

ON MEASURING THE INTELLIGIBILITY OF SYNTHETIC SPEECH IN NOISE — DO WE NEED A REALISTIC NOISE ENVIRONMENT?

Tuomo Raitio¹, Marko Takanen¹, Olli Santala¹, Antti Suni², Martti Vainio², and Paavo Alku¹

¹Department of Signal Processing and Acoustics, Aalto University, Finland

²Institute of Behavioural Sciences, University of Helsinki, Finland

ICASSP, Kyoto, Japan

March 27 2012

Introduction

- Assessing the intelligibility of synthetic speech is important for creating synthetic voices for real environments
- Real listening environments contain some sort of ambient noise
 - Spatially distributed noise sources
- However, usually synthetic speech is assessed in dichotic listening:
 - Speech and noise are played through headphones
 - No spatial cues for speech/noise
- Is this enough to get a realistic measure of speech intelligibility?
- **This paper addresses the question whether a realistic noise environment should be used to test the intelligibility of synthetic speech**

Speech Intelligibility

Speech intelligibility depends on 3 main factors:

1. Level and type of speech

- Distance, speaker, type of speech (e.g., normal or Lombard speech)



2. Level and type of noise

- SNR, spectral and temporal properties, spatial distribution



3. Acoustic environment

- Room response, reverberation, loudspeakers/headphones

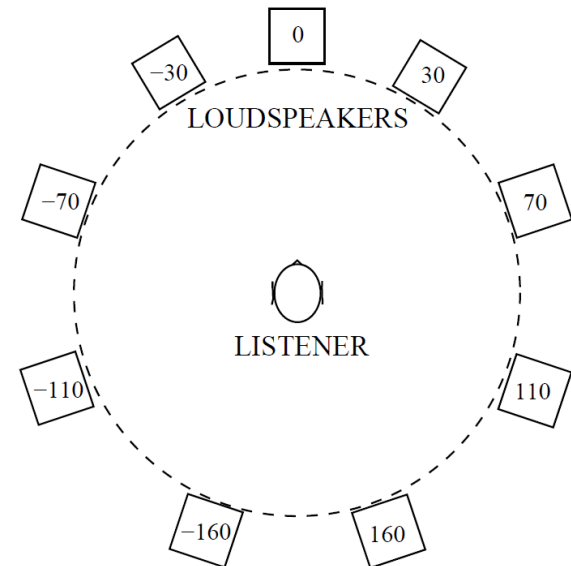


Experiments

- The effect of sound reproduction setup on speech intelligibility and quality/suitability was evaluated
- This was done by conducting the same listening test using three different setups:
 1. **Multichannel loudspeaker setup**
 2. **Stereo headphone setup**
 3. **Mono headphone setup**

1. Multichannel Loudspeaker Setup

- 9 DSP-equalized loudspeakers (Genelec 8260A) at 2.4–2.6m from listener
- Listening room (ITU-R BS.1116-1), average reverberation time 0.3 s
- B-format microphone recordings converted for nine loudspeaker setup with Directional Audio Coding
- Speech reproduced with the center speaker



2. Stereo Headphone Setup

- High-quality headphones (Sennheiser HD580) used for sound reproduction
- B-format microphone noise recordings
- W, X, and Y channels converted to **stereo noise**, sound arriving from -90 and + 90 degrees
- Speech played to both ears



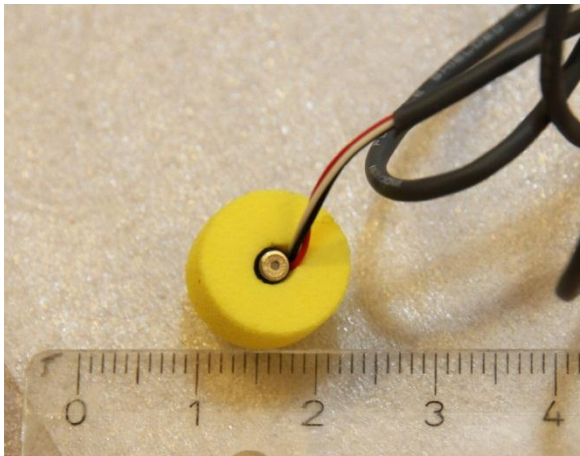
3. Mono Headphone Setup

- High-quality headphones (Sennheiser HD580) used for sound reproduction
- B-format microphone noise recordings
- Omnidirectional channel (W) used for creating **mono noise**
- Speech and noise played to both ears



Loudness Equalization Between Setups

- Sound reproduction levels between loudspeaker and headphone setups were equalized by in-ear microphone measurements with a human subject
- A-weighted sound pressure levels of the signals were normalized at the blocked ear canal entrance



Photos: Javier Gomez Bolanos

Noise Material



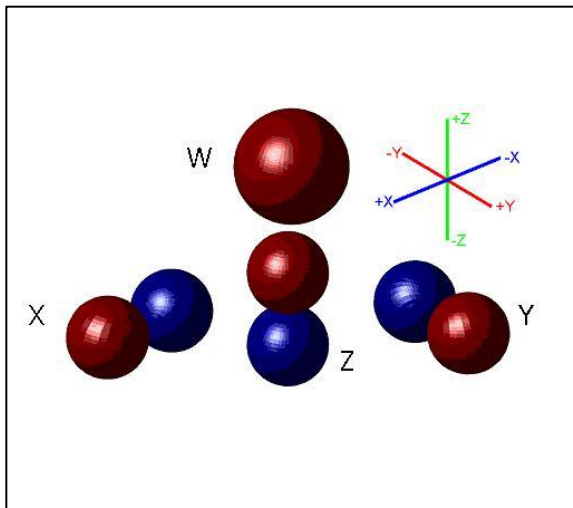
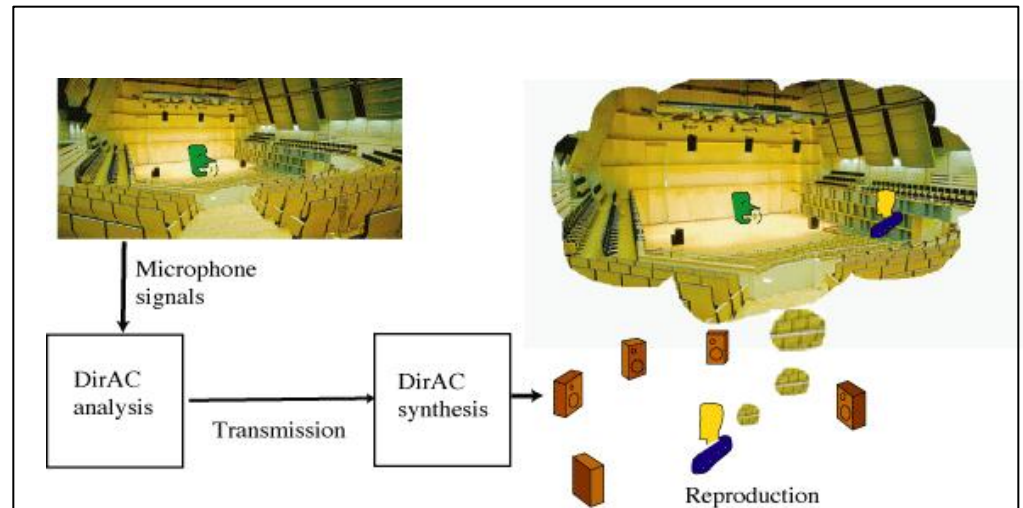
- Two types of noise: **Street noise**  **School noise** 
- **B-format microphone recordings** consisting of W, X, Y, and Z channels
- Converted to 9-channel, stereo, and mono using **Directional Audio Coding**

Illustration of B-format microphone technique

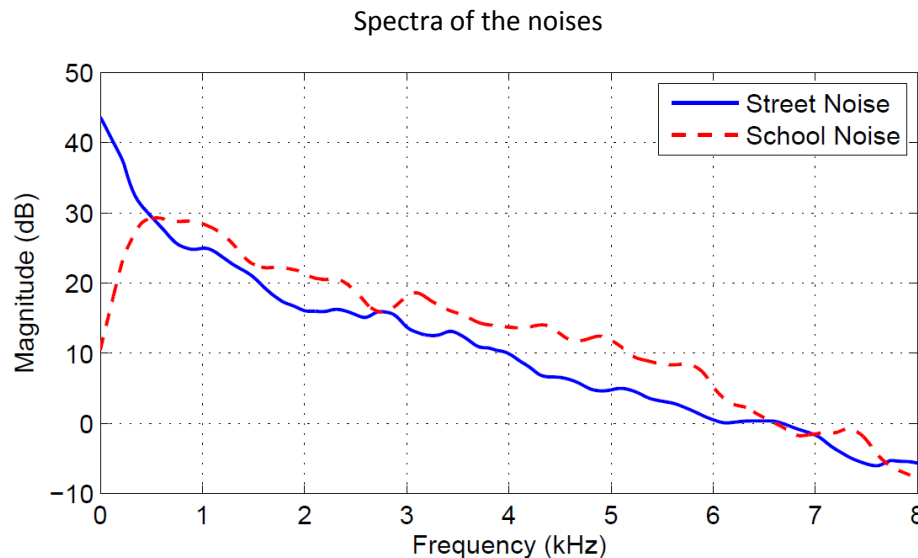


Directional Audio Coding (DirAC)







Noise Material

- Both noises were reproduced with two A-weighted SPLs:
 - **Moderate (63 dB)**
 - **Loud (70 dB)**
- The average SNRs of speech and noise were -1 dB and -8 dB



Speech Material

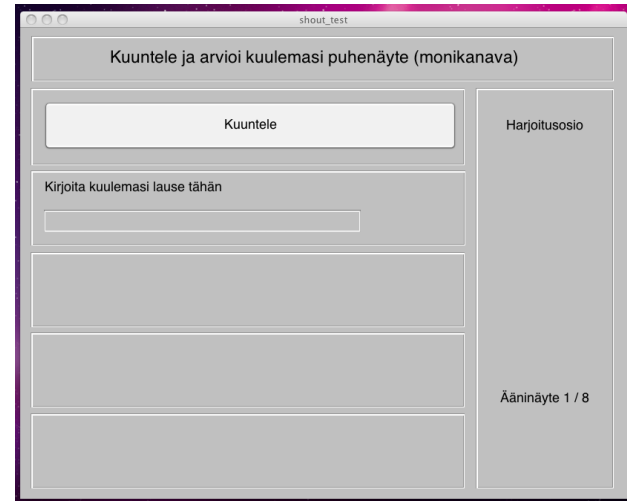
- Four types of speech signals were used
 1. Natural normal speech ($L_A = 59$ dB) 
 2. Synthetic normal speech ($L_A = 61$ dB) 
 3. Natural Lombard speech ($L_A = 63$ dB) 
 4. Synthetic Lombard speech ($L_A = 63$ dB) 
- Finnish male speaker whose normal and Lombard speech was recorded
- Synthetic voices were created using GlottHMM speech synthesis system

**The loudness of all speech samples was normalized by ITU-T P.56
(energy of active speech level)**

Intelligibility Test

- 144 short Finnish sentences in random order
- Sentences presented in the presence of masking noise
- Task of the listener was to type in the heard sentence
- 17 listeners

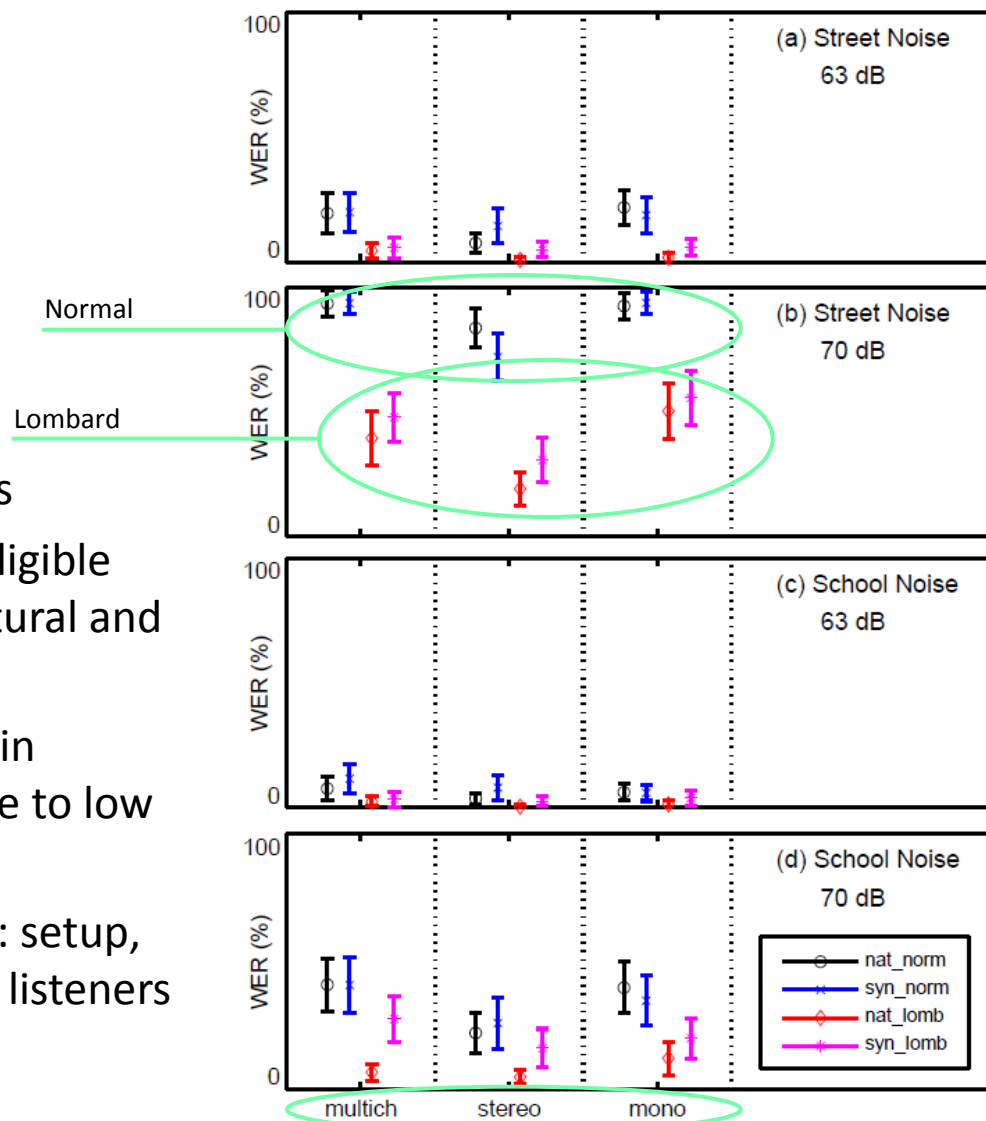
→ Word error rates (WERs) averaged for each setup



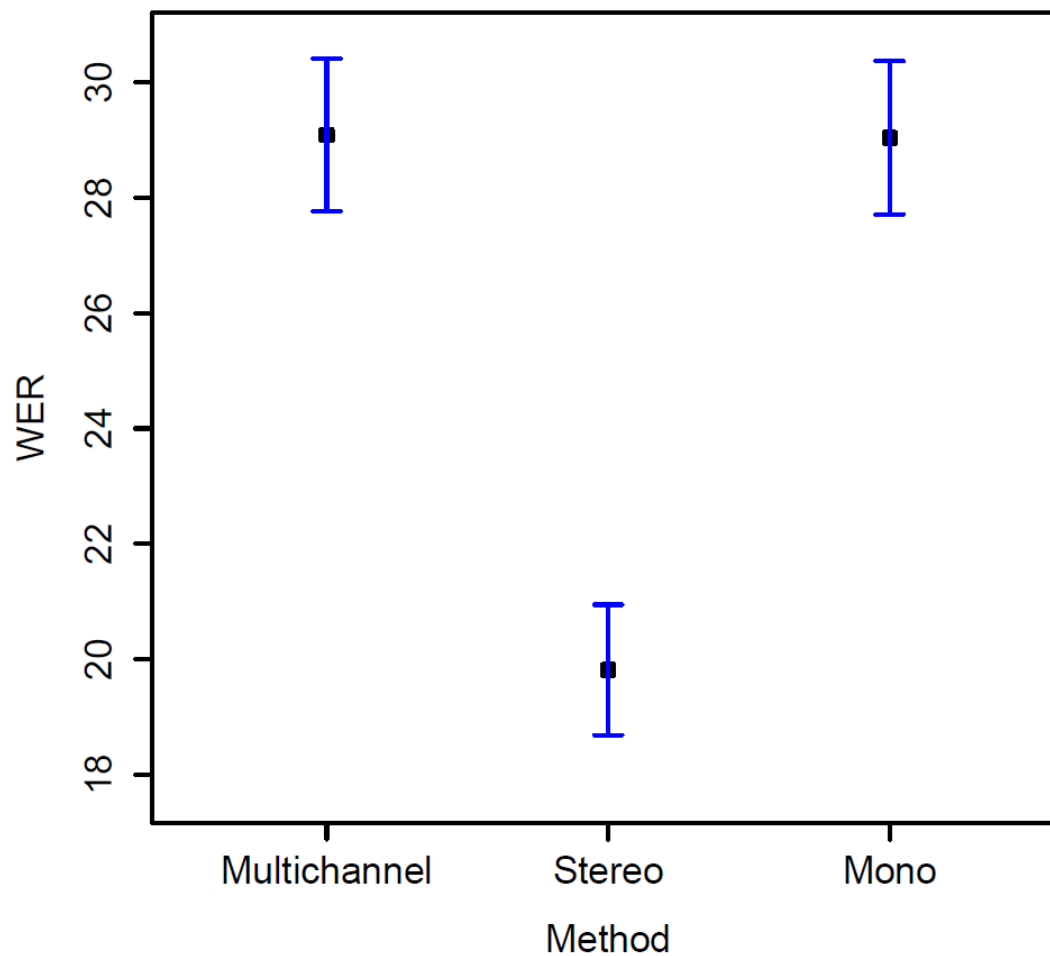
Listening test user interface

- Natural normal speech
- Synthetic normal speech
- Natural Lombard speech
- Synthetic Lombard speech

- WER trend similar in all setups
- Lombard speech is more intelligible than normal speech (both natural and synthetic)
- Street noise is more effective in masking than school noise due to low frequencies
- 5-way ANOVA (fixed variables: setup, speech type, noise type, SNR; listeners as random variable)

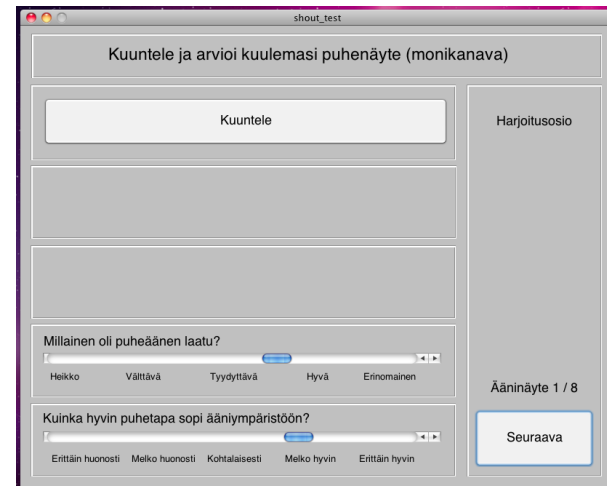


Results - WER

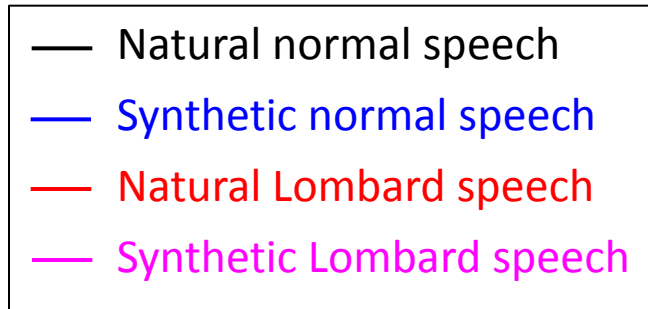


Quality and Suitability

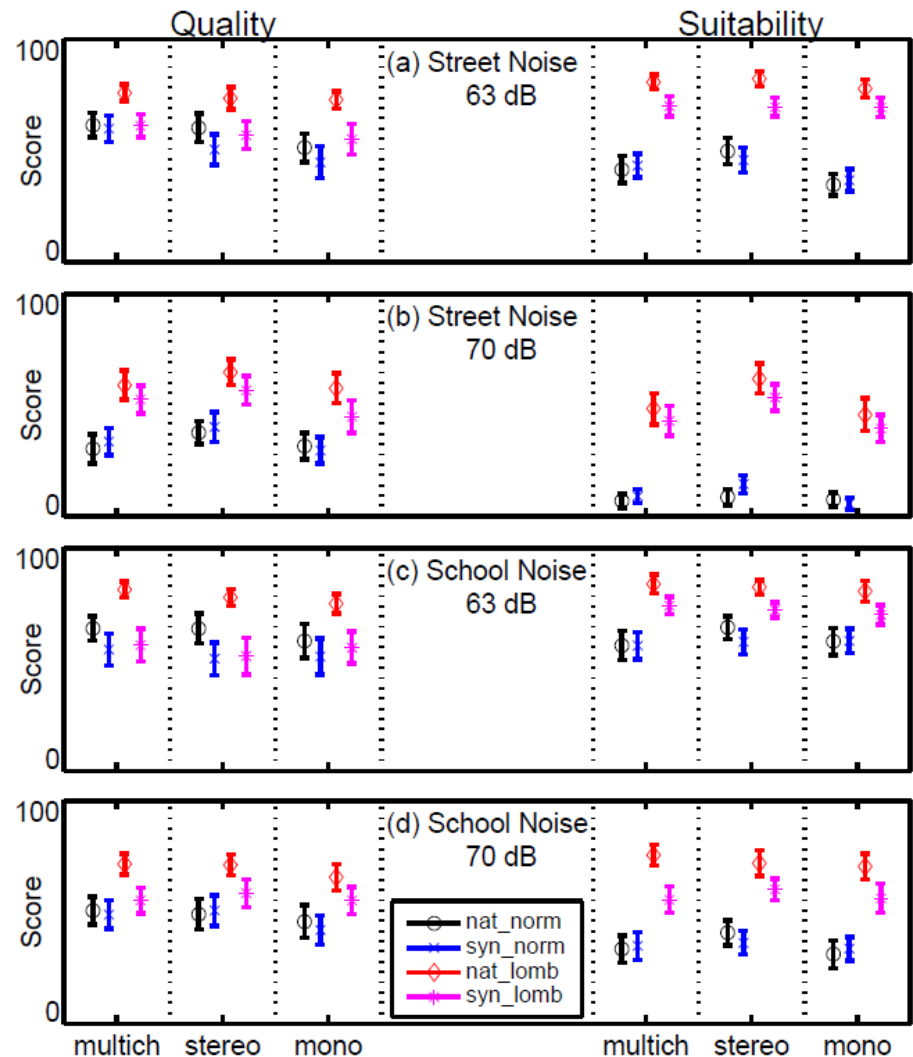
- Quality: How would you rate the quality of the speech sample?
- Suitability: How suitable was the speaking style considering the noise environment?
- Ratings averaged for each reproduction setup



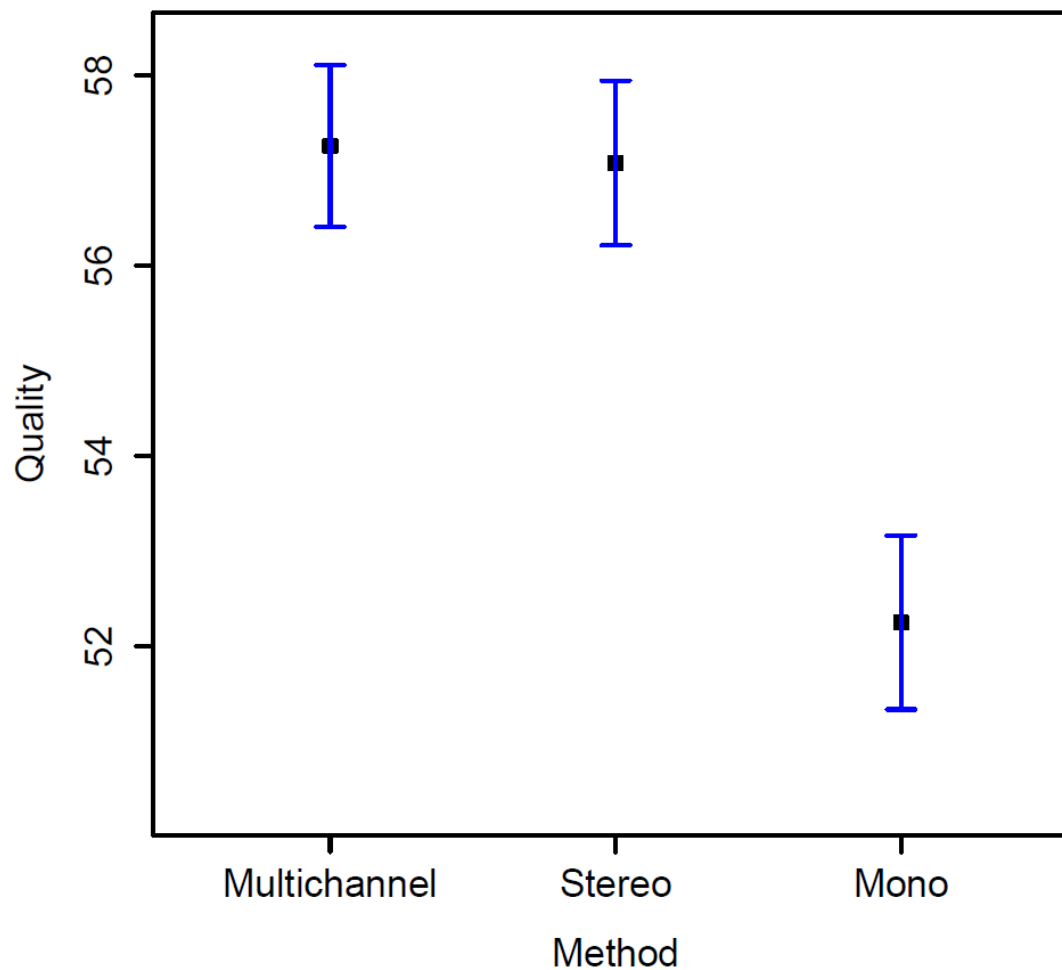
Listening test user interface



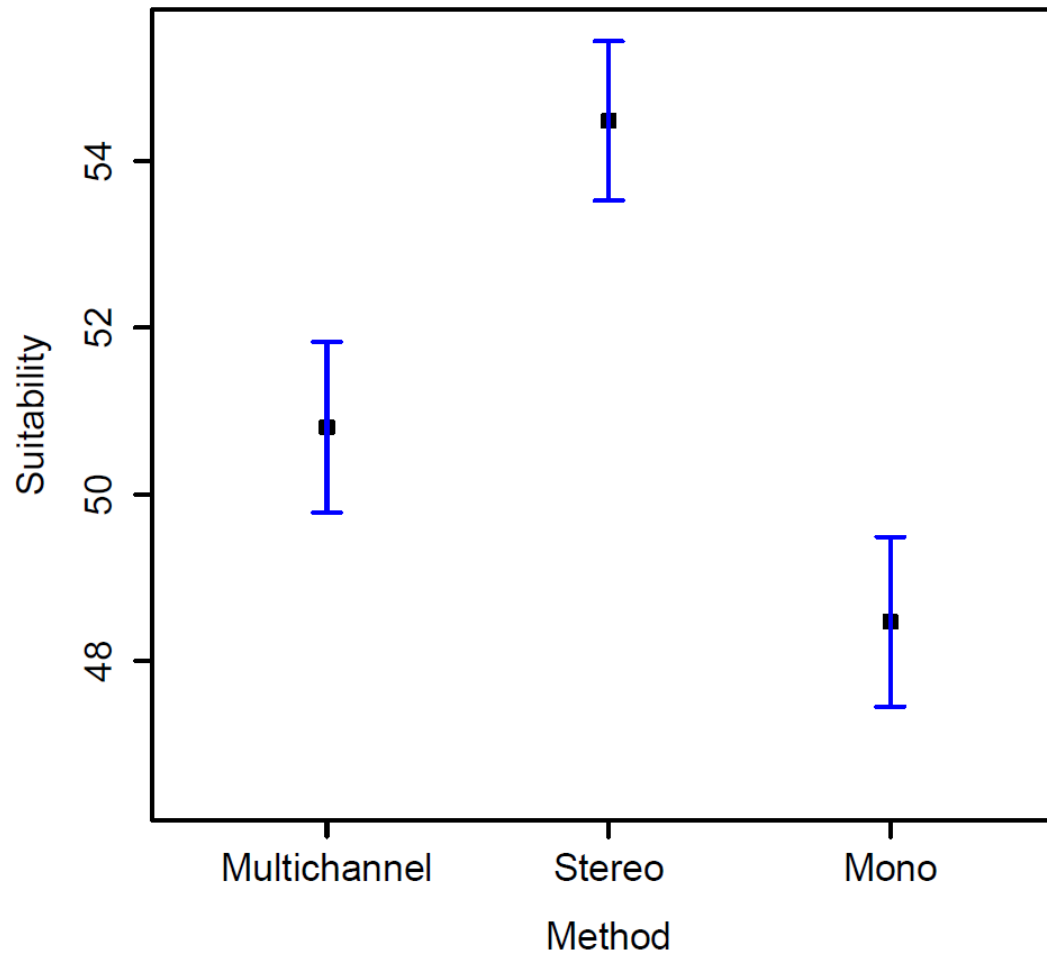
- Natural Lombard speech is rated higher in quality than other systems
- Lombard speech is rated more suitable



Results – Quality



Results – Suitability



Conclusions

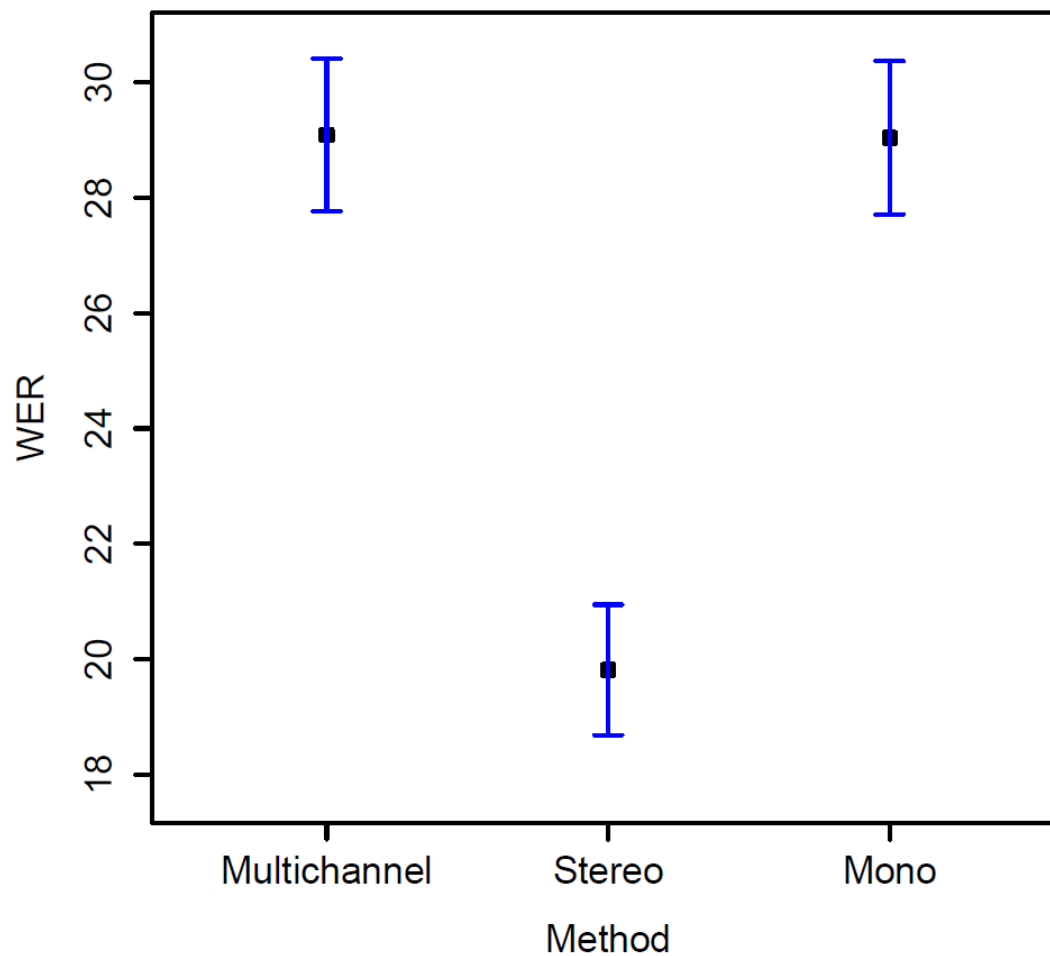
- All methods show similar results:
 - Lombard speech is more intelligible
 - WER is higher with loud noise
 - WER is higher in street noise (due to low frequencies)
- But some interesting differences:
 - Stereo setup resulted in better WER (9 percentage points difference)
 - Speech with mono setup was rated lowest in quality
 - Speech with stereo setup was rated most suitable
- Conclusion: Differences between setups exist!
- Question: Which one to use?

Discussion

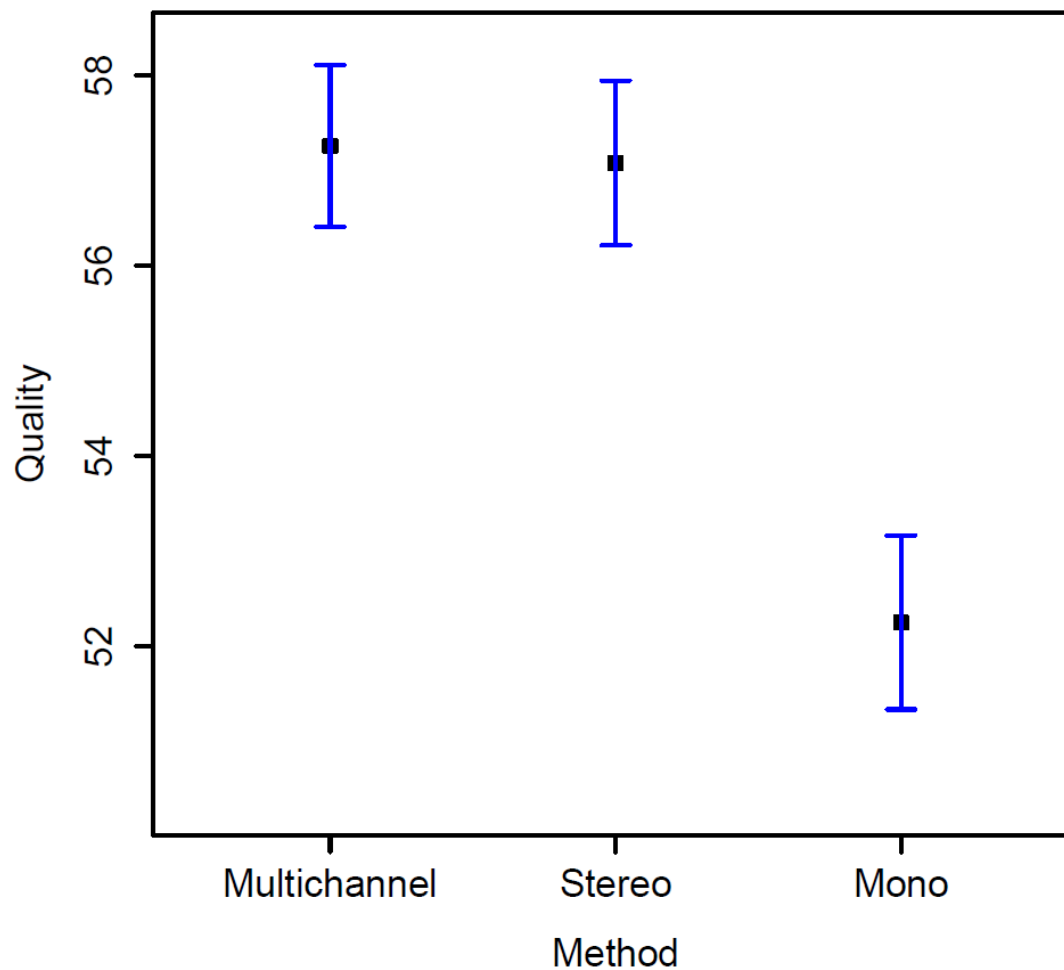
- Why stereo setup resulted in lowest WER?
 - According to previous studies, multichannel loudspeaker setup should have been the most intelligible
 - Did room response have effect to intelligibility?
 - Low reverberation (avg. 0.3 s)
- Why speech with mono headphone setup was rated lowest in quality?
 - Artifacts of synthetic speech are perceived easier with headphones compared to loudspeaker reproduction
 - But same phenomenon with natural speech samples!
- Why speech with stereo setup was rated most suitable for the noise environment?
 - Lowest WER!

Thank You!

Results - WER



Results – Quality



Results – Suitability

