Analysis and Detection of Differences in Spoken User Behaviors between Autonomous and Wizard-of-Oz Systems

Mikey Elmers, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan

Background

Conversational Robots

- **Growing** presence in daily life
- **Gap** between conversational robots and human-like interaction

User Spoken Behaviors

- **Fillers** “ano”/“eto” (Japanese) and “uh”/“um” (English)
- **Backchannels** “hai”/“un” (Japanese) and “mhmm”/“uh-huh” (English)
- **Disfluencies** lengthening, truncation, repair, word fragmentation
- **Laughter**

Research Question

How do user spoken behaviors differ when interacting with an autonomous system versus a Wizard-of-Oz (WoZ) system?

Descriptive & Inferential Statistics

Attentive Listening

	Autonomous	WoZ	p-value
Length	9.91 (8.95)	10.63 (9.75)	<0.001
Speaking Rate	6.04 (2.27)	6.56 (2.47)	<0.001
Fillers/second	0.26 (0.67)	0.32 (0.88)	<0.001
Backchannels/second	0.34 (1.09)	0.42 (1.29)	<0.001
Disfluencies/second	0.11 (0.67)	0.10 (0.79)	<0.001
Laughs/second	0.04 (0.32)	0.06 (0.41)	<0.001
Filler Count	26.88%	30.03%	<0.001
Backchannel Count	10.69%	11.77%	<0.001
Disfluency Count	8.12%	6.69%	<0.001
Laugh Count	2.40%	4.00%	<0.001

Job Interview

	Autonomous	WoZ	p-value
Length	14.00 (13.33)	11.46 (11.27)	<0.001
Speaking Rate	7.29 (2.50)	7.77 (2.91)	<0.001
Fillers/second	0.48 (0.95)	0.46 (1.56)	<0.001
Backchannels/second	0.39 (1.27)	0.87 (2.03)	<0.001
Disfluencies/second	0.07 (0.63)	0.09 (0.88)	>0.05
Laughs/second	0.01 (0.13)	0.03 (0.29)	<0.001
Filler Count	46.1%	30.1%	<0.001
Backchannel Count	9.51%	17.10%	<0.001
Disfluency Count	6.49%	5.80%	>0.05
Laugh Count	0.49%	2.51%	<0.001

Table 1. Feature comparison for each scenario + condition combination. Mean (Standard Deviation) in the first two columns and p-value in the third column.

Variable Importance

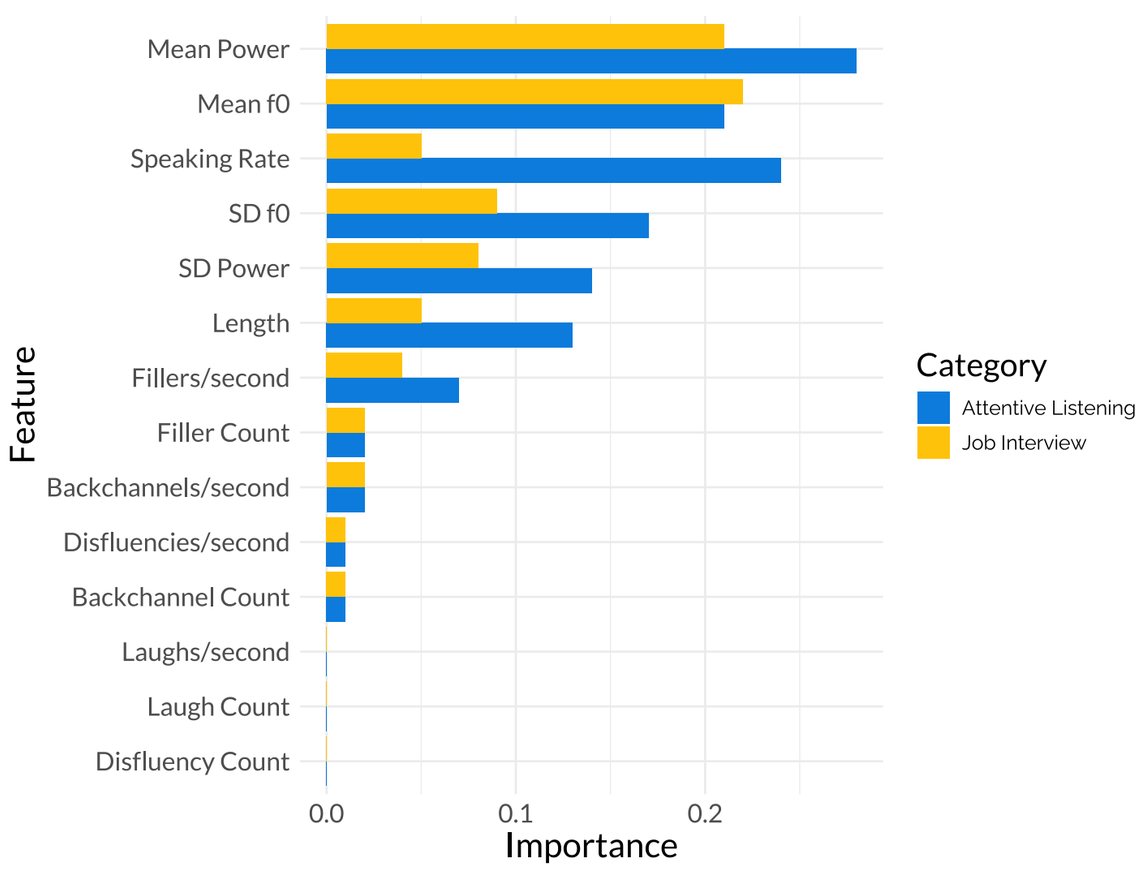

Figure 2. Permutation-based variable importance analysis for the random forest models.

Method

Experiment

- **Users** interacted with ERICA in Japanese
- **Evaluated** user spoken material (i.e., not system responses)
- **Scenarios** attentive listening and job interview
- **Conditions** autonomous and WoZ

Transcription

- **Inter-pausal unit (IPU)** pause > 200 ms
- **Linguistic** count and #/second for fillers, backchannels, disfluencies, and laughter
- **Length** # of tokenized characters per IPU
- **Speaking rate** IPU length / IPU duration

Figure 1. Illustration of experimental setup. The top frame is a side profile view of a subject (left side) and ERICA (right side). The bottom-left and bottom-right frames depict the operator’s activity during the WoZ condition.

Model Evaluation Metrics

Attentive Listening

Model	Accuracy	Precision	Recall	F1
Baseline	0.64	0.64	**1.00**	0.78
Logistic Regression	0.66	0.69	0.86	0.76
Support Vector Machine	**0.71**	0.73	0.87	**0.79**
Random Forest	0.70	**0.74**	0.81	0.77

Job Interview

Model	Accuracy	Precision	Recall	F1
Baseline	0.51	0.49	**1.00**	0.66
Logistic Regression	0.55	0.54	0.53	0.54
Support Vector Machine	0.67	0.66	0.68	0.67
Random Forest	**0.69**	**0.69**	0.67	**0.68**

Table 2. Prediction if user is interacting with autonomous or WoZ system. Binary classification task with linguistic and acoustic features (14 in total) as input and autonomous/WoZ prediction as the output. Baseline model predicts majority class for all instances.

Conclusion

Limitations

- **Data** collection over years
- **Topic** differences
- **Immediate** context
- **Evaluation** did not include age, gender, or multi-modal cues

Summary

- **User’s spoken behaviors** differ between autonomous and WoZ systems
- **Acoustic** + **linguistic** predictive model performed best