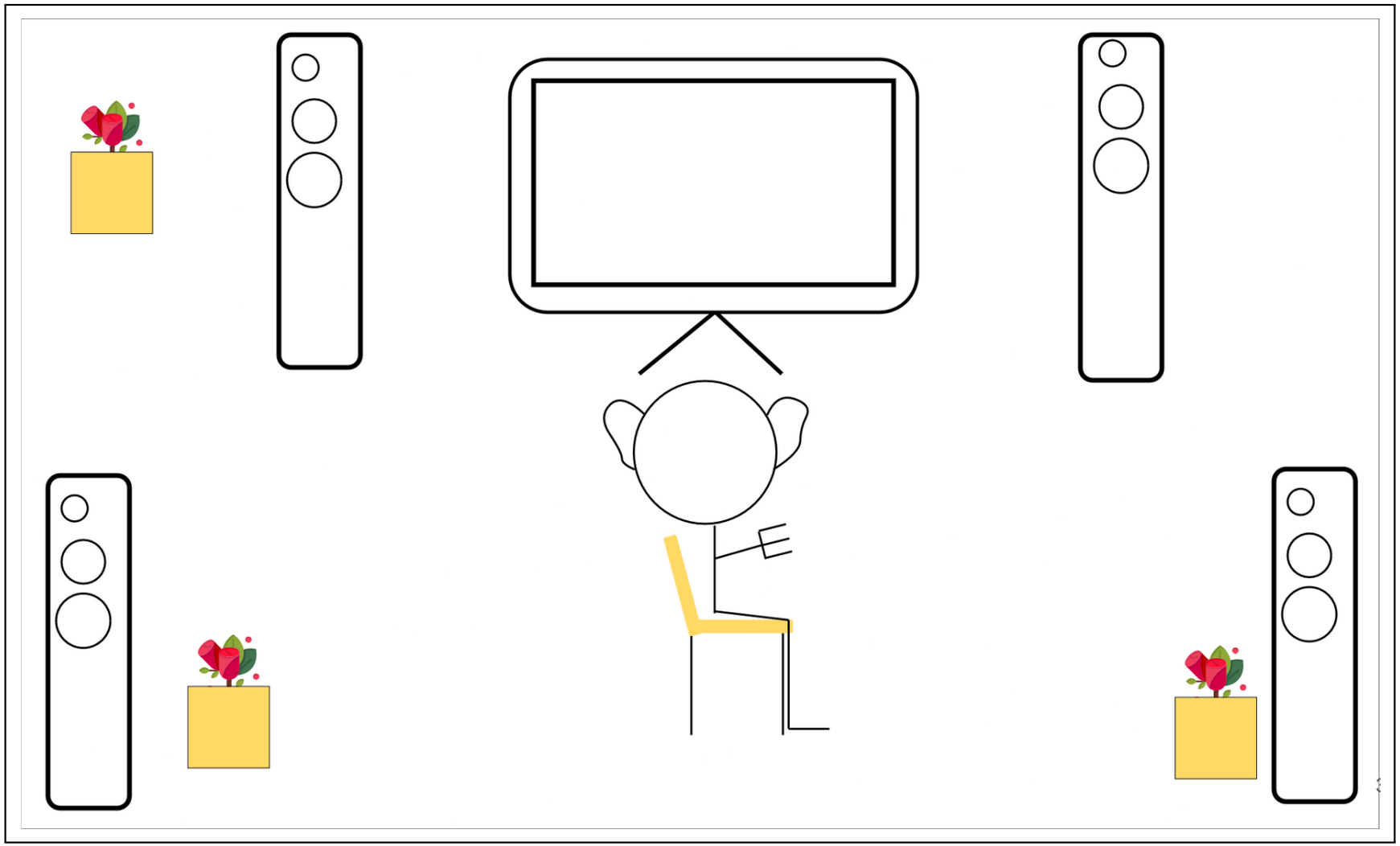
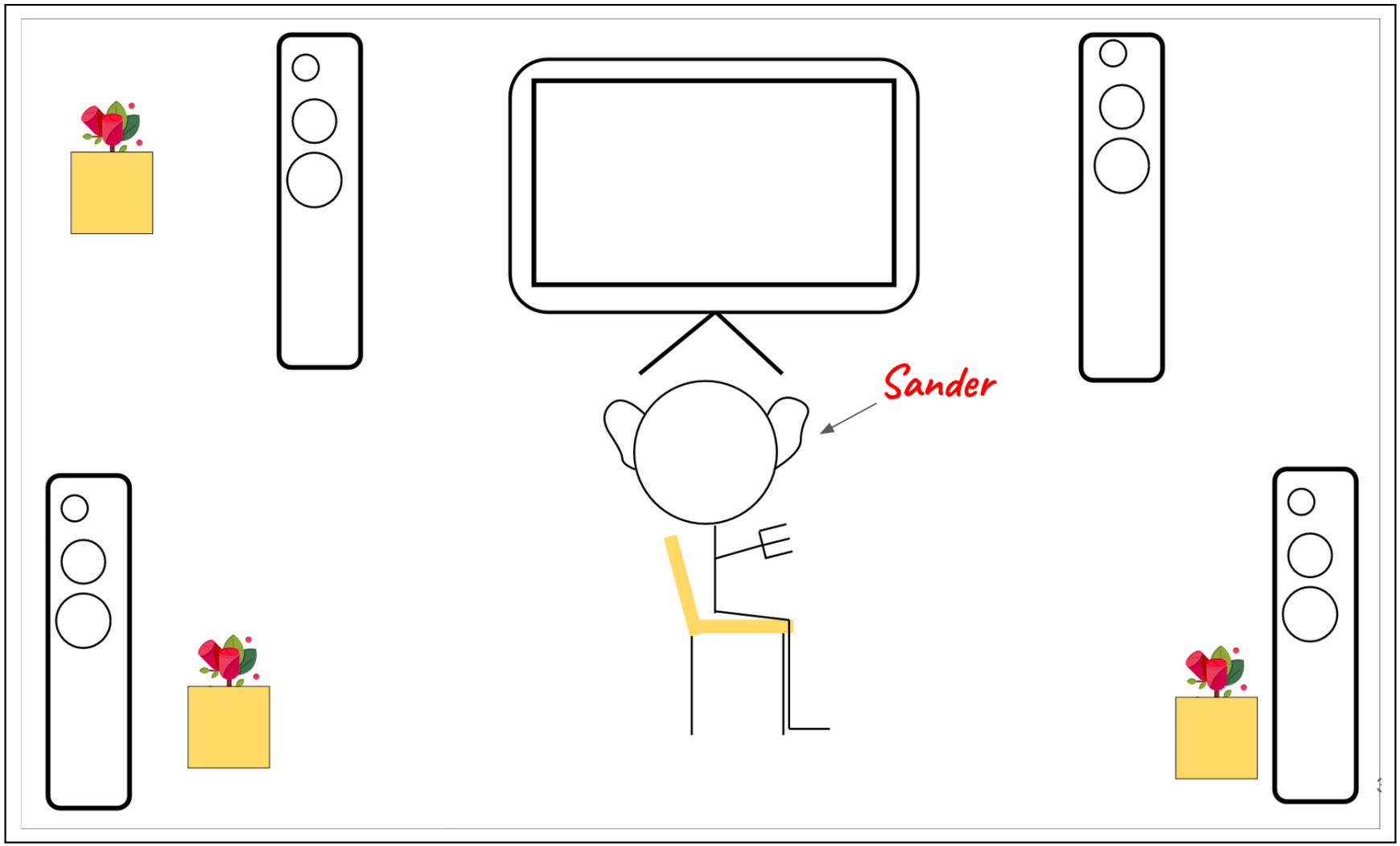
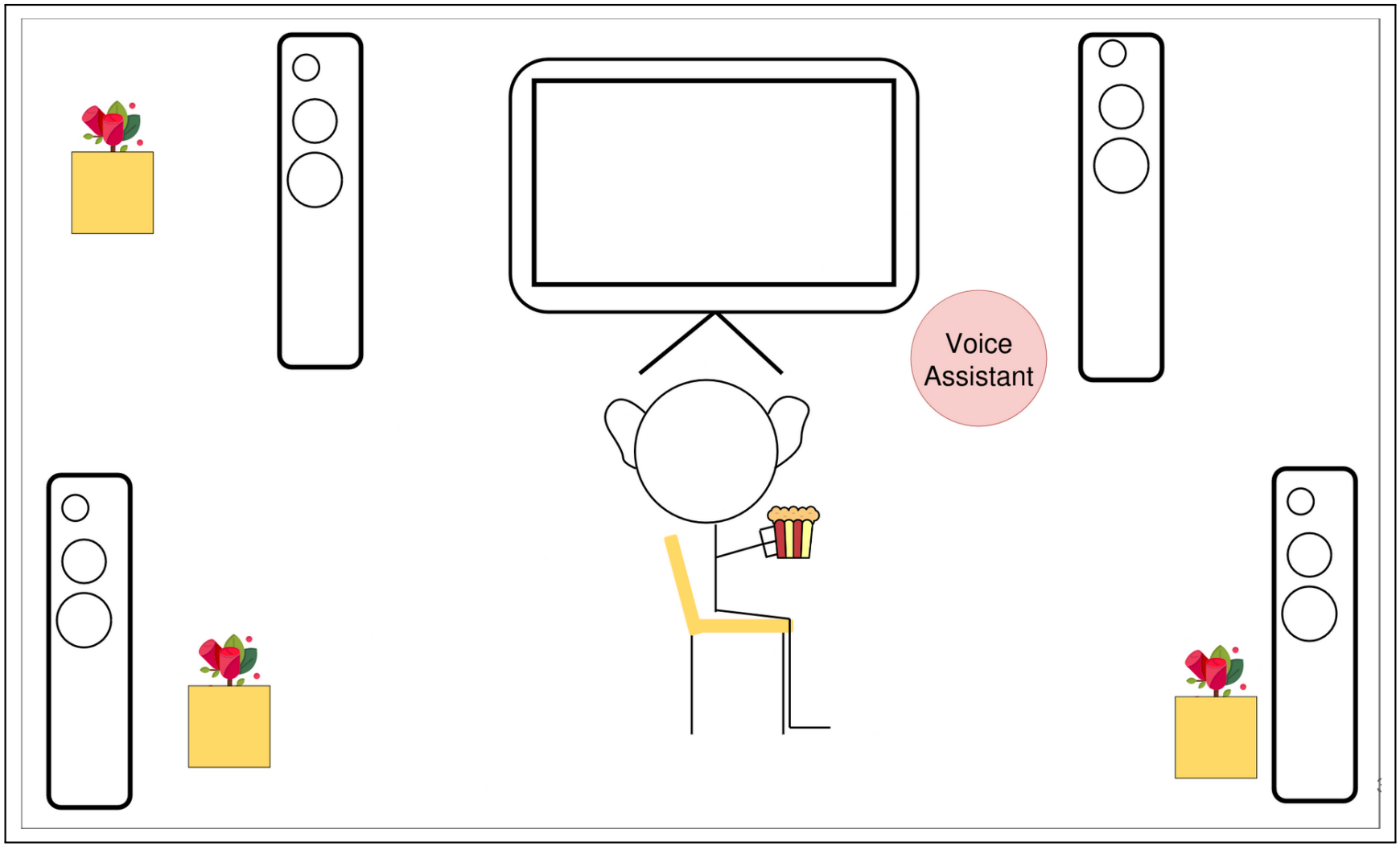
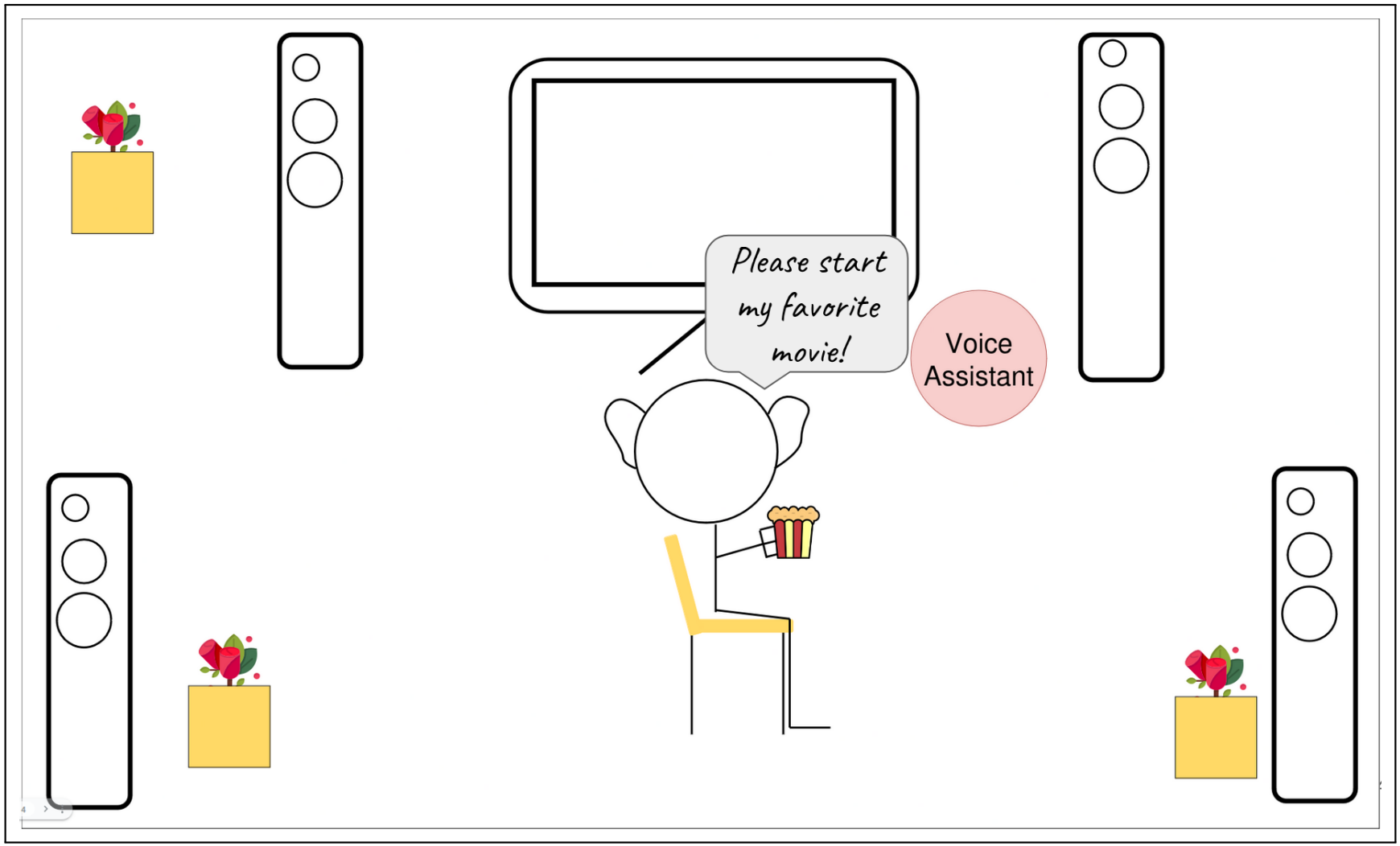
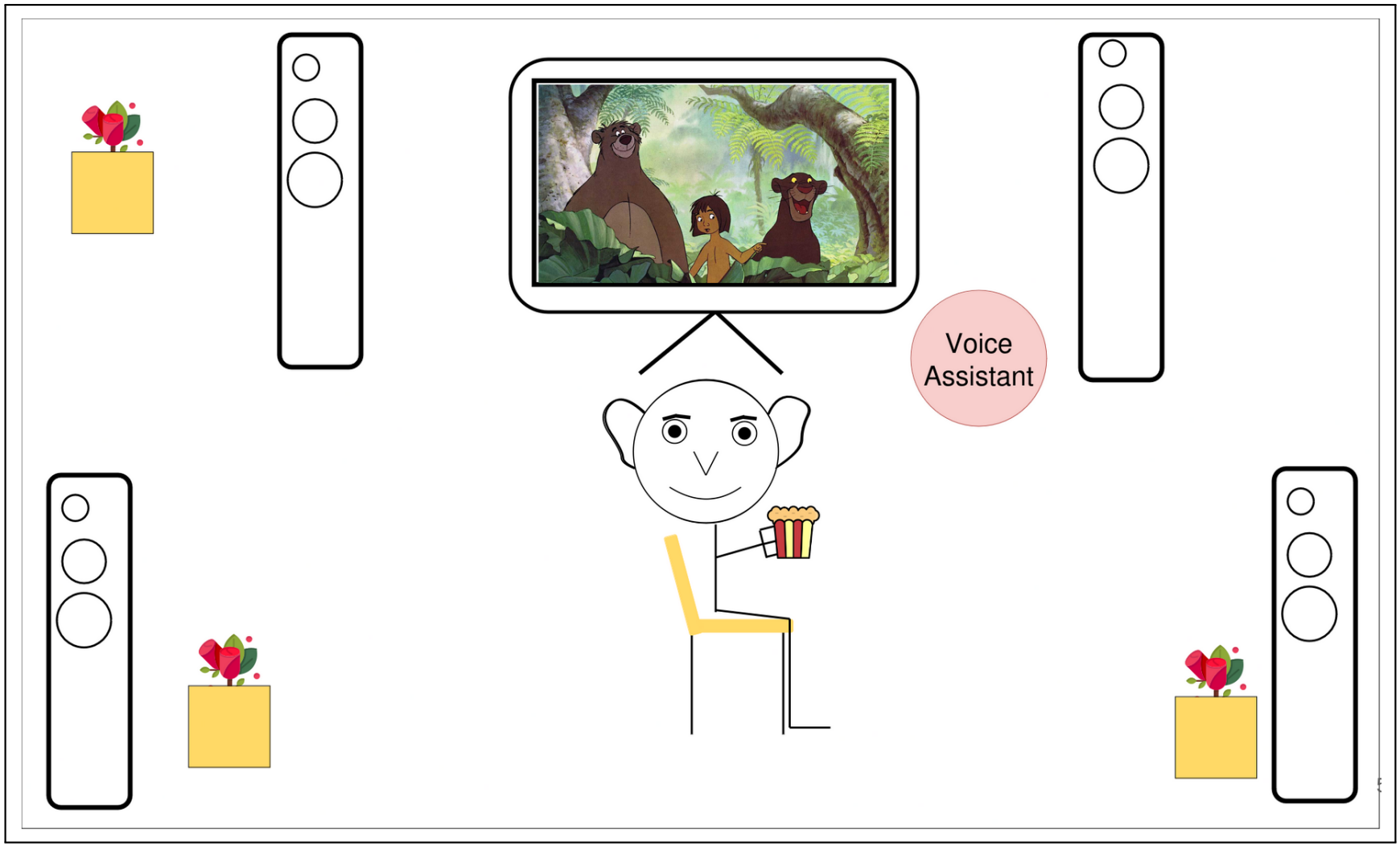
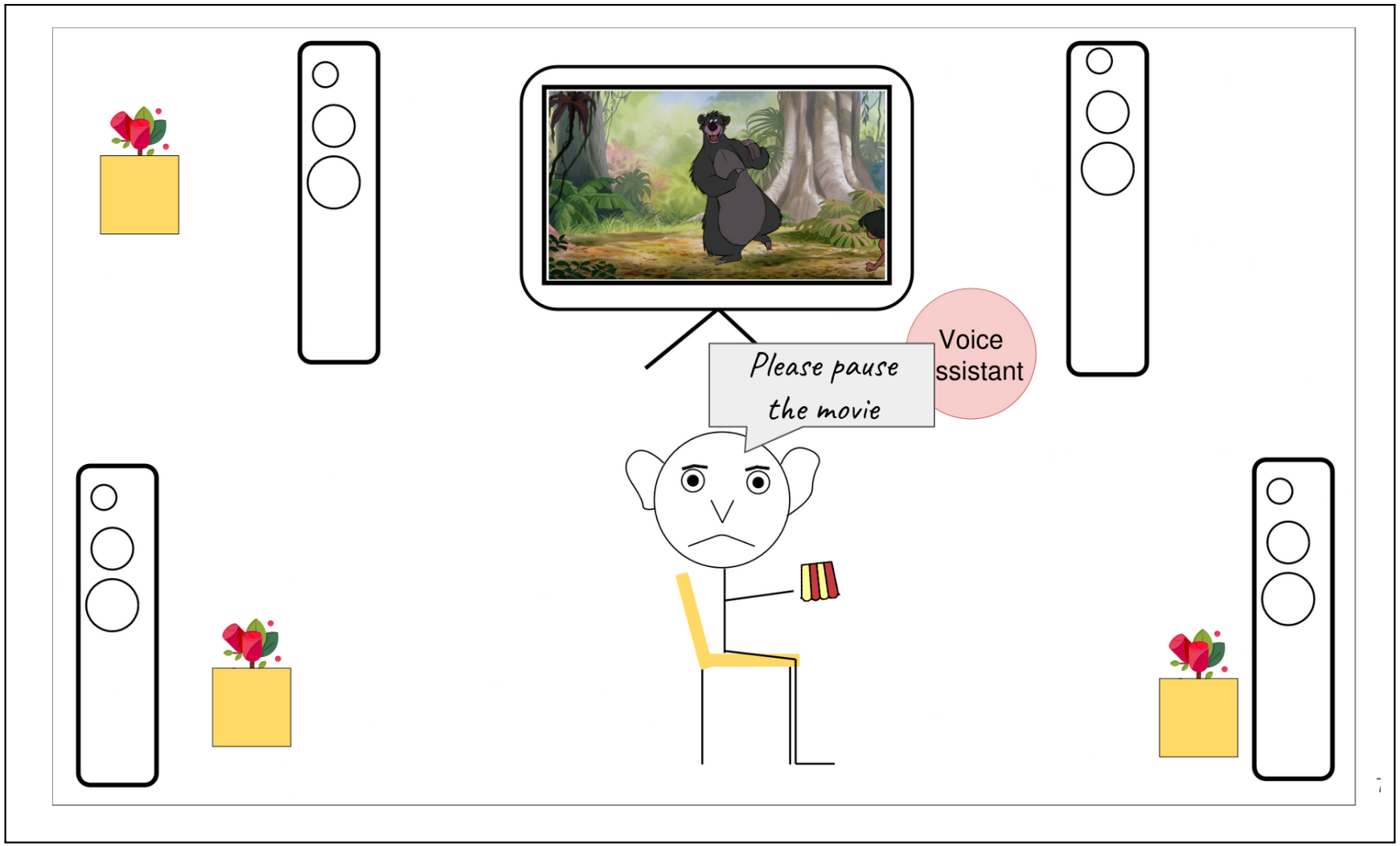


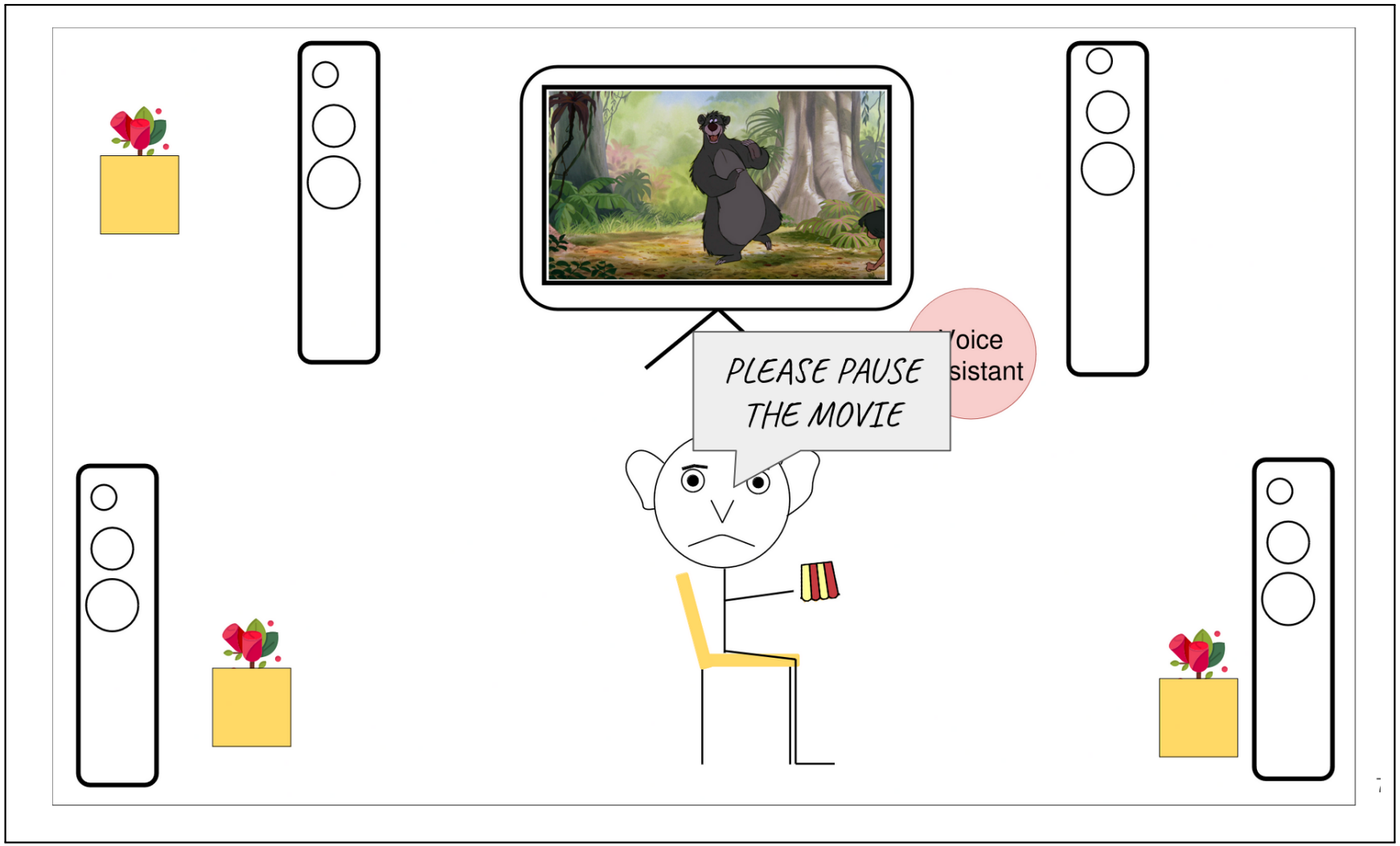

Sander

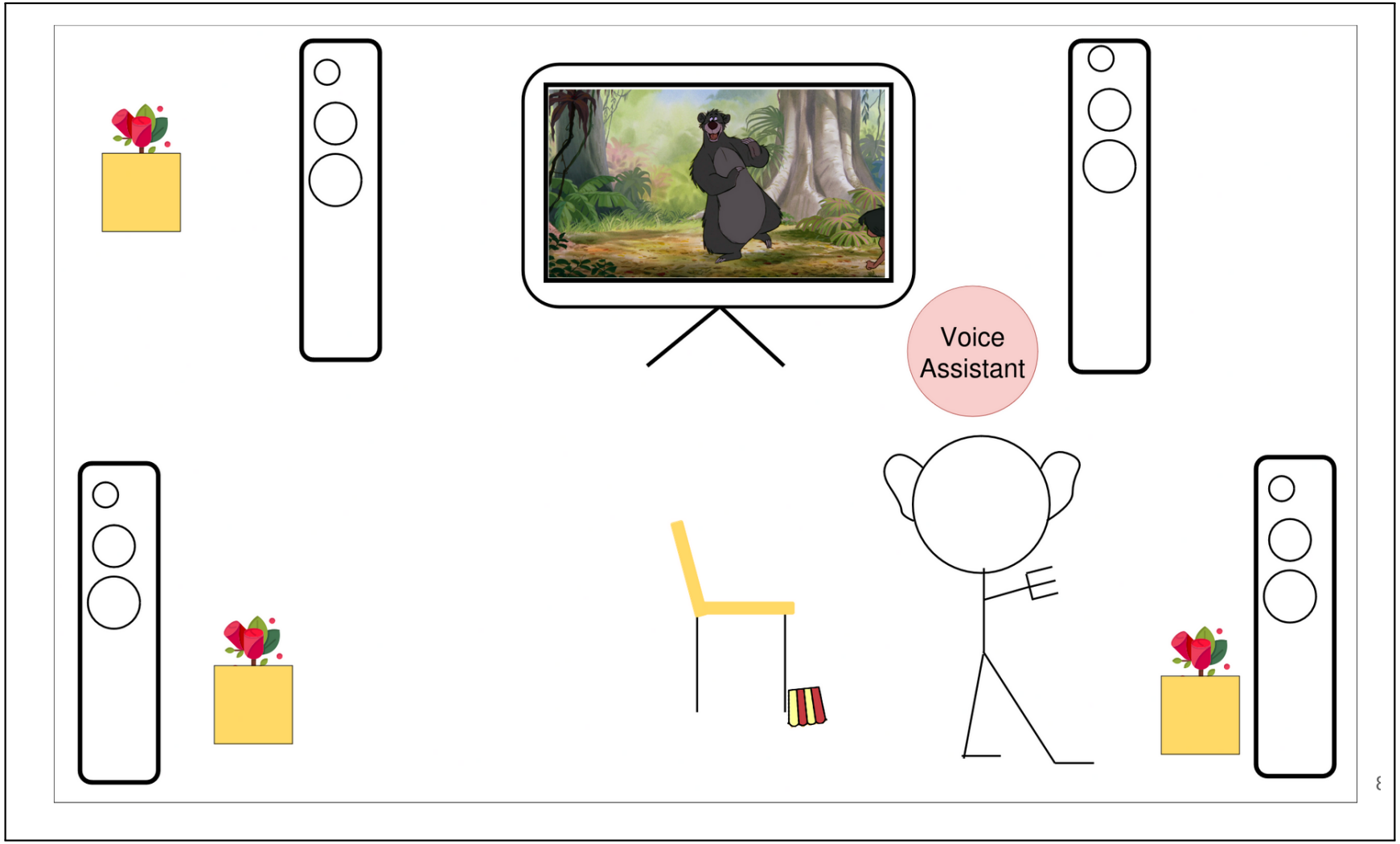

Voice
Assistant


Please start my favorite movie!
Voice Assistant


Voice
Assistant


Please pause the movie
Voice ssistant


PLEASE PAUSE
THE MOVIE
oice
sistant


Voice
Assistant

Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications

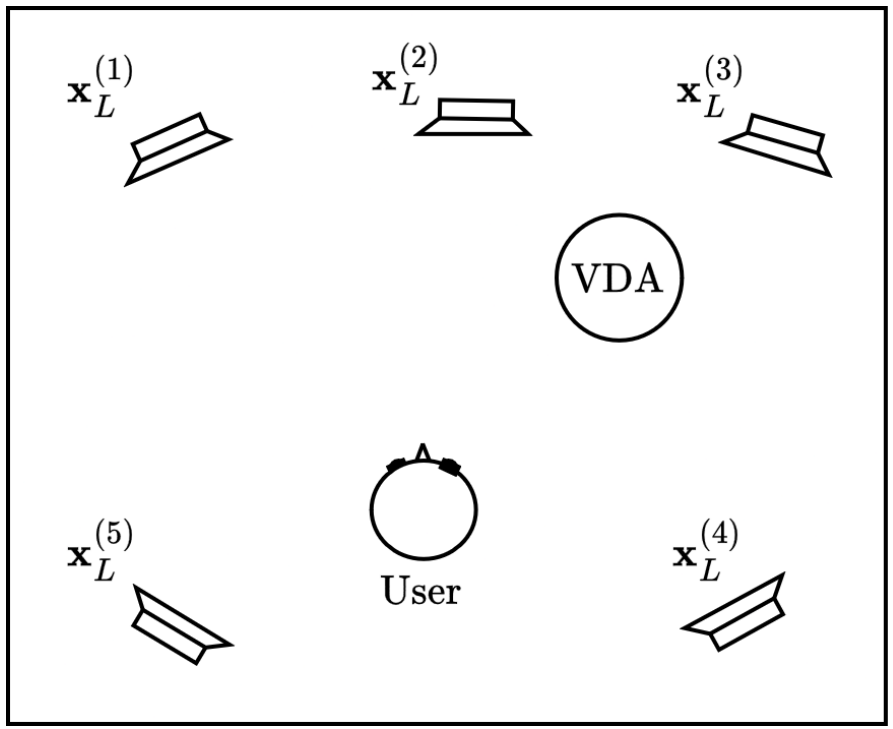
Dimme de Groot, Baturalp Karslioglu, Odette Scharenborg and Jorge Martinez

Delft University of Technology, The Netherlands

March 17, 2025

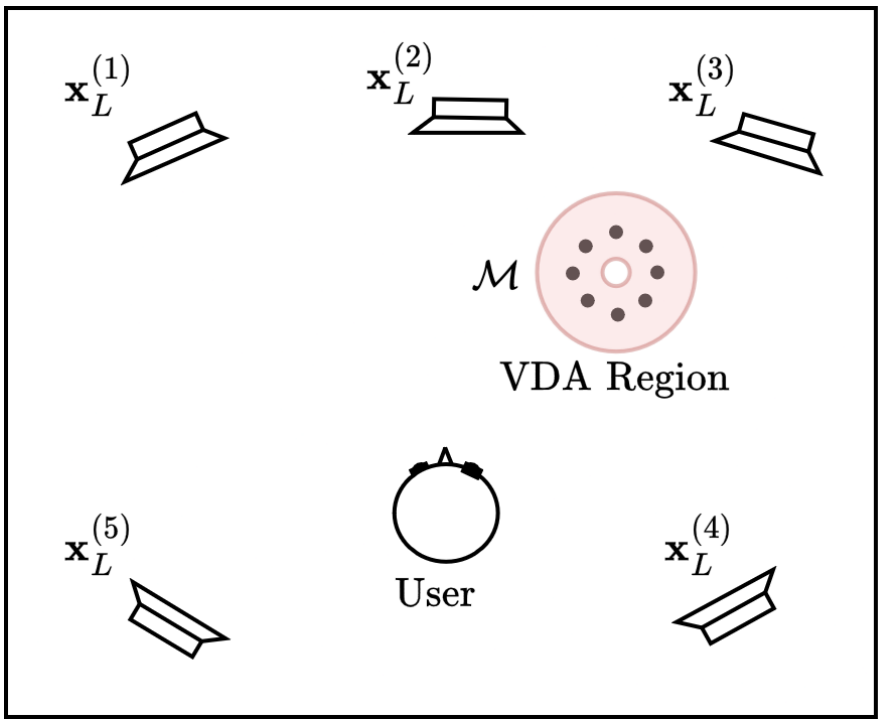
Problem & Idea

- **Problem:** acoustic interference from loudspeakers degrades speech recognition performance of voice driven application (VDA)



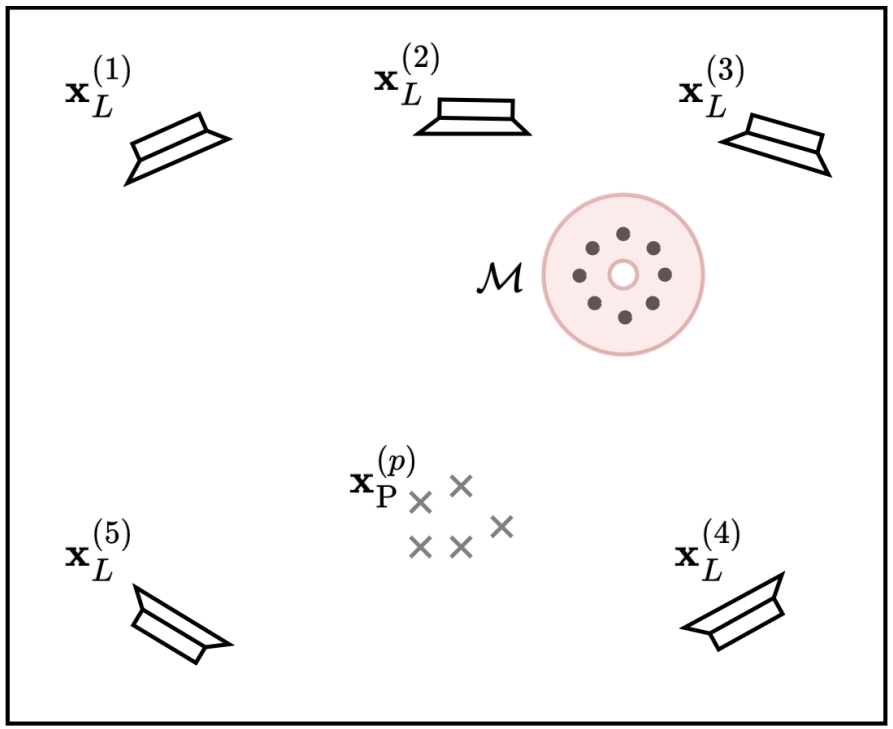
Problem & Idea

- **Problem:** acoustic interference from loudspeakers degrades speech recognition performance of voice driven application (VDA)
- **Idea (a):** use *loudspeaker spotforming* to generate a *low-acoustic-energy* region near the voice-driven application.



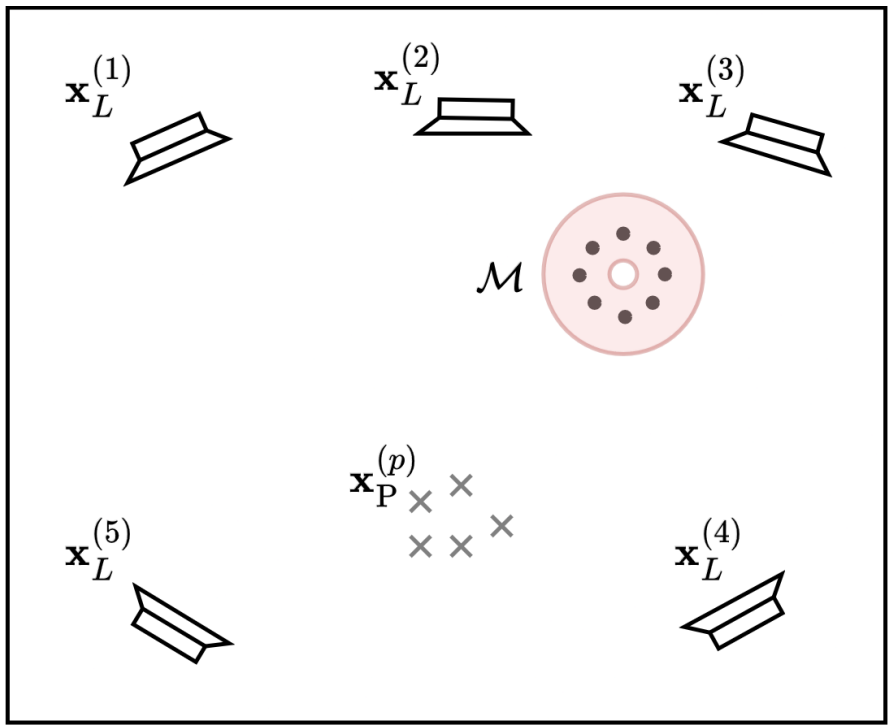
Problem & Idea

- **Problem:** acoustic interference from loudspeakers degrades speech recognition performance of voice driven application (VDA)
- **Idea (a):** use *loudspeaker spotforming* to generate a *low-acoustic-energy* region near the voice-driven application.
- **Idea (b):** keep the *perceived distortion* around the user constrained.



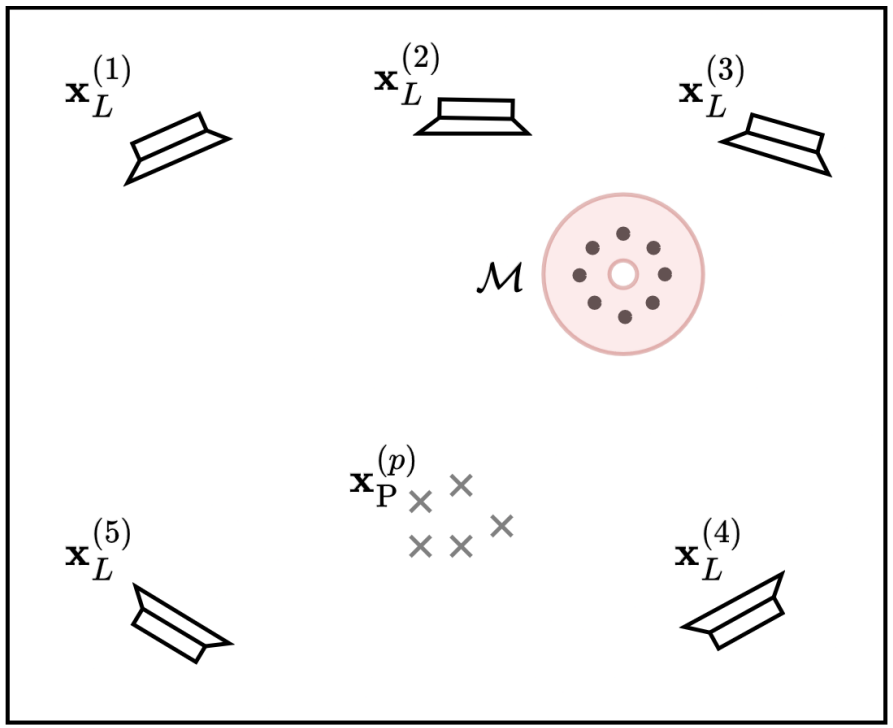
Our Solution & Toy Optimization Problem (1)

- Numerically compute spatial covariance matrix $\mathbf{R}_{\mathcal{M}}(\omega)$ over region $\mathcal{M}$. Each matrix entry $\{\mathbf{R}_{\mathcal{M}}(\omega)\}_{ll'}$ represents the covariance between two loudspeakers l and l'.



Our Solution & Toy Optimization Problem (1)

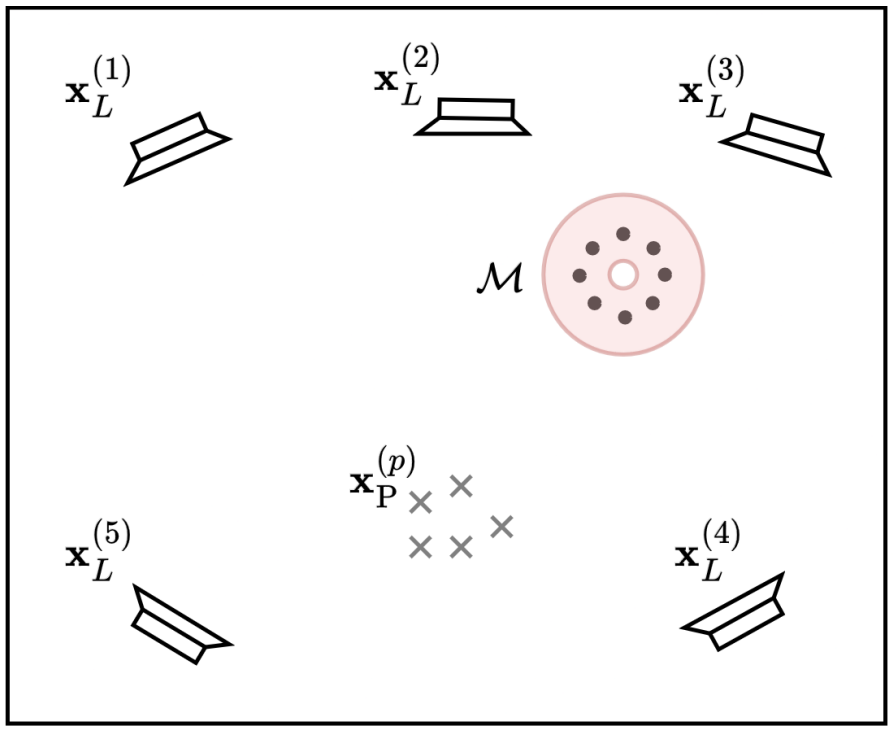
- Numerically compute spatial covariance matrix $\mathbf{R}_{\mathcal{M}}(\omega)$ over region $\mathcal{M}$. Each matrix entry $\{\mathbf{R}_{\mathcal{M}}(\omega)\}_{ll'}$ represents the covariance between two loudspeakers l and l'.
- Use a measure of auditory masking $D\left(s_{\text{ref}}^{(p)}, s_{\text{rec}}^{(p)}\right)$ to constrain perceived distortion between the reference received signal $s_{\text{ref}}^{(p)}$ and the estimated received signal $s_{\text{rec}}^{(p)}$ at virtual locations $\mathbf{x}_{\text{P}}^{(p)}$ for all p.



Our Solution & Toy Optimization Problem (2)

$$\begin{aligned} \min \quad & \mathbf{s}_{\text{play}}^{H} \mathbf{R}_{\mathcal{M}} \mathbf{s}_{\text{play}} \\ \text{s.t.} \quad & D\left(s_{\text{ref}}^{(p)}, s_{\text{rec}}^{(p)}\right) < d \;\; \forall p \end{aligned}$$

- Acoustic energy in region $\mathcal{M}$ is minimised through $\mathbf{R}_{\mathcal{M}}(\omega)$.
- While distortion between reference and received signal (precomputed based on direct path) at virtual control points is bounded by $D\left(s_{\text{ref}}^{(p)}, s_{\text{rec}}^{(p)}\right) < d$.

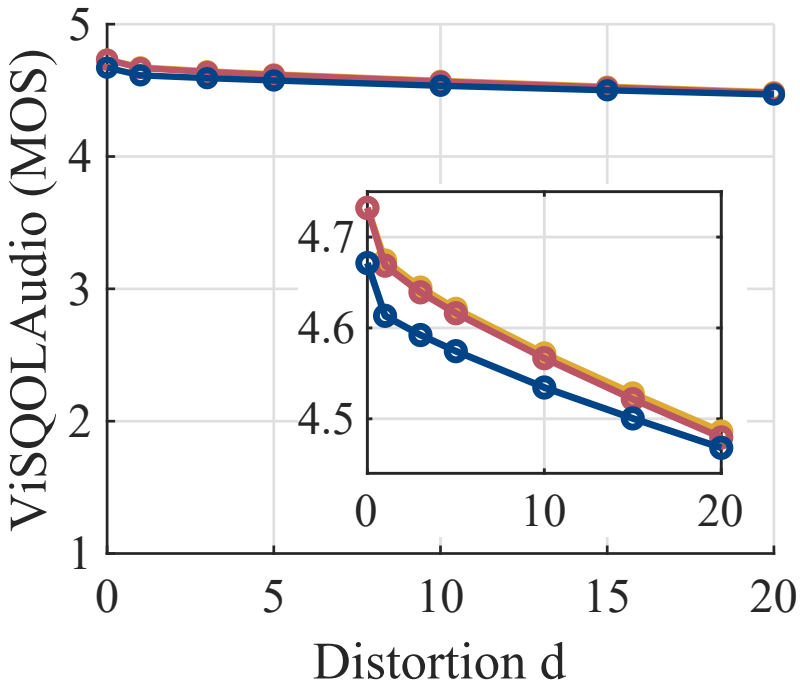


Results (1): no speech recognition

- Three conditions:
  - Simulated anechoic conditions (●)
  - Simulated reverberant conditions ($T_{60} \approx 220$ ms, ●)
  - Real room ($T_{60} \approx 220$ ms, ●)

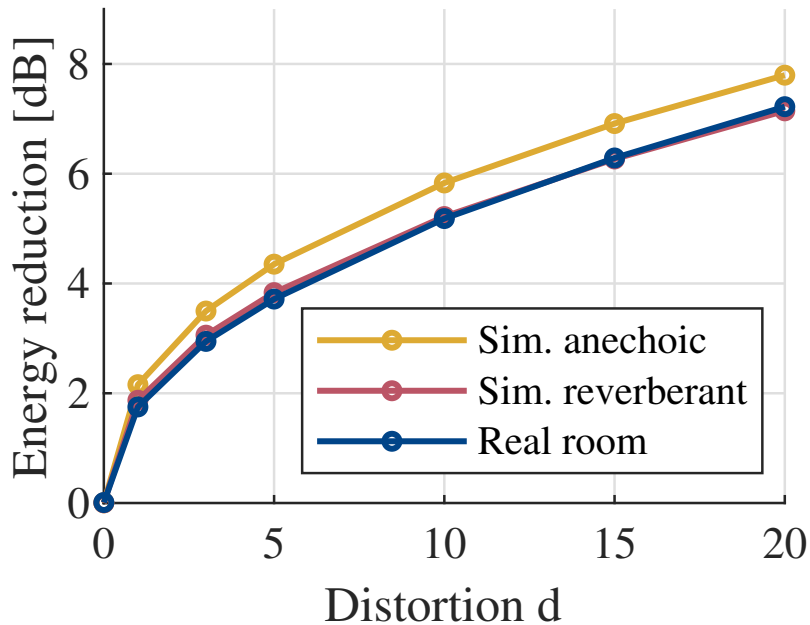
Results (1): no speech recognition

- Three conditions:
  - Simulated anechoic conditions (●)
  - Simulated reverberant conditions ($T_{60} \approx 220$ ms, ●)
  - Real room ($T_{60} \approx 220$ ms, ●)
- Speech Quality in user region (ViSQOLAudio) as function of distortion d.



Results (1): no speech recognition

- Three conditions:
  - Simulated anechoic conditions (●)
  - Simulated reverberant conditions ($T_{60} \approx 220$ ms, ●)
  - Real room ($T_{60} \approx 220$ ms, ●)
- Speech Quality in user region (ViSQOLAudio) as function of distortion d.
- Energy reduction at microphones of VDA as function of distortion d.



Results (2): ASR performance - Word Error Rate

- We use a circular microphone array with 8 microphones
  - Microphone Spotforming (●)
  - Single microphone (●)
  - MVDR beamforming (●)

Results (2): ASR performance - Word Error Rate

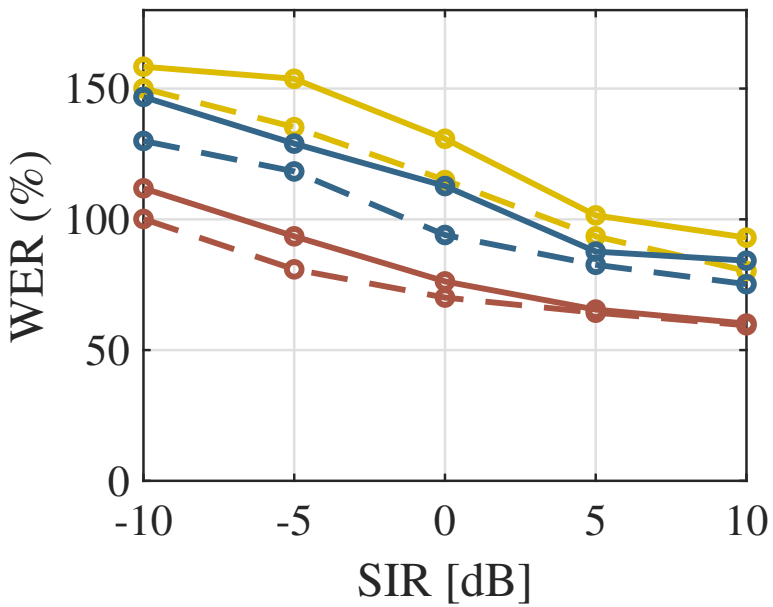
- We use a circular microphone array with 8 microphones
  - Microphone Spotforming (●)
  - Single microphone (●)
  - MVDR beamforming (●)
- We use distortion $d = 5$

Results (2): ASR performance - Word Error Rate

- We use a circular microphone array with 8 microphones
  - Microphone Spotforming (●)
  - Single microphone (●)
  - MVDR beamforming (●)
- We use distortion $d = 5$
- We consider each of the microphone algorithms including (−−) and excluding (−) loudspeaker spotforming.

Results (2): ASR performance - Word Error Rate

- We use a circular microphone array with 8 microphones
  - Microphone Spotforming (●)
  - Single microphone (●)
  - MVDR beamforming (●)
- We use distortion $d = 5$
- We consider each of the microphone algorithms including (−−) and excluding (−) loudspeaker spotforming.
- Due to time constraints, only the results for the real room are presented.



Conclusion & Discussion

- We were able to create a low-acoustic-energy region which improved ASR performance in a real-life scenario,

Conclusion & Discussion

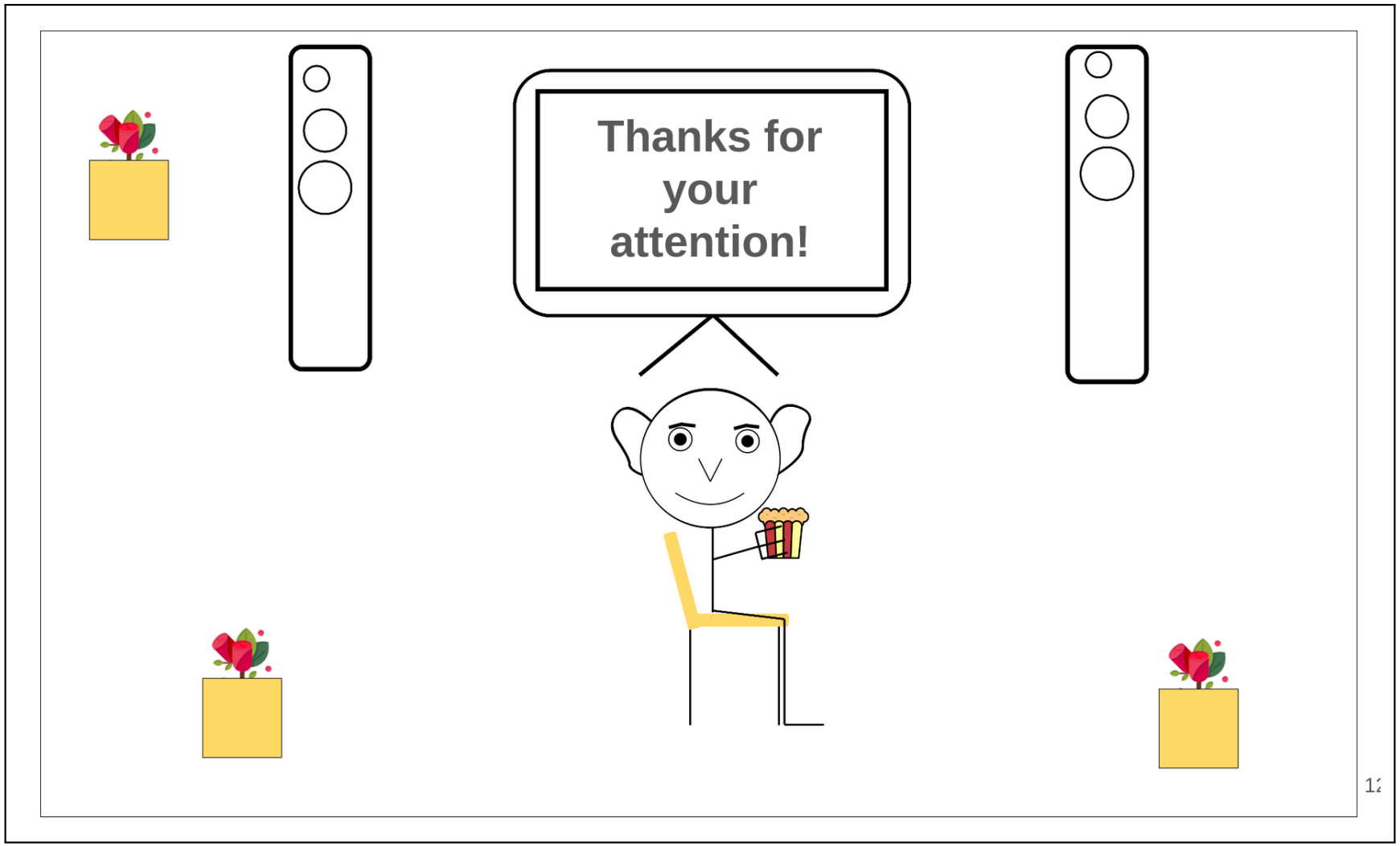
- We were able to create a low-acoustic-energy region which improved ASR performance in a real-life scenario,
- At the same time, the distortion introduced remained limited (according to an objective audio quality measure).

Conclusion & Discussion

- We were able to create a low-acoustic-energy region which improved ASR performance in a real-life scenario,
- At the same time, the distortion introduced remained limited (according to an objective audio quality measure).
- In future work, subjective tests should be used to confirm this.

Conclusion & Discussion

- We were able to create a low-acoustic-energy region which improved ASR performance in a real-life scenario,
- At the same time, the distortion introduced remained limited (according to an objective audio quality measure).
- In future work, subjective tests should be used to confirm this.
- Additionally, the computational complexity is high.


Thanks for your attention!