Towards Embodied Speech Recognition

Tanis Sarbatananda
Dr. Jesse Thomason
Tejas Srinivasan

8/4/2022

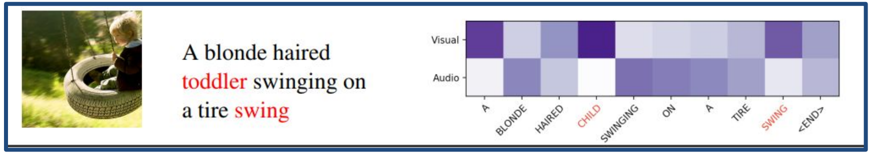
Introduction

- Area of improvements on current speech recognitions
- Expanding my technical skills and learn Python
- Increase the accessibility of speech recognition when encountering corrupted instructions and handle different accents

Literature Review

- *MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. Percept. Psychophys. 24, 253–257. doi: 10.3758/BF03206096*

- *Srinivasan, Tejas, et al. "Multimodal speech recognition with unstructured audio masking." arXiv preprint arXiv:2010.08642 (2020).*



Pick up the spoon on the table ⟷ Pick up the soon on the table

- *Kolve, Eric, et al. "Ai2-thor: An interactive 3d environment for visual ai." arXiv preprint arXiv:1712.05474 (2017).*

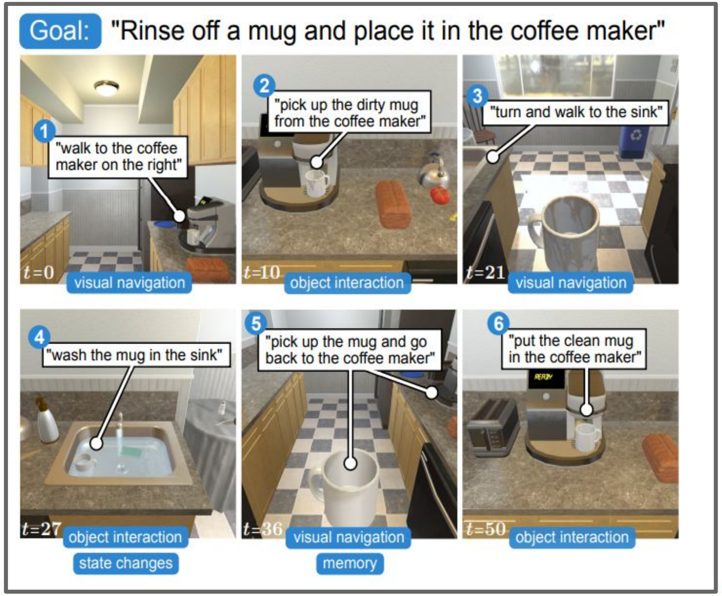
- Krol, Jacob. *Amazon's Astro Home Robot Puts Alexa on Wheels — but Is It Worth $1,000?*, Cnn Underscored, 28 Sep. 2021, https://www.cnn.com/cnn-underscored/electronics/astro-amazon-robot-hands-on.

Research Questions

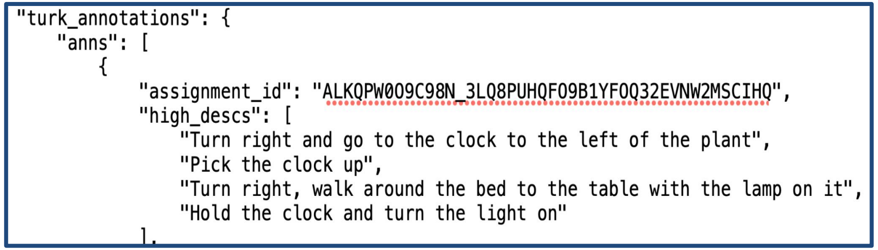
- What information do robots need to perform tasks effectively?
- What are some aspects that at-home-robots can improve?
- If a speech signal is corrupted, what kind of information can be utilized to compensate?
- Will the combination of speech signal and visual scene information increase the accuracy of automatic speech recognition?

Research Methods

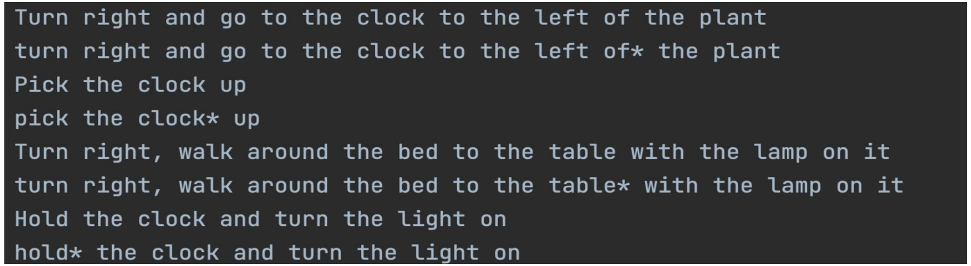
- AI2-THOR (virtual environment)
- ALFRED (text-based language directives)


Goal: "Rinse off a mug and place it in the coffee maker"
1 "walk to the coffee maker on the right"
t=0
visual navigation
2 "pick up the dirty mug from the coffee maker"
t=10
object interaction
3 "turn and walk to the sink"
t=21
visual navigation
4 "wash the mug in the sink"
t=27
object interaction
state changes
5 "pick up the mug and go back to the coffee maker"
t=36
visual navigation
memory
6 "put the clean mug in the coffee maker"
t=50
object interaction

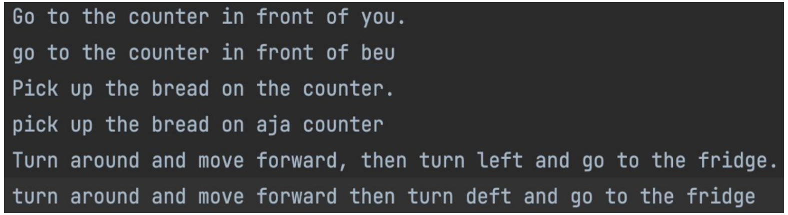
Findings/ Data



```
"turk_annotations": {
    "anns": [
        {
            "assignment_id": "ALKQPW009C98N_3LQ8PUHQF09B1YF0Q32EVNW2MSCIHQ",
            "high_descs": [
                "Turn right and go to the clock to the left of the plant",
                "Pick the clock up",
                "Turn right, walk around the bed to the table with the lamp on it",
                "Hold the clock and turn the light on"
            ],
```

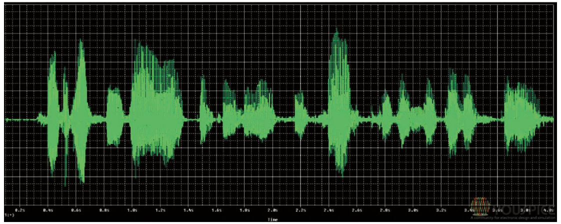


```
Turn right and go to the clock to the left of the plant
turn right and go to the clock to the left of* the plant
Pick the clock up
pick the clock* up
Turn right, walk around the bed to the table with the lamp on it
turn right, walk around the bed to the table* with the lamp on it
Hold the clock and turn the light on
hold* the clock and turn the light on
```



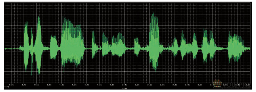
```
Go to the counter in front of you.
go to the counter in front of beu
Pick up the bread on the counter.
pick up the bread on aja counter
Turn around and move forward, then turn left and go to the fridge.
turn around and move forward then turn deft and go to the fridge
```

ESPnet



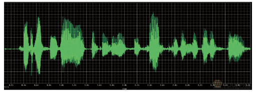
Findings/ Data

SPOON S P UW1 N



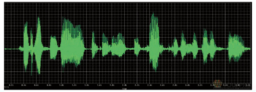
HARPOON HH AA0 R P UW1 N

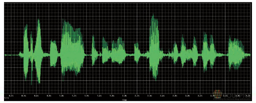
LAMPOON L AE0 M P UW1 N



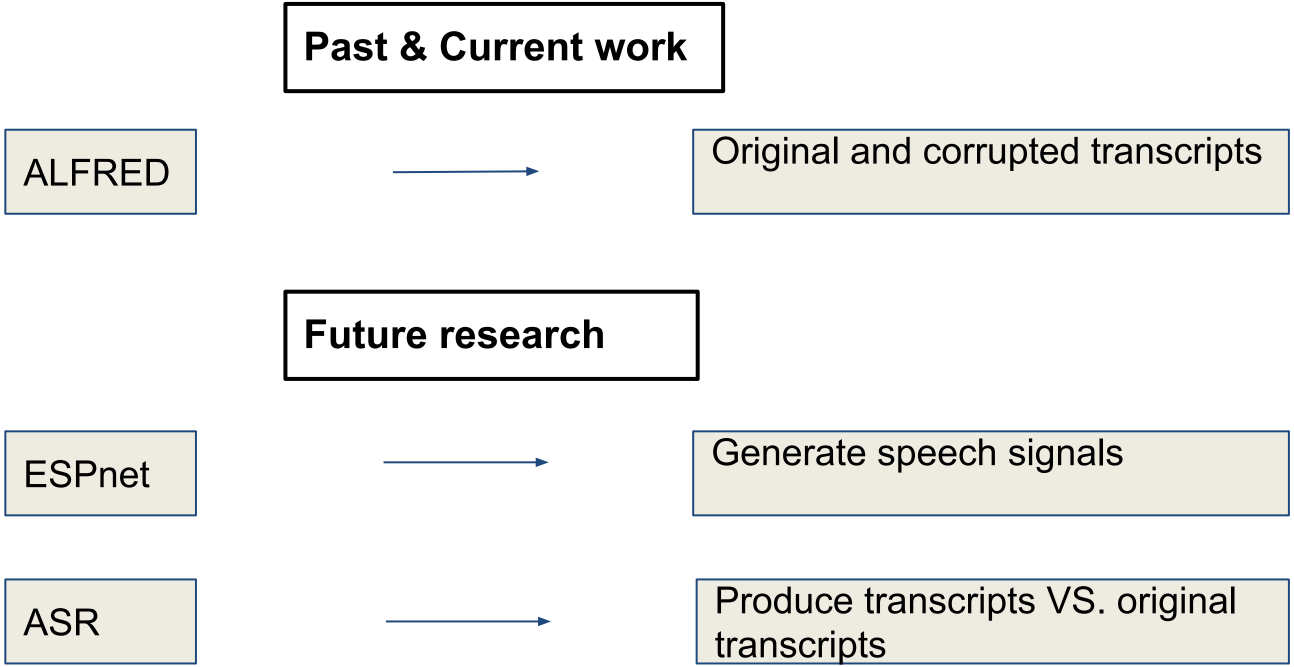
Pick up the spoon on the table

Pick up the soon on the table





Future Research



Past & Current work

ALFRED → Original and corrupted transcripts

Future research

ESPnet → Generate speech signals

ASR → Produce transcripts VS. original transcripts

Discussion/ Conclusion

- Contribution: extracted JSON file & made corruptions on instructions
- Issue: Corrupted instructions can reduce the accuracy of speech recognition
- Goal: Robots/agents recover corrupted instructions
- Approach: Visual information to increase accessibility of speech recognition

References

- *Hayashi, Tomoki, et al. "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit." ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020.*
- *Kolve, Eric, et al. "Ai2-thor: An interactive 3d environment for visual ai." arXiv preprint arXiv:1712.05474 (2017).*
- Krol, Jacob. *Amazon's Astro Home Robot Puts Alexa on Wheels — but Is It Worth $1,000?*, Cnn Underscored, 28 Sep. 2021, https://www.cnn.com/cnn-underscored/electronics/astro-amazon-robot-hands-on.
- *MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. Percept. Psychophys. 24, 253–257. doi: 10.3758/BF03206096*
- Mankad, Sapan. *Audacity: Yet Another Tool for Speech Signal Analysis*, Open Source For You, 2 Nov. 2016, https://www.opensourceforu.com/2016/11/audacity-speech-signal-analysis/.
- *Srinivasan, Tejas, et al. "Multimodal speech recognition with unstructured audio masking." arXiv preprint arXiv:2010.08642 (2020).*
- *Shridhar, Mohit, et al. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.*

Acknowledgements

- Dr. Jesse Thomason
- Dr. Ragusa & Ms. Lilian
- LACC -Economic Development & Workforce Education
- ASSURE -NSF
- Tejas Srinivasan