Learning Backchanneling Behaviors for a Social Robot via Data Augmentation from Human-Human Conversations

Michael Murray[1] Nick Walker[1] Amal Nanavati[1] Patricia Alves-Oliveira[1] Nikita Filippov[1] Allison Sauppe[2] Bilge Mutlu[2] Maya Cakmak[1]

[1] University of Washington, [2] University of Wisconsin–Madison

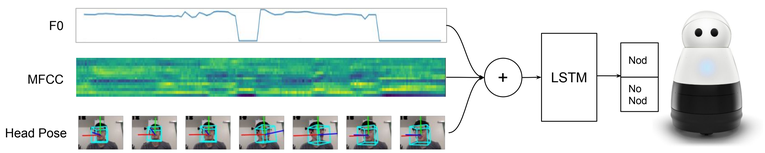
Motivation

The goal of this work is to improve backchanneling behaviors in a robot during a conversation with a human.

Learning based approaches have shown promising results. However, learning backchanneling behavior requires annotated video datasets of human-human conversations which are expensive and time-consuming to collect.

In this work, we present a *data augmentation* scheme to learn more robust backchanneling behaviors for a social robot from human-human conversations.

Method

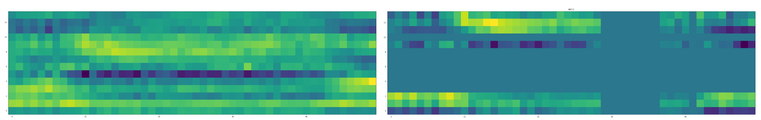


Our model is a recurrent neural network that maps sequences of input from the speaker to output actions on the robot. The input is a combination of three features designed to summarize the speaker's acoustic and visual properties:

- **Fundamental frequency (F0)** - an estimation of the speaker's pitch
- **Mel-frequency cepstrum coefficients (MFCC)** - represents the short-term power spectrum of the sound
- **Head pose** - used as a summary visual feature informed by work showing gaze as a cue for backchannels

We aim to enable accurate prediction while maintaining the feasibility of deploying the model on computationally constrained robotics platforms.

Data Augmentation



We augment the features by warping them across time and masking blocks of utterances over time. An example of these augmentations applied to the MFCC feature can be seen in the above figure. These augmentations are intended to improve the model's robustness to partial loss of information (frequency, segments of speech, or head pose information) and deformations across time.

Data Collection

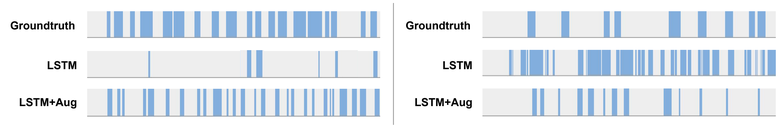
Using a custom peer-to-peer video chat tool, we recorded audio and video of conversational interactions between two people. Each interaction was structured around a prompt, with one person tasked to respond to the prompt while the other person was listening and performing natural backchannels.

Evaluation

Method	Training (Seen Speakers) Acc.	Prec.	Rec.	F_1	Validation (Unseen Speakers) Acc.	Prec.	Rec.	F_1
Rule-based	-	-	-	-	0.60	0.20	0.14	0.23
LSTM	0.80	**0.31**	**0.29**	**0.60**	0.62	0.18	0.16	0.24
LSTM + Augmentation	0.80	0.29	0.25	0.57	**0.64**	**0.23**	**0.24**	**0.29**

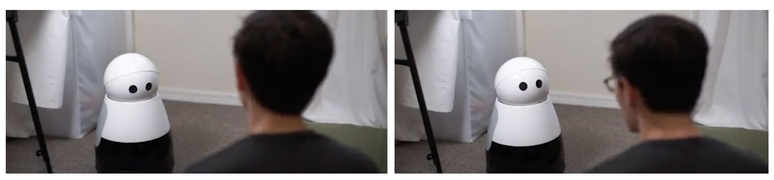
Our evaluation on the dataset shows both LSTM based approaches outperforming the rule based approach. Using data augmentation achieves highest performance on the validation set due to less overfitting to the training set. Reported numbers are averages across all validation folds.

Qualitative Results



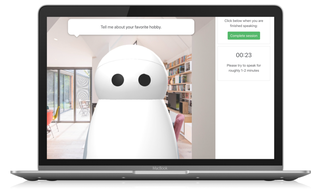
These visual timelines indicate nods over the duration of two example sessions. On the left, we can see that the LSTM nods too infrequently and the more robust data augmentation method is closer to the ground truth. While on the right, the LSTM nods too frequently.

On-Robot Demonstration



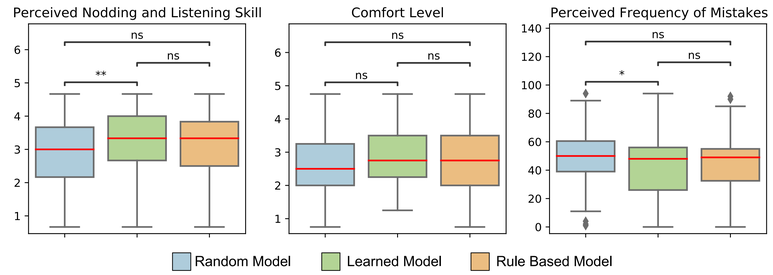
Our user evaluation was done with a virtual robot to adapt to the COVID-19 pandemic. However, we also demonstrated that our approach can be deployed on a physical Mayfield Kuri robot.

User Study

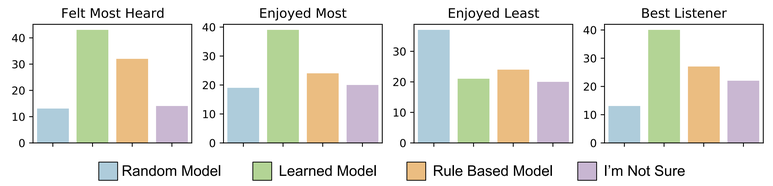


Our study involved participants talking to a virtual Mayfield Kuri robot that uses either a learned model, a rule-based model, or a random baseline to determine when to nod. We used a virtual robot to adapt to the COVID-19 pandemic. Our study was completed by **102 participants**.

Likert Scale Responses



Forced Choice Responses



Open Ended Responses

Some participants noticed the timing of nods, e.g., one said the learned model appeared "***better programmed to nod at more appropriate intervals***";

Others answered based on how they felt after the interaction; e.g., comments about the learned model included that it "***seemed natural and normal***", "***felt like he was really listening***", and "***seemed more life like.***".

Some users expressed optimism about the utility of a backchanneling robot; e.g. one user said "***I would use this to practice job interviews***" and another suggested to "***make these robots some kind of therapy app***".

Other participants felt that the experience of talking to the robot was negative, highlighting important difficulties of social robot design; e.g. one person stated that "***Speaking to the robot felt dehumanizing.***".