Prediction of Ice-breaking Between Participants Using Prosodic Features in the First Meeting Dialogue

Hirofumi Inaguma
Kyoto University
School of Informatics
Kyoto, Japan
inaguma@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Kyoto University
School of Informatics
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Shizuka Nakamura
Kyoto University
School of Informatics
Kyoto, Japan
shizuka@sap.ist.i.kyoto-u.ac.jp

Katsuya Takanashi
Kyoto University
School of Informatics
Kyoto, Japan
takanasi@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto University
School of Informatics
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

In the human-human first meeting dialogue, people tend to have a chat before their main topics to break tension or the "ice." This phenomenon is called "ice-breaking." For realizing this kind of natural conversations in dialogue systems, we address prediction of ice-breaking using prosodic features in dialogue. This will allow for the systems to change conversation topics smoothly. At first, we statistically analyze relationships between prosodic features and ice-breaking events, to select the useful feature sets showing significant effects. Then, prediction of ice-breaking is conducted by a logistic regression model with these features, which shows a promising result.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI);

Keywords

Ice-breaking; prosodic features; dialogue

1. INTRODUCTION

Recently, various chat-like dialogue systems using a human-like robot have been developed for the purpose of natural conversation with humans [11]. In this kind of dialogue systems, users are expected to meet a robot for the first time. In the human-human first meeting, people often consider a flow of conversation and have a chat before their main topics to relieve a feeling of tension. This phenomenon, which is called "ice-breaking," [10] is expected to relax them. In order to realize this kind of natural conversation, it is necessary to design a series of flow of conversation in dialogue systems. However, there are few previous studies focusing on ice-breaking in dialogue systems. By predicting an occurrence of ice-breaking, it is expected for the dialogue system to change conversation topics naturally and smoothly as we do because it can know when to change topics. Moreover, by generating ice-breaking at the right timing, it is also expected to increase a feeling of rapport toward the dialogue system.

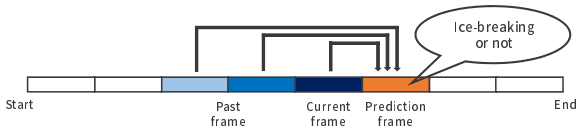


Figure 1: Prediction the occurrence of ice-breaking

This study investigates how prosodic features in dialogues affect the occurrence of ice-breaking. Various kinds of prosodic features are computed for the different time-windows. Then, the changes in their values among the time-windows are statistically analyzed by t-test. Prosodic features used for the analysis are related with utterance features, their overlaps, silence, F0, power, and speech rate. We focus on laughter in a conversation, and the event of ice-breaking is defined by an occurrence of a shared laughter [7, 2, 5, 6] by two participants. The occurrence of ice-breaking is predicted by a logistic regression model using the above-mentioned prosodic features (See Figure 1).

The remaining part of this paper is organized as follows. Section 2 describes dialogue data. In Section 3, the event of ice-breaking is defined and how to extract features is described. In Section 4, relationships between prosodic features and the occurrence of ice-breaking is statistically analyzed. Finally, the prediction of ice-breaking is conducted in Section 5, and conclusions are described in Section 6.

ASSP4MI'16, November 16 2016, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3005467.3005472

Figure 2: A snapshot of dialogue recording: a subject (left) and the android ERICA (right) which is remotely-operated by a human.

2. DIALOGUE DATA

2.1 Android ERICA

Dialogue data were recorded with an android ERICA [8] as shown in Figure 2. ERICA was remotely operated by an amateur actress. There are a lot of actuators on ERICA's body and ERICA can generate various facial expressions and multi-modal behaviors such as gazing and nodding.

2.2 Data Collection

Fifteen sessions of the first meeting with ERICA were recorded. Subjects were fifteen university students (three females and twelve males). They entered a room one by one and talked freely with ERICA, . Each session consisted of two phases. The former part was a chat about their daily lives and hobbies[1] (phase 1), and the latter part was focused on the main topic on an android like ERICA[2] (phase 2). The operator was instructed to take about three minutes for phase 1 and about seven minutes for phase 2. The actual timing to change phases was determined by the operator.

2.3 Annotation

For the purpose of analyzing conversation before a main topic, phase 1 in each session was focused in this study. Dialogue data were annotated with ELAN[3] [12]. Three kinds of labels were anotated as follows: 'utterance,' 'topic,' and 'laughter'. Each utterance was segmented based on pauses (longer than 200 msec). Each laugher was confirmed to include laughing voices and facial expressions.

3. ICE-BREAKING AND FEATURE EXTRACTION

3.1 Definition of Ice-breaking

Ice-breaking can occur when both participants get relaxed. Apparently, they are relaxed when they laugh together. Therefore, laughing together can indicate the occurrence of ice-breaking. This kind of laughing is called "shared laughter." [7, 2, 5, 6] So, in this study, we focus on laughter in a conversation, and the event of ice-breaking is defined by an occurrence of a shared laugher by two participants. As shown in Figure 3, one or more shared laughter occurred in each session and almost all of them are distributed in the topic termination. This is consistent with previous studies [7, 2, 5]. Thus, it is reasoned that ice-breaking is an indicator of a topic change. On the other hand, the participants did not always move into the main topic after the occurrence of ice-breaking. The timing of the occurrence of ice-breaking was different depending on subjects, and even when ice-breaking occurred in the beginning of conversation, they kept chatting for a while after ice-breaking, and there were some other shared laughters before a main topic.

[1]Topics (phase 1): greeting, self-introduction, hobbies, dream, academics, club, circle, part-time jobs, hometown, skills, living etc.

[2]Topics (phase 2): knowledge about her, impression of her, a sense of discomfort to her, how her body is moved etc.

[3]https://tla.mpi.nl/tools/tla-tools/elan/

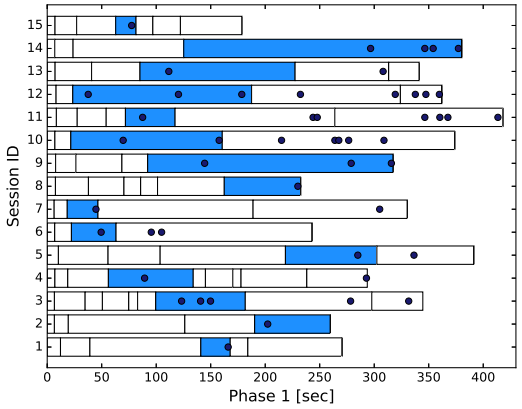


Figure 3: The distribution of topic and shared laughter in each session: the elapsed time (X axis) and the ID of each session (Y axis). Each point represents the occurrence of shared laughter. Colored bars represent topics with the first ice-breaking.

Table 1: Classification of prosodic features

	Feature[Unit]
Utterance features	Num[num/sec], Dur[sec], Freq
Overlaps	Num[num/sec], Dur[sec], Freq
Silence	Num[num/sec], Dur[sec]
F0	Mean[Hz], SD, Max[Hz], Min[Hz], Range[Hz]
Power	Mean[dB], SD, Max[dB], Min[dB], Range[dB]
Speech Rate	Mean[mora/sec]

Num: number, Dur: total duration, Freq: frequency,
SD: standard deviation, Range: mean range

3.2 Feature Extraction

Wrede et al. [13] reported relationships between prosodic features and high interest level in dialogue, which was called "hot spot." Getica-Prez et al. [4] tried to detect a heightened interest of the whole group in the meeting with prosodic and visual information. In addition, Bonin et al. [3] analyzed relationships between topic change and social signals, which were composed of laughter, overlaps, silence, and nodding. Luz et al. [9] tried to detect topic boundaries with overlaps and silence. Moreover, there were some previous studies analyzing relationships between topic boundaries and shared laughter [7, 2, 5]. Therefore, prosodic features can also correlate with the occurrence of ice-breaking and it is expected that prosodic features can be indicators to predict the occurrence of ice-breaking.

There are many features such as linguistic information, back-channels, facial expressions, gazing, nodding, posture, and gesture. Unlike these features, prosodic features are uniquely identified and can be easily measured in real time. Thus, we focus on them in this study. Prosodic features are categorized into six groups shown in Table 1. Here, it is presumed that there were bias in acoustic features such as F0, power, and speech rate because each speaker had their own inherent speech style. So, these features are normalized for each speaker to compensate the effects of individuality. Methods of computing these prosodic features are described as follows.

Table 2: Results of t-test ($*$ p-value$<$0.05, $**$ p-value$<$0.01, $***$ p-value$<$0.001)

	Feature	First time-window vs. First ice-breaking			First time-window vs. Last time-window		
		20 sec	30 sec	40 sec	20 sec	30 sec	40 sec
Utterance features	Num(R/H)	0.474 /0.044*	0.208 /0.005**	0.019* /0.195	0.159 /0.005**	0.914 /0.028*	0.970 /0.079
	Dur(R/H)	0.193 /0.000***	0.000***/0.002**	0.000***/0.000***	0.008**/0.006**	0.001**/0.000***	0.004**/0.005**
	Freq	0.034*	0.002**	0.005**	0.115	0.144	0.163
Overlaps	Num	0.495	0.304	0.901	0.217	0.468	0.314
	Dur	0.529	0.527	0.904	0.423	0.238	0.171
	Freq	0.454	0.408	0.691	0.319	0.480	0.241
Silence	Num	0.318	0.004**	0.503	0.960	0.932	0.232
	Dur	0.401	0.226	0.567	0.002**	0.074	0.154
F0	Mean(R/H)	0.008**/0.294	0.003**/0.911	0.301 /0.948	0.023*/0.334	0.090 /0.333	0.175 /0.392
	SD(R/H)	0.500 /0.014*	0.072 /0.099	0.054 /0.200	0.270 /0.021*	0.022*/0.027*	0.329 /0.026*
	Max(R/H)	0.411 /0.617	0.252 /0.677	0.879 /0.929	0.261 /0.216	0.519 /0.265	0.216 /0.384
	Min(R/H)	0.223 /0.483	0.443 /0.620	0.087 /0.039*	0.181 /0.543	0.232 /0.735	0.101 /0.031*
	Range(R/H)	0.105 /0.550	0.003**/0.142	0.013*/0.826	0.150 /0.958	0.110 /0.408	0.567 /0.105
Power	Mean(R/H)	0.001**/0.773	0.004**/0.799	0.053 /0.473	0.003** /0.333	0.041* /0.333	0.065 /0.098
	SD(R/H)	0.060 /0.714	0.014* /0.878	0.024*/0.692	0.203 /0.994	0.020* /0.308	0.065 /0.025*
	Max(R/H)	0.479 /0.787	0.436 /0.487	0.613 /0.288	0.144 /0.428	0.653 /0.597	0.536 /0.637
	Min(R/H)	0.004**/0.028*	0.164 /0.240	0.048*/0.841	0.093 /0.366	0.337 /0.393	0.975 /0.051
	Range(R/H)	0.287 /0.438	0.001**/0.215	0.036*/0.046*	0.000***/0.821	0.000***/0.154	0.000***/0.006**
Speech Rate	Mean(R/H)	0.001**/0.858	0.001**/0.641	0.020*/0.497	0.972 /0.946	0.000***/0.946	0.386 /0.438

R: Robot WOZ, H: Human Subject

Utterance features

The number, total duration, frequency of utterances is computed in each time-window. The number and total duration are normalized by the length of each analysis time-window. The frequency is the number of utterances of each speaker divided by the total number of utterances of the two speakers in each time-window.

Overlap features

Overlap is defined as an overlapping part of utterances by the two participants longer than 100 msec. The number, total duration, frequency of overlaps is computed in each time-window. The number and total duration are normalized by the length of each analysis time-window. The frequency is the number of overlaps divided by the total number of utterances of the two speakers in each time-window.

Silence features

Silence is defined as a silent part longer than 200 msec. The number and total duration of silence parts are computed in each time-window and both of them are normalized by the length of each analysis time-window.

F0 and Power

Both F0 and power are extracted by using speech analysis software Praat[4] [1] at 50 msec intervals, and they are assigned into each corresponding utterance. Then, the mean and range (difference between maximum and minimum) of these values in each utterance are computed. Moreover, we compute the mean, standard deviation, maximum, and minimum of mean values, and mean of range values in each analysis time-window. Note that we use values of logarithmic F0 and power.

Speech Rate

Speech rate in each utterance is computed as the number of the mora divided by the duration of the utterance. Furthermore, the mean of this value is computed in each analysis time-window. Note that we exclude utterances consisted of only laughing voice.

[4]http://www.fon.hum.uva.nl/praat/

4. RELATIONSHIPS BETWEEN PROSODIC FEATURES AND ICE-BREAKING

4.1 Analysis Method

As the analysis unit, two units can be considered: topic and the fixed time time-window. We would like to predict the occurrence of ice-breaking in real time, but it is difficult to compute the topic unit in real time because topic lengths are different. Therefore, to predict the occurrence of ice-breaking in real time, it is suitable to compute prosodic features for the fixed time-windows. Then, the shorter the length of each time-window is, the smaller the number of utterances included in each time-window is. On the other hand, the longer the length of each time-window is, the less significant the changes of values between time-windows can be. Considering this trade-off, the fixed time-windows of different lengths are designed as follows: 20, 30, and 40 sec.

Relationships between the prosodic features and the occurrence of ice-breaking is statistically analyzed by t-test. At first, each session is divided into the analysis time-windows (20, 30, and 40 sec), and the features in Table 1 are computed in each time-window. In order to investigate how these features in dialogue affect the occurrence of ice-breaking, a paired t-test is conducted between the first time-window and the time-window where ice-breaking occurred for the first time, and between the first and last time-windows. Note that in case that an utterance crosses a border of a time-window, it was divided by the border. If the length of the last time-window was less than 20% of that of the other time-windows, it was excluded from the analysis.

4.2 Results

As shown in Table 2, among the listed features, statistically significant differences were observed in the utterance features, F0 (excluding Max and Min), power (excluding Max and Min), and speech rate for any length of the analysis time-windows. With regard to the utterance features, it is known that both the number and frequency of utterances have a relationship with hot spots in meetings [13]. These results are consistent with this. Furthermore, the acoustic features such as F0, power, and speech rate mostly of a robot WOZ were useful. For the purpose of analyzing the role of laughter, Gupta et al. [6] classified utterances in interviewing sessions between a client and a counselor to five groups: utterance with no laughter, utterance with solo client laughter, utterance with

solo counselor laughter, utterance with client-lead shared laughter, utterance with counselor-lead shared laughter. Likewise, in this study, we classify utterances with laughter to four groups and count them (see Table 3). Results show that the number of robot-lead shared laughter is more than that of human-lead, and the number of solo human laughter is more than that of robot. It can be said that this is caused by operator's attitude of trying to spark conversation among participants. On the other hand, among the different lengths of the analysis time-windows, we observed the most useful features in 30 sec. time-window. So, in the following section, we adopt 30 sec. time-window as the length of analysis time-window and try to predict the occurrence of ice-breaking.

Table 3: Classification of laughter

	So-R	So-H	Sh-R	Sh-H
Total	73	189	40	9

So-R: Solo robot laughter
So-H: Solo human laughter
Sh-R: Robot-lead shared laughter
Sh-H: Human-lead shared laughter

5. PREDICTION OF ICE-BREAKING

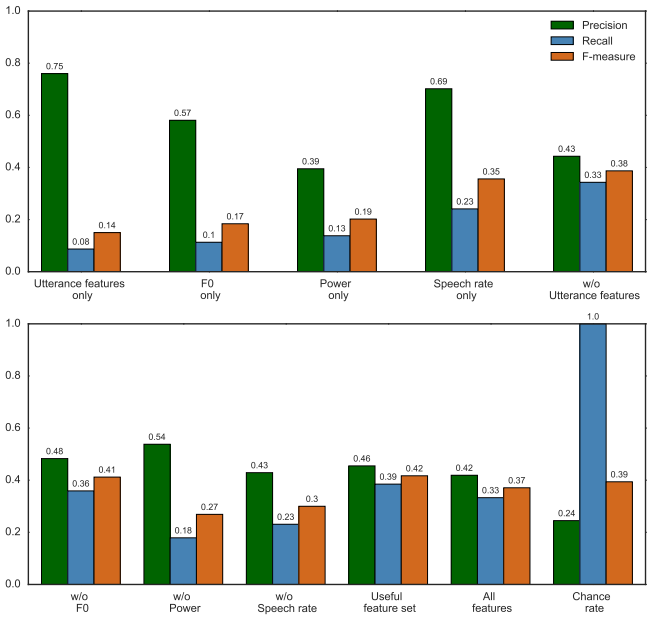
5.1 Experimental Setting

All or a part of the prosodic features showing significant differences in the previous statistical analysis in Section 4.2 are used to predict the occurrence of ice-breaking. In this study, we formulate the prediction task as a classification of a degree of relaxation, which is defined as 1 when ice-breaking occurred and otherwise 0, for each time-window. Prediction is conducted by a logistic regression model. Inputs for the model are the feature vectors computed in each analysis time-window, gradient ($\mathbf{\Delta}$) vectors of the features in the preceding two time-windows, and the elapsed time. Prediction output is a degree of relaxation (ice-breaking or not). For evaluation measures, we adopt precision, recall, and F-measure of prediction against the annotation of the time-windows. Five-fold cross validation is carried out, using the 15 dialogue sessions.

5.2 Results

In the previous statistical analysis, the following features showed significant effects (hereafter, useful feature set): utterance features, F0 (excluding Max and Min), power (excluding Max and Min), and speech rate. So, in order to investigate which feature is important to predict the occurrence of ice-breaking, the feature sets for prediction were categorized as follows: the useful feature set, a single feature in the useful feature set, the useful feature set excluding one feature, all features (adding overlap and silence), and the chance rate (in the case of prediction that ice-breaking occurs in all time-windows).

The prediction results are shown in Figure 4. When using a single feature, both recall and F-measure were low although precision was relatively high, so prediction was not accurate at all. On the other hand, by using multiple features, relatively high accuracies were obtained. F-measure got the maximum when using the useful feature set (F-measure: 0.42). In the case of using all features and excluding one feature, F-measure decreased. In addition, by using the useful feature set, accuracies were improved compared to the chance rate (F-measure: 0.39). Therefore, it is concluded that we can predict ice-breaking reasonably by using prosodic features and components in the useful feature set are important for prediction. This is consistent to the results of analysis by t-test in Section 4.2.



Useful feature set: { Utterance features, F0, Power, Speech rate }
All features: Useful feature set + {Overlaps, Silence }
Chance rate: In the case of prediction that ice-breaking occurs in all time-windows

Figure 4: Prediction accuracy (precision, recall, and F-measure) with various feature sets (30 sec. time-window)

6. CONCLUSIONS

For predicting the occurrence of ice-breaking in the first meeting dialogue, relationships between prosodic features and ice-breaking events were statistically analyzed. As results of t-test, we found the following useful features showing significant effects: utterance features (the number, total duration, the frequency of utterances), F0, power, and speech rate. Then, we conducted machine learning by a logistic regression model and confirmed that F-measure got the highest value when using these features, which was better than chance rate (in the case of prediction that ice-breaking occurs in all time-windows). Furthermore, when adding or excluding some features, F-measure decreased. Therefore, we conclude that we can predict ice-breaking reasonably by using these significant features and these are important for prediction. In the future, we plan to increase the amount of data for analysis. Moreover, we will investigate the effects of multi-modal behaviors such as facial expressions.

7. ACKNOWLEDGMENTS

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

8. REFERENCES

[1] P. Boersma and D. Weenink. Praat: doing phonetics by computer. 2010.

[2] F. Bonin, N. Campbell, and C. Vogel. Laughter and topic changes: Temporal distribution and information flow. In *Proceedings of Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 53–58, 2012.

[3] F. Bonin, N. Campbell, and C. Vogel. The discourse value of social signals at topic change moments. In *Proceedings of INTERSPEECH*, 2015.

[4] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *Proceedings of ICASSP*, volume 1, pages 489–492, 2005.

[5] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Laughter and topic transition in multiparty conversation. In *Proceedings of SIGDIAL*, pages 304–308, 2013.

[6] R. Gupta, T. Chaspari, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan. Analysis and modeling of the role of laughter in motivational interviewing based psychotherapy conversations. In *Proceedings of Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] E. Holt. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics*, 42(6):1513–1525, 2010.

[8] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara. Talking with erica, an autonomous android. In *Proceedings of SIGDIAL*, 2016.

[9] S. Luz and J. Su. The relevance of timing, pauses and overlaps in dialogues: detecting topic changes in scenario based meetings. In *Proceedings of INTERSPEECH*, pages 1369–1372, 2010.

[10] Y. Rogers and H. Brignull. Subtle ice-breaking: encouraging socializing and interaction around a large public display. In *Workshop on Public, Community. and Situated Displays*, 2002.

[11] A. Vinciarelli, A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal, et al. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Journal of Cognitive Computation*, 7(4):397–413, 2015.

[12] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, page 5th, 2006.

[13] B. Wrede and E. Shriberg. Spotting “ hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of EUROSP*, pages 2805–2808, 2003.