Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications

Dimme de Groot[@,*] Baturalp Karslioglu[*] Odette Scharenborg[*] Jorge Martinez[*]

[@]d.c.c.j.degroot@tudelft.nl, [*]Delft University of Technology

We developed a *loudspeaker beamformer* which can be used to create a *low-acoustic-energy* region around a voice driven application, while constraining the *quality degradation* around the user.

Problem

- Acoustic interference from the loudspeaker playback system can degrade the speech recognition performance of a voice driven application (VDA), (Fig. 1, left).
- This interference can be reduced through techniques such as acoustic echo cancellation,
- This does not work well if only a low-rate link is available between the VDA and the playback system.

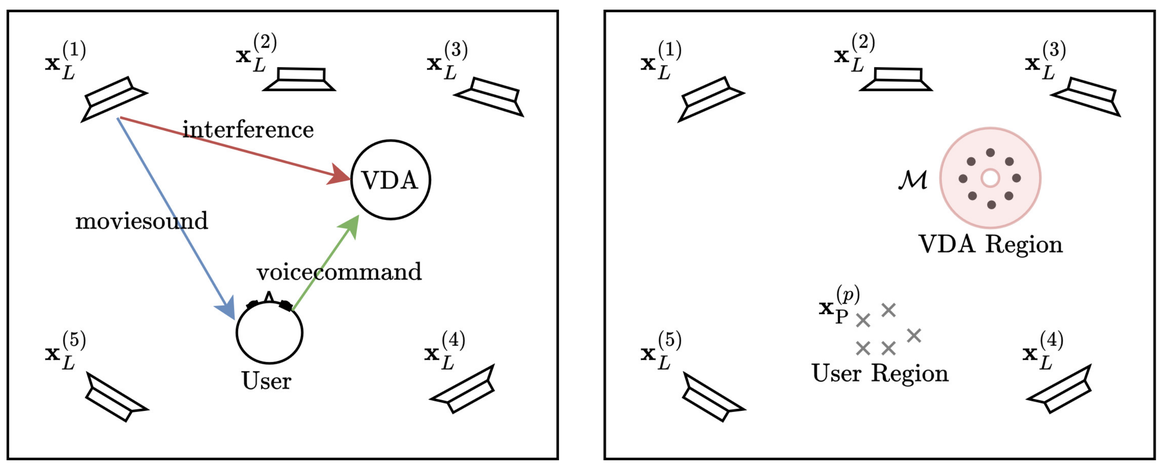


Figure 1: ***Left:*** *the user is watching a movie. From the perspective of the VDA, the movie sound is acoustic interference degrading the speech recognition performance.* ***Right:*** *we define a region around the voice assistant (red, $\mathcal{M}$) and a number of virtual control points ($\mathbf{x}_P^{(p)}$, $\times$) around the user. The acoustic energy due to the loudspeakers at the red region is minimised while the quality degradation at the virtual control points is constrained.*

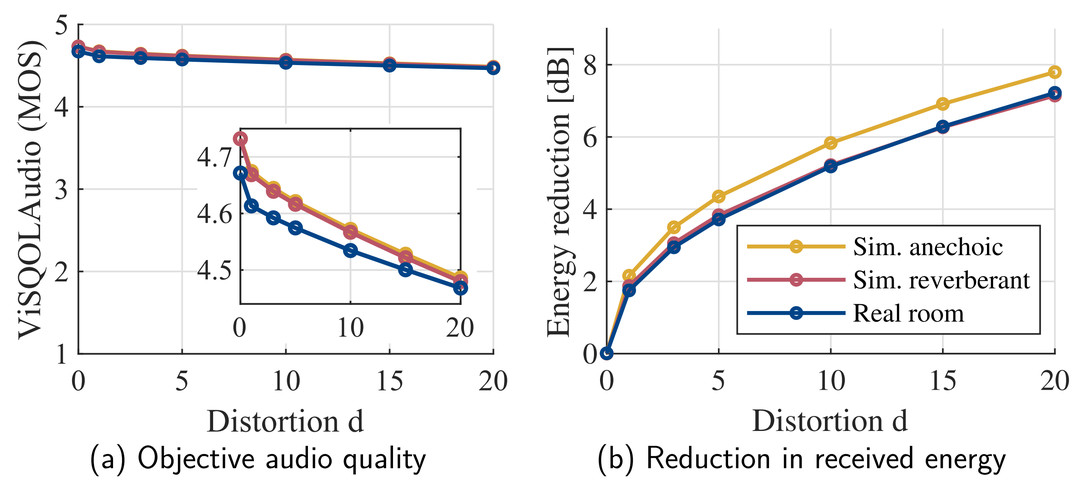
Contribution

- **Main idea:** if we reduce the acoustic interference from the loudspeakers to the VDA, the speech recognition performance will improve. However, the distortion introduced to the user should be limited.
- We develop a loudspeaker beamformer which creates a low-acoustic-energy region around the VDA (Fig. 1, right).
- The loudspeaker beamformer is constrained to keep the estimated perceived distortion low at a number of virtual control points around the listener (Fig. 1, right; *Van de Par 2005*).
- The loudspeaker beamformer is formulated as a convex optimisation problem.

Results

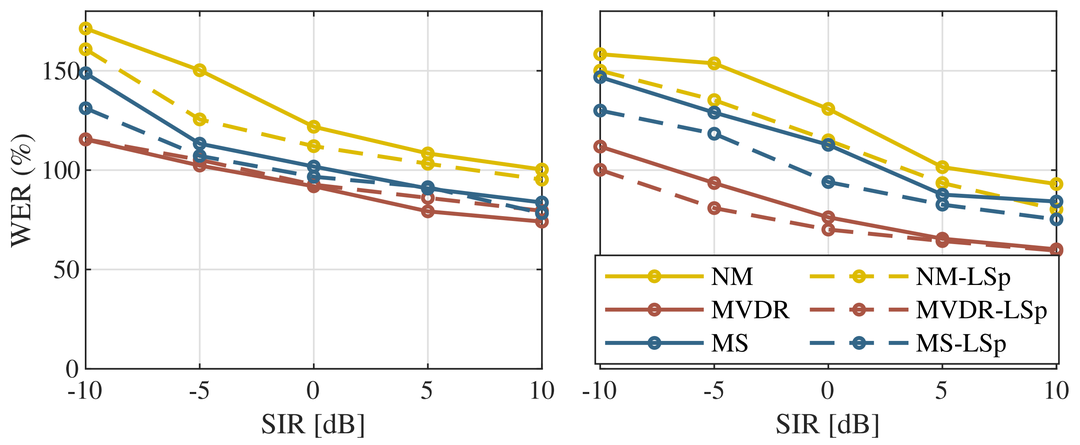
We evaluate the loudspeaker beamformer performance as:

1. **Fig. 2:** objective audio quality (ViSQOLAudio) at a number of points around the user, as a function of the allowed distortion d (parameter of the beamformer).
2. **Fig. 2:** energy reduction at a number of points inside the VDA region, as a function of the allowed distortion d.
3. **Fig. 3:** ASR performance measured as word-error-rate (WER). Here the distortion is fixed at $d = 5$, and three different microphone algorithms are considered: single-microphone (NM), MVDR beamformer and a microphone spotformer (MS; *Martinez 2015*).



(a) Objective audio quality (b) Reduction in received energy

Figure 2: The results for objective audio quality around the user and the achieved energy reduction around the VDA.



(a) Simulated reverberant (b) Real room

Figure 3: The word error rate for different signal-to-interference ratios in a simulated reverberant and a real room. The results are presented without (solid line) and with (dashed line) loudspeaker beamforming.

Conclusion

- We developed a loudspeaker beamformer which improves the speech recognition performance of a voice driven application.
- We showed that the algorithm performs well in a real room.
- The major limitation of the algorithm is the computational complexity.

Van de Par et al., *A Perceptual Model for Sinusoidal Audio Coding based on Spectral Integration*, 2005.

Martinez et al., *A robust region-based near-field beamformer*, 2015.

QR: https://doi.org/10.1109/ICASSP49660.2025.10889702

TU Delft - Intelligent Systems