# Detecting the End of Speaking Turns to Enhance Social Robots' Participation in Group Conversation

Minji Kim — Carnegie Mellon University, Henny Admoni — Research Advisor

## Problem Statement

Research in end-of-turn detection has advanced significantly, and yet studies on its performance beyond in-person pair interactions have not been conducted. Development of a more general system would open up new avenues for HRI practices. For example it could aid a robot in responding more realistically in a group interaction. It could also aid virtual agents from responding at appropriate times as well.

Thus, I aim to investigate whether it's possible to make a real-time model that is generalizable to different interactions such as pair vs. group as well as in-person vs. virtual interactions?

## Background Research

As a speaker is ready to hand off their speaking turn, notable changes occur in various acoustic features, such as frequency contour and intensity level of their voice. Inspired by this fact, acoustic features, typically in conjunction with linguistic features, have been used to train machine learning models to detect end-of-turns in conversation.

Although these end-of-turn models have been successful, they've been tested mainly on in-person pair conversations and usually have limitations which prevent them from operating in real-time. In my research, I aim to develop a more general model that will address these limitations.

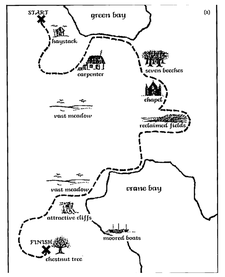
**Micbot Experiment [2]**

Amongst other use cases, this model could be used to manipulate group facilitator robots. Currently most robots in group conversation rely mostly on Wizard of Oz or other limited computation models, but a end-of-turn detection model could provide a more sophisticated means of robot manipulation.

## Dataset Collection

Two datasets were used to evaluate model performance.

- Map Task Corpus: In-person pair interactions where one individual directs another individual to discover a particular route on their map [1].
- Desert Survival Video Corpus: This corpus was developed as part of this project. It consists of recordings over video call of a group of individuals collaborating on the ranking of items according to importance for desert survival.




**Example Map from Map Task**

| Item | My Ranking | Team Ranking |
|---|---|---|
| Flashlight | | |
| Pocket Knife | | |
| Air map of the area | | |
| Plastic raincoat (large size) | | |
| Compass | | |
| Compress Kit with Gauze | | |
| .45 caliber pistol (loaded) | | |
| Parachute (red & white) | | |
| A cosmetic mirror | | |
| A pair of sunglasses per person | | |
| 1 lightweight overcoat per person | | |
| 1 Litre of water per person | | |

**Item List from Desert Survival Task**

## Experimentation and Analysis of Data

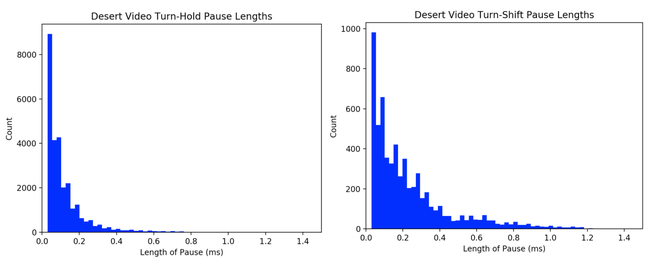
Voice activity was extracted from the Desert Survival audio data using python VAD software. That extracted voice activity was then analyzed for turn-holds versus turn-shifts to use as positive and negative labels for the data respectively. Turn-holds are when the same speaker holds the turn while turn-shifts are when the speaker shifts before and after the pause. The map task data was pre-labeled. Experiment groups within dataset and transferred across datasets were used to evaluate the robustness of the model at detecting end-of-turns in different scenarios.

| Training Set | Testing Set |
|---|---|
| Map Task Training | Map Task Testing |
| Video Training Data | Video Test Data |
| Map Task Complete + Video Training | Video Test |
| Map Task Training + Video Complete | Map Task Testing |

**Experiment Groups for Model Evaluation**

## Pause Model

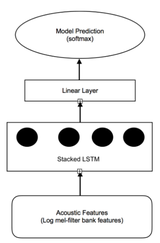
A more simplified model that uses the average between the median pause lengths of a turn-shift/turn-hold as a marker. Any pause longer than that marked duration would classify as a turn-shift while everything else would be classified as a turn-hold.




**Histograms used to analyze pause length in the Map Task corpus**

## LSTM Model

Log mel-filter bank features were extracted at 50 ms intervals from the audio to train a LSTM model. To compensate for an imbalanced dataset, windowed classification was used, classifying segments of 500 ms at a time for containing a turn-hold/turn-shift.




**LSTM structure**

## Evaluation of Models

The pause model evaluation metrics are demonstrated on the right. In general there were about four times as many pause holds than pause-shifts in the data. Unfortunately, the LSTM model was unable to train in general scenarios and did majority classification beyond very small datasets.

| | Map:Map | Video:Video | Map+Video:Video | Map+Video:Map |
|---|---|---|---|---|
| Accuracy | .154 | .215 | .207 | .156 |
| Precision | .423 | .317 | .306 | .425 |
| Recall | .487 | .550 | .573 | .447 |
| F1 Score | .452 | .391 | .390 | .436 |

**Evaluation Metrics for Pause Model**

## Discussion

Judging from the results, there does appear to be some indication that turn-shifts come after longer pause-lengths than turn-holds do. However, the pause model is limited due to the general tendency for both pause-holds and pause-shifts to be more likely to have shorter pauses than longer pauses as seen in the histograms. On another note, fairly equal results are seen across experimental groups. This indicates that the pause model is generalizable to a variety of scenarios: group vs individual and virtual vs. in-person.

## Conclusion

The pause model would benefit from fusion with other models since pause duration doesn't appear to be enough to distinguish between turn-holds versus turn-shifts.

The LSTM model must be further investigated. However, presently it is unclear whether log mel-filter bank features alone are enough to classify turn-holds versus turn-shifts in general scenarios.

## Future Expansions

- Improve the end-of-turn detection model by analyzing more features. Non-verbal features such as gaze and linguistic information could also be integrated into the model.
- The model could be made into a fusion model, which accounts for predictions from the pause model as well as the acoustic model.
- The turn-taking model could be used to aid robots in facilitating group conversations more effectively.

[1] Anderson, A.H., et al. The hcrc map task corpus. *Language and Speech 3* (1991), 351-366.
[2] Tennent, H., et al. Micbot: A peripheral robot object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction* (2019), pp. 133-142.