Speaker recognition (e.g. Baldis, 2001).
- Focal assurance
  - Participants can better recall general concepts of other participants (Baldis, 2001).
  - Efficient share of load by two parts of working memory (Logie, 1995; Baddeley, 1987):
    - Visual – spatial (*visual-spatial sketch*).
    - Verbal – semantic (*phonological loop*).

Quality

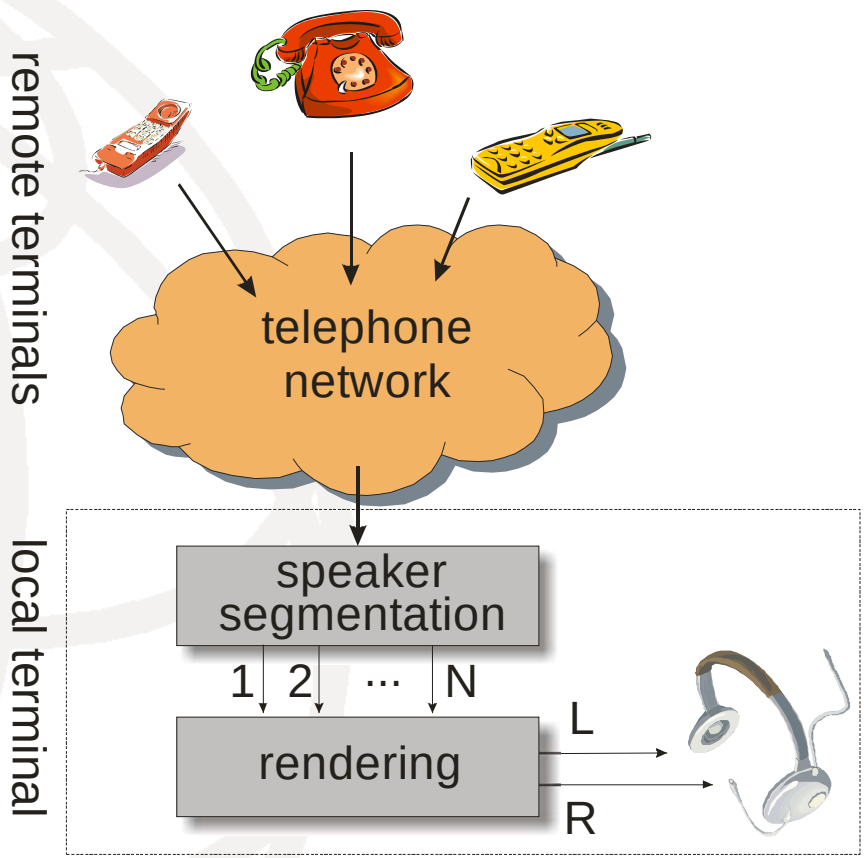
- "Result of judgment of perceived composition with respect to desired composition". (Jekosch, 2000, 2005)
- Quality in listening situation
  - Timbral reproduction more important than spatial feautures (Rumsey et al., 2005; Silzle, 2007).
  - Spatial reproduction typically preferred over non-spatial reproduction (Baldis, 2001).
  - May depend on whether sources keep their location, i.e. headtracked headphone or loudspeaker presentation vs. non-headtracked headphones (Kilgore et al., 2003).

Overview

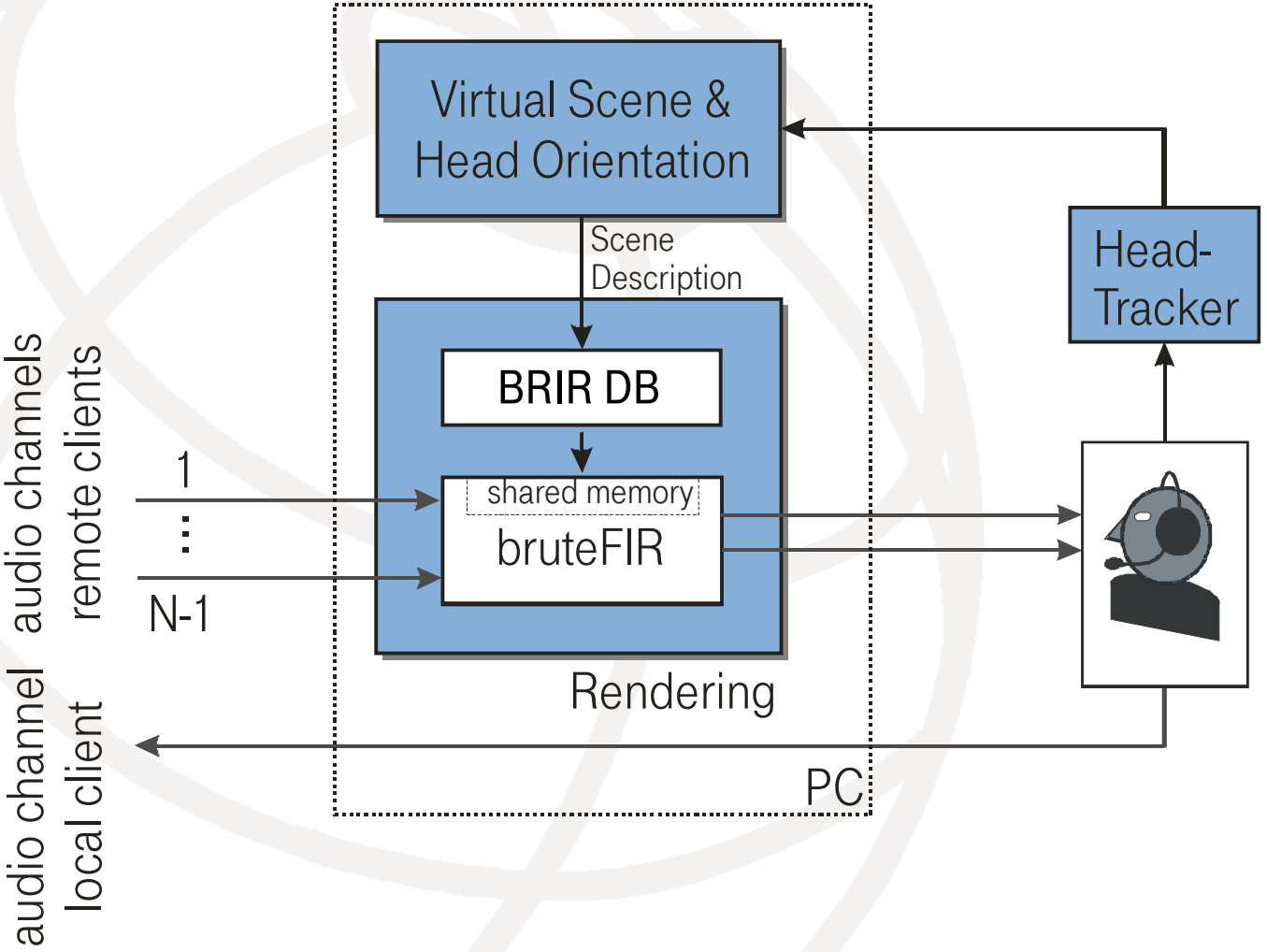
- Introduction
- Aspects of 3D conferencing & user perception
  - Intelligibility
  - Usability & task-performance
  - Quality
- **Listening test**
- Conversation tests
- Conclusion

Listening test
General goal

- Evaluation of downward-compatible spatial teleconferencing based on automatic speaker clustering
  (Raake, Spors, Ahrens, Ajmera, 2007)
- NB speech!



Listening test
Binaural reproduction



Listening test
Test set-up

- German digit utterances concatenated from various speakers (VeriDat database: Turk & Schiel, 2003).
- 5 sequences (1x two speakers, 2x three speakers, 2x four speakers); durations: 40 s - 1 min.
- Fs = 8 kHz (downward-compatibility to NB-telephony).

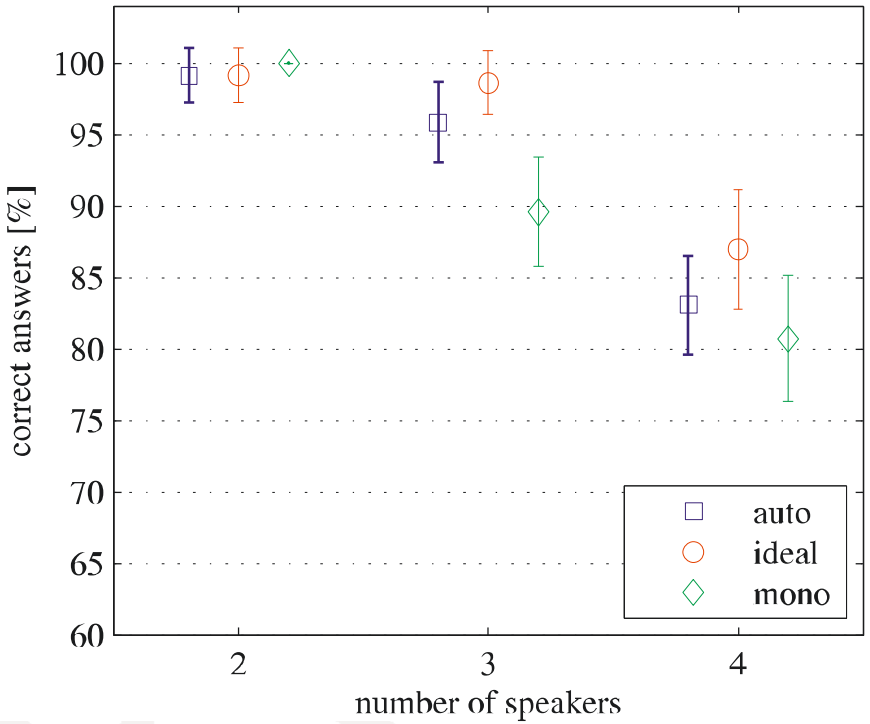
- Three presentation methods
  - Diotic ("mono").
  - Binaural, automatic segmentation ("auto").
  - Binaural, ideal segmentation ("ideal").

- Symmetrical locations, azimuth $\alpha \in \{60°, -60°, 30°, -30°, 0°\}$

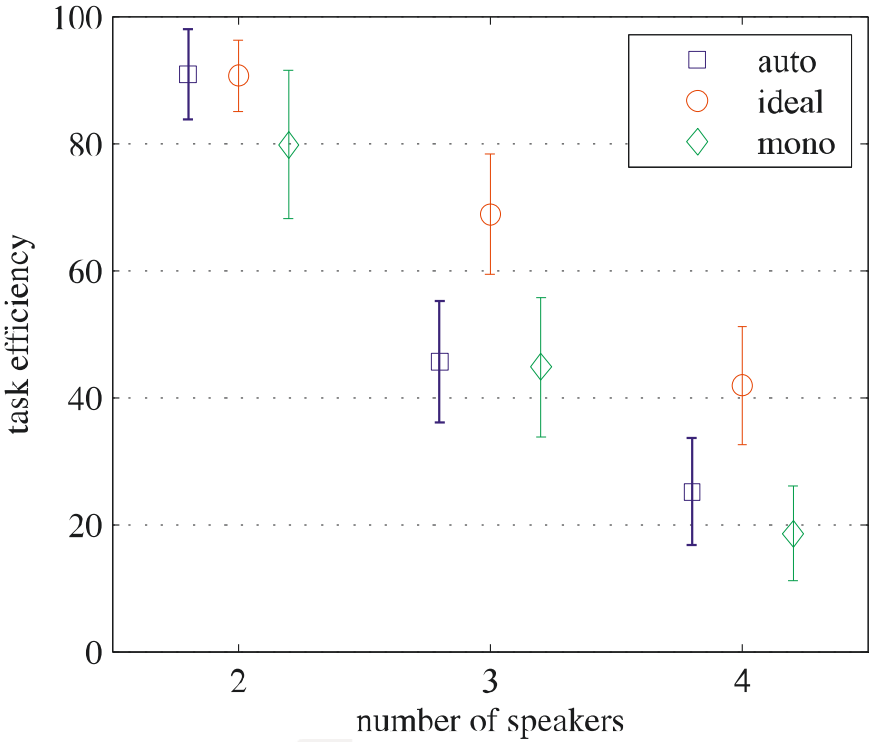
- Tasks (GUI on touch-screen)
  - Report speakers & speaker change points during sequence.
  - Judgments of pleasantness & task efficiency after sequence.

Listening test
Results for task performance

Measured performance

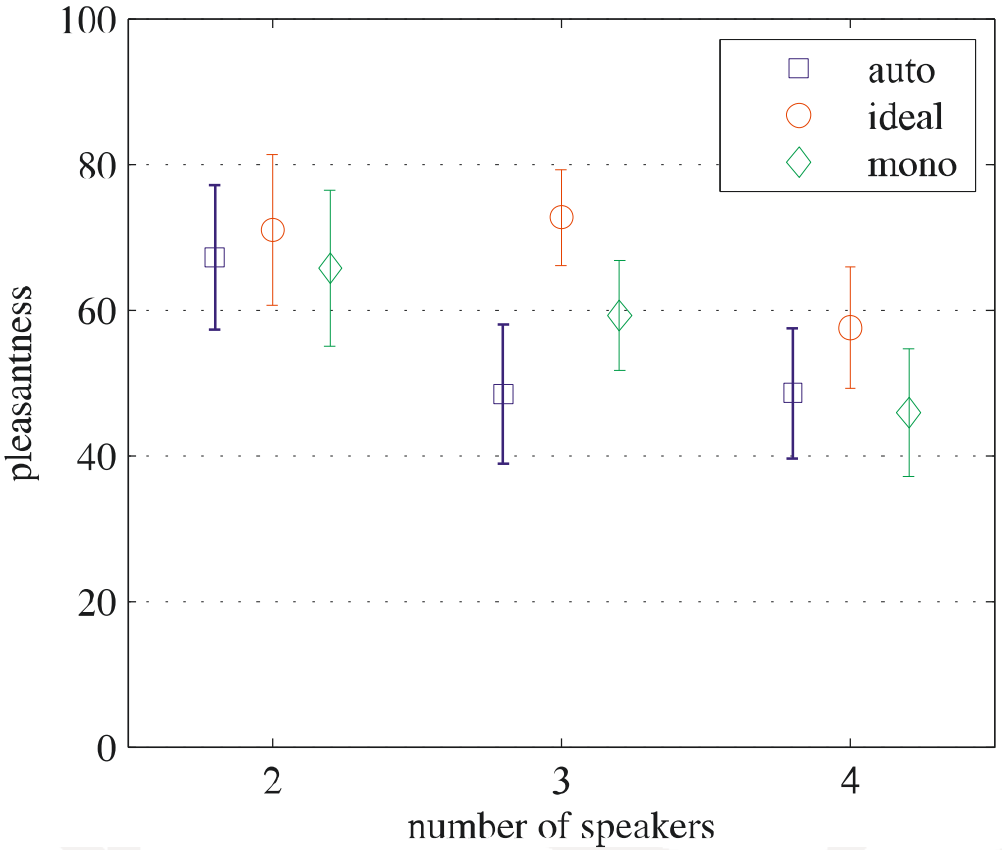


Perceived performance



- 3- & 4-speaker cases: Spatial representation helps considerably to correctly detect speaker changes.
- Real & perceived change detection efficiency
  - 1. Ideal, 2. auto, 3. mono.

Listening test
Results for pleasantness



- ANOVA: "Presentation mode" & "number of speakers" significant factors.
- Ranking: 1. Ideal, 2. mono, 3. auto (misclassifications).
- Significant advantage only for 3 speakers.
- Note: very demanding task!

Overview

- Introduction
- Aspects of 3D conferencing & user perception
  - Intelligibility
  - Usability & task-performance
  - Quality
- Listening test
- **Conversation tests**
- Conclusion

Conversation tests

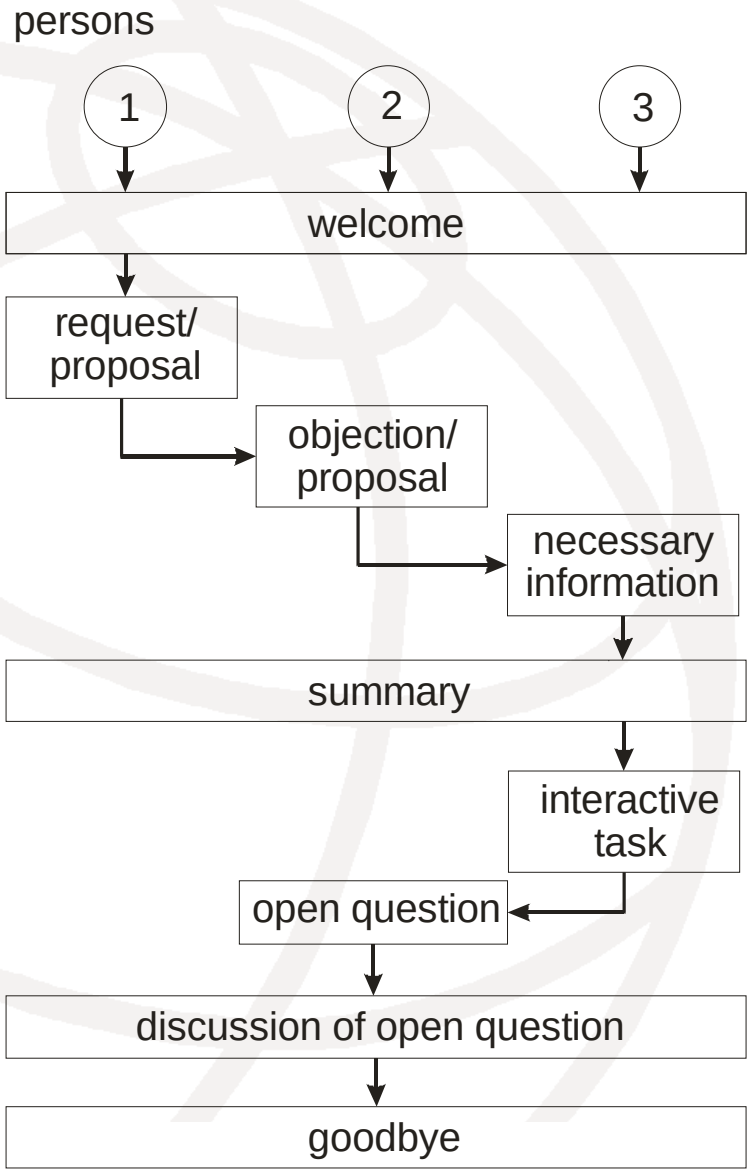
- Main advantage of conversation tests:
  - Reflect actual application of telephony or conferencing in ecologically more valid (more natural) way.
- Main limitations:
  - Time-consuming.
  - Often involve unnatural test scenarios.
  - Lower resolution than listening tests.
- **Aim: Scenarios for conferences, 3 subjects.**

Requirements based on SCTs (Short Conversation Test scenarios)

- Naturalness (topic and environment)
  - Natural conversation tasks.
  - Natural beginning and end.
  - Limited distraction from the quality-perception and -judgment task.
- Balance (conversation flow)
  - No fixed sender- and receiver-roles.
  - Short periods of monologues.
  - Realistic amount of double- or triple-talk.
  - Same repartition of speech activity between participants.
  - Limited overall duration.
- Comparability (between scenarios)
  - Similar instructions, dialogue-structures, durations.

(adopted from Möller, 2000)

"3CT scenarios (3CTs)" Target conversation flow



3CTs development

- Identification of appropriate conferencing topics in email-poll (all Lab collaborators)
  - Business conferences.
  - Spare-time conferences.
- Workshop (experienced conferencing users)
  - Additional topics.
  - Rate topics.
- Scenario formation.
- Informal scenario evaluation (no technical system).
- Scenario refinement.

3CTs

- Each scenario described on 2 sheets.
- 1st sheet identical for all participants
  - Overall situation, topics, roles & names.
- 2nd sheet individual for the 3 participants
  - Information for 3 participants complementary.
  - Necessary to complete conversation task.
- Example topics for business scenarios:
  - Planning of a business meeting.
  - Selection of titles for a new music CD compilation.
  - Organization of an arts exhibition.

3CTs – example

Name: **Meyerhof**

Firma: **Brauerei Starkbier, Unternehmenskommunikation**

Sie wählen die Nummer von Burger aus der Marketingabteilung des Fernsehsenders Fußball Kanal:

Begrüßung

Thema 1:

Sie haben sich bereits geeinigt auf:

- **1-stündiges Meeting nächsten Montag**
- **Treffpunkt: Besprechungsraum beim Fernsehsender Fußball Kanal**
- **Thema: Sponsoring-Angebot besprechen**

Noch offen: Um wie viel Uhr?

Ihr Terminkalender am Montag	
9.00 – 10.00 Uhr	
10.00 – 11.00 Uhr	
11.00 – 12.00 Uhr	
12.00 – 13.00 Uhr	
13.00 – 14.00 Uhr	
14.00 – 15.00 Uhr	
15.00 – 16.00 Uhr	nicht verschiebbares Seminar
16.00 – 17.00 Uhr	

Einigung auf _____ *Uhr*

Thema 2:

Austausch der E-Mail-Adressen:

Name	**Abteilung**	**E-Mail-Adresse**
Meyerhof	Unternehmenskommunikation	h.meyerhof@starkbier.de

Sonstiges:

Diskussion

Verabschiedung

Conversation tests
Scenario evaluation

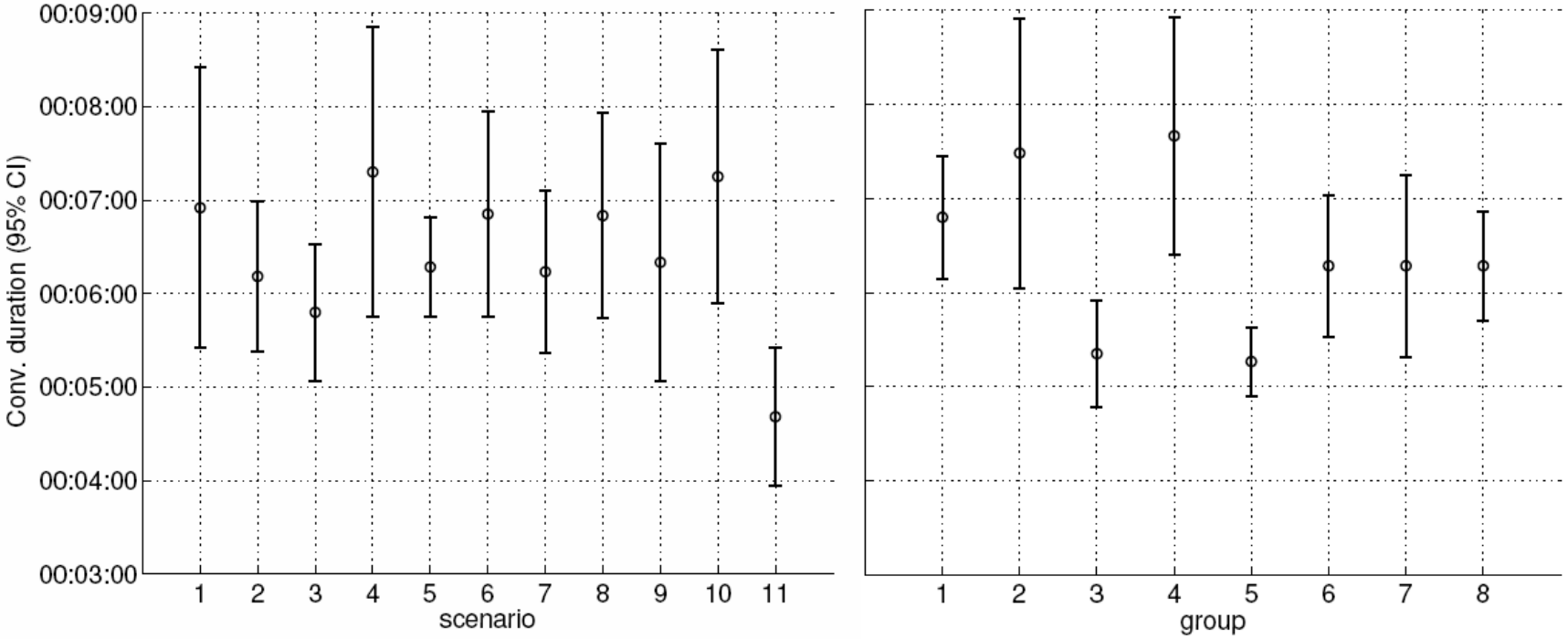
- Goals:
  - Evaluate scenarios.
  - First results on quaity due to spatialized audio.
- 2 test runs.
- 24 subjects per run (8 groups of 3 subjects).
- 1st run
  - Overall quality (Continuous version of the 5-point Absolute Actegory Rating Scale, ACR; yields Mean Opinion Score – MOS; ITU-T Rec. P.800)
  - Conversation effort (CR-10 category-ratio scale; Borg, 1982)
  - Recordings per subject (3 individual tracks): Call duration, turns, etc.

Conversation tests Conditions

#	bandwidth	*TELR* [dB]	*T* [ms]	presentation
1	NB	65	0	diotic
2	NB	65	0	dichotic
3	WB	65	0	diotic
4	WB	65	0	dichotic
5	FB	65	0	diotic
6	FB	65	0	dichotic
7	NB	35	100	diotic
8	NB	35	100	dichotic
9	FB	35	100	diotic
10	FB	35	100	dichotic

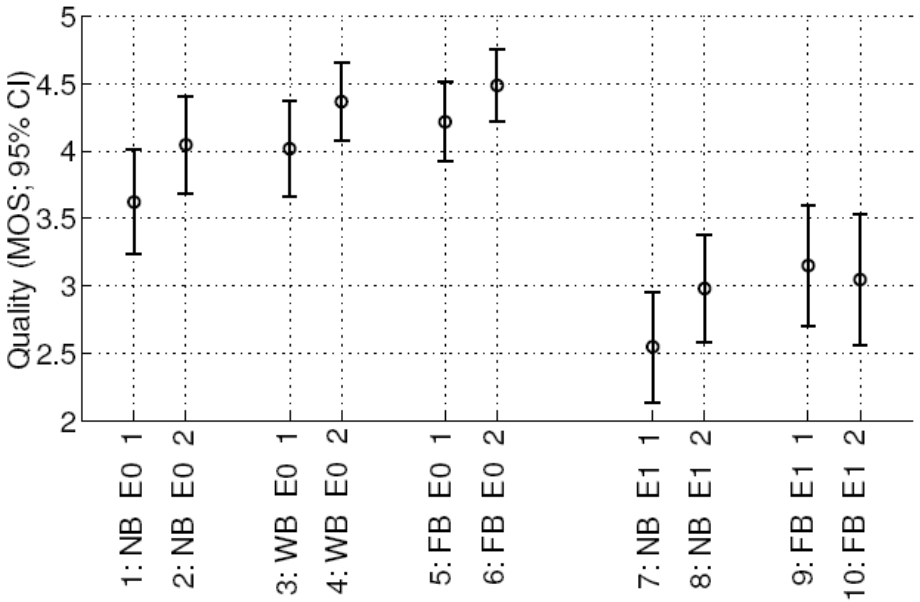
TELR Talker Echo Loudness Rating (echo attenuation)
T Mean one-way delay
NB 300 – 3400 Hz
WB 50 – 7000 Hz
FB 20 – 22000 kHz
Note: System like in listening test, but no head-tracking!

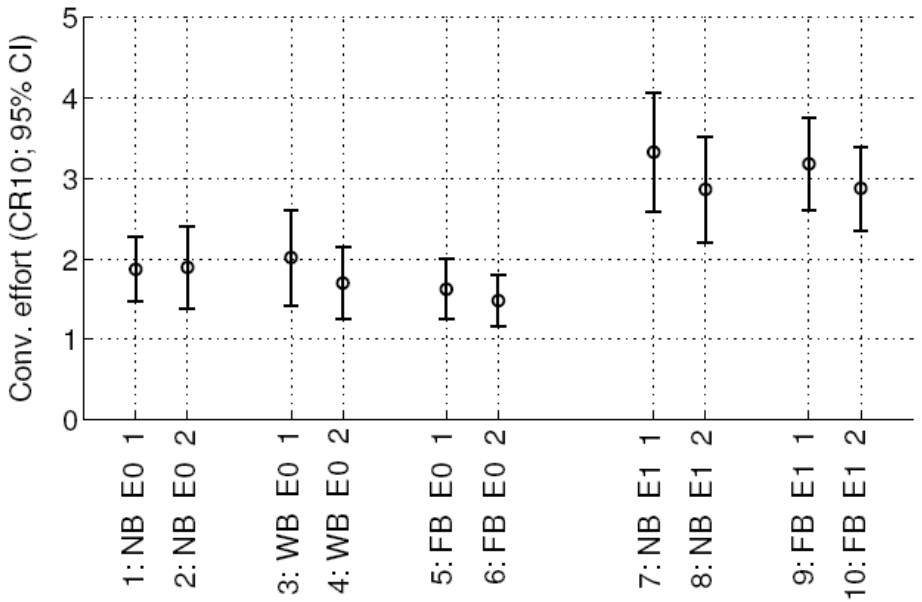
1st conversation test
Call duration



- Average durations between 5:50 to 7:20 minutes, mean 6:25 min.
- Scenario statistically significant factor.
- Subject group: Higher impact.
- No significant impact due to condition (!).
- Similar conversation durations for 10 actual test scenarios.
- Good match with the scenario design goal:
  For SCTs (2-poeple) 2–3 min duration → 3 participants ≈ 3 x 2 min.

1st conversation test
Quality & conversation effort





- Ratings little dependent on diotic vs.dichotic presentation.
- ANOVA:
  - Condition: Highly significant.
  - Scenario: Weak impact.
  - Subject group: No impact on quality, but highly significant impact on conversation effort.

Legend for conditions

"N: XX YY P"

N: condition number

XX: bandwidth

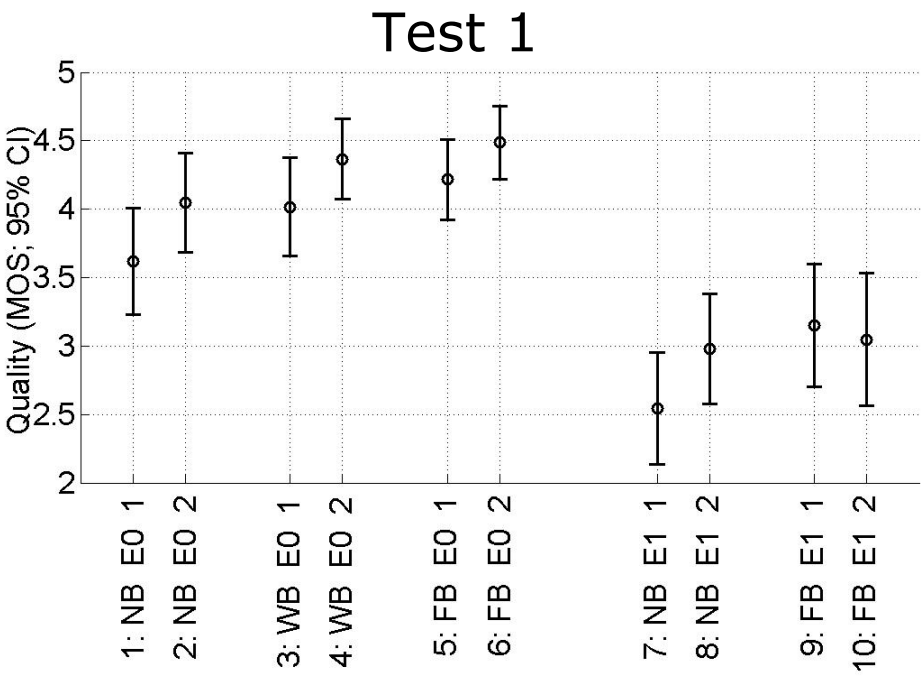
YY: E0 ≡ no talker echo
E1 ≡ talker echo

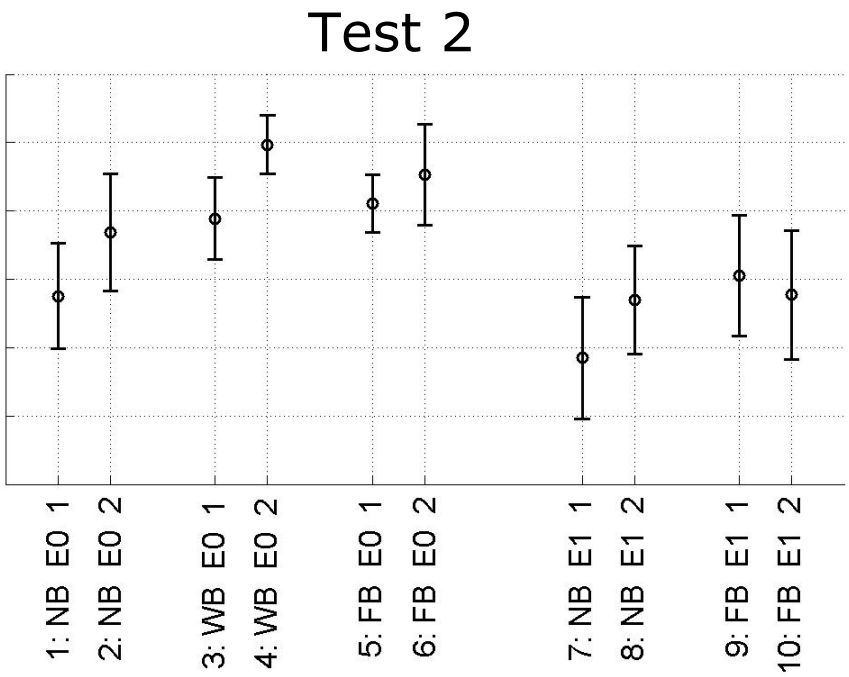
P: 1 ≡ diotic
2 ≡ dichotic (spatial)

2nd conversation test
Set-up

- Differences to 1st run:
  - Simplified scenarios.
  - Paid, external subjects.
  - New instructions highlighting potential spatial presentation.
  - Rating: Overall quality.
  - Additional questions after each scenario & test: Memory, focal assurance.

2nd conversation test
First results





- Differences to 1st run:
  - Quality under echo slightly higher.
  - Again no significant difference between diotic & dichotic for FB.
  - Significant advantage between diotic & dichotic for WB.

Conclusions & Outlook

- Conclusions
  - Human performance increased with spatial audio.
  - Depending on task and presentation, listening quality judged higher than for non-spatial audio.
  - New method for assessing conversational quality.
  - Conversations: Advantage of spatial audio measurable, but subtle.
- Future work
  - Further analysis of recordings (turns, etc.).
  - Analysis of memory test of test 2.
  - Comparison: New listening test with recordings, with headtracking & including memory test.

Thank you!
Questions?