1F2-2

会話中に表出する言語・非言語情報のデータマイニングに基づく説明会話の客観的評価指標の獲得

Analysis of relationship among the effect of narrative, verbal cues and nonverbal cues in conversation

米航[*1] 岡田将吾[*1] 新田克己[*1]
MI HANG Shogo Okada Katsumi NITTA

[*1] 東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

This paper addresses the analysis of the relationship among the effect of narrative, verbal cues and nonverbal cues in face-to-face conversation. Towards this goal we define verbal and nonverbal cues by aggregating automatically extracted cues at the individual and dyadic levels. Four coders from the third party evaluate whether the narration was understandable by questionnaire. Our results show that the group speaking length and feedback of listeners (number of interjections and nodding times) have significant correlations with the third party evaluation of narrative effect in face-to-face conversation.

1. Introduction

It's well known that good communication is the foundation of any successful relationship. Face-to-face conversation is a fundamental social interaction. The automatic analysis of face-to-face conversational focuses on developing computational systems and sensor technology that can automatically analyze human conversational behavior by observing via sensing devices such as cameras and microphones [1]. The aim of the automatic analysis of face-to-face conversational is to infer and possibly predict aspects of the underlying social context, including both individual attributes and interactions with other people in the group.

Narrative has existed in every known society. Our goal is to automatically analyze the relationship among the effect of narrative, verbal cues and nonverbal cues in face-to-face conversation. In order to address this question, we extract verbal and nonverbal cues of participants from a narrative task. Then we ask the third party to give evaluation scores of narrative task. Our data is from Cartoon Narrative Task which is held by Okada group[2]. In this task, a group is composed of three unacquainted participants, where two participants, who have watched cartoon video, explain it to the other participant [2].

2. Related Work

The automatic analysis of conversational is a fundamental area in social psychology and nonverbal communication.

In a conversation, an addressee is the person at whom the speech is directed [3]. In social psychology, it is known that the addressing phenomenon occurs through different communication channels, including speech, gaze, and gesture, e.g. listeners manifest attention by orienting their gaze to speakers, who in turn use gaze to indicate whom they address, and to ensure visual attention from addressees to hold the floor [4].A good part of the body of work on automatic analysis of head pose as a surrogate for gaze and of visual focus of attention (VFOA) in group conversations [5] could be applied towards the automatic identification of addressees in multi-party cases. In brief, the goals of the existing works in addressing are to identify what participants in a conversation the current speaker is talking to, and to explore the connections between addressee modeling and other conversational activities.

In our work, using the evaluation scores of narrative task made by the third party to automatically analyze the relationship among the effect of narrative, verbal cues and nonverbal cues in face-to-face conversation is an unexplored problem.

3. Narrative Dataset

We use 4 group data from the Cartoon Narrative Task. [2].This task has changed the setting of a dyadic narrative interaction designed by McNeill [6], in McNeill's research, a participant is asked to narrate from memory a cartoon story to a participant. The name of the cartoon story is "Canary Row", and this story has been used for gesture analysis in narrative tasks [6].In this task, a group is composed of three unacquainted participants, where two participants (A and B), who have watched the video, explain it to the other participant(C). The three participants have never watched the video or listened to the story.24 women participants aged between 20 and 25 years was recruited to collect the dataset. 8 sessions dataset has been collected in cooperation with these participants. Average time length of recorded datasets is 11 minutes (total is 700 minutes). Both manual annotation and autonomous annotation have been used to get the dataset and primitive nonverbal patterns (speech, gesture, head gesture, and head direction) from every participant has been annotated as binary on/off or three variables by using some pattern recognition techniques [2].

Contact: MI HANG, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology. 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, 226-8503, JAPAN. Tel:045-924-5205 , Fax:045-924-5218,E-mail:mi_h@ntt.dis.titech.ac.jp

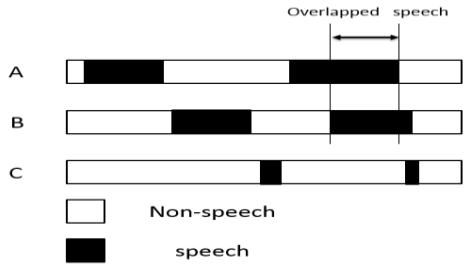

Figure 1 : Speaking turn audio nonverbal features

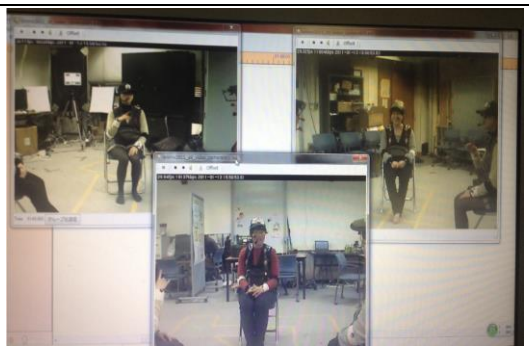

Figure 2 : The video of Cartoon Narrative Task

4. Cues Extraction and Third Party Evaluation Data

In our work, cartoon "Canary Row" was divided into 9 scenes according to different scenarios cartoon. We extract cues and get them scene by scene in every group.

4.1 Verbal Cues

While spoken language constitutes a very strong communication channel in group conversations, it is known that a wealth of information is conveyed nonverbally in parallel to the spoken words.

By using morphological analysis tool Chasen[7],we can get words cues and interjections cues in each scene by analyzing the text of participants' speak contents which are defined below:

Number of Words (*NW*): Cumulates the total number of four kind of words which are noun, verb, adjective and adverb spoken by two narrators.

Number of Interjections (*NI*): Cumulates the total number of interjections spoken by listener.

Number of Filler (*NF*): Cumulates the total number of filler spoken by two narrators.

4.2 Nonverbal Cues

Nonverbal signals include features that are perceived aurally – through tone of voice and prosody and visually – through body gestures and posture, eye gaze, and facial expressions.

From the speech segmentation(Figure1), we compute speaking length cues, overlap cues and turn-taking cues in each scene defined below:

Group Speaking Length (*GSL*): Cumulates the total time that two narrator speak according to their binary speaking status in each scene.

Overlap Length (*OL*): Cumulates the total time of overlap between two narrators A and B when they narrated to listener C in each scene.

Turn-taking Length (*TL*): We compute total time of turn-taking between two narrators in each scene.

From the gaze and nod segmentation, we compute gazing cues and nodding cues.

Conjugate Gaze Length (*CGL*): Conjugate Gaze length means the total time of narrators look at listener at the same time.

Nodding Times (*NT*): We define the times of nodding from listener as Nodding Times.

Here we explain the above definition of verbal and nonverbal cues in detail. Taking feature *NW* for an example.

Let define *NW* as below:

$$NW = \left\{ NW_{gs} \right\} \left(1 \le g \le 4, 1 \le s \le 9 \right)$$

where NW_{gs} denote the number of words spoken by two narrators in the scene S of group g .The feature *NW* is a 36-dimensional vector.

Therefore, the other features(*NI*,*NF*,*GSL*,*OL*,*TL*,*CGL*,*NT*) have the same definition as *NW* mentioned above .

4.3 Nonverbal Cues

We asked 4 coders (we code them with 1,2,3,4)as third party to make ten-grade evaluation about narrative task scene by scene in every session by watching narrative task videos which are captured by cameras (Figure 2).These coders are asked to watch the cartoon video firstly before they watch the Narrative Task video. These coders are Chinese students who are studying their master course in Japan and almost have the same level of Japanese to understand the speech in the Narrative Task video. They are asked to score every scene according the effect of narrative speech in the video. Let define ES_{pgi} as the evaluation scores of scene i in group g made by student p.(p is the number of students, g means the number of group and i is the number of scene. $1 \le p \le 4, 1 \le g \le 4, 1 \le i \le 9$).For example, ES_{215} means the evaluation score of scene 5 in group 1 made by coder 2. Then we define $MES_{gi} = \sum_{p=1}^{4} ES_{pgi} \Big/ 4$ as the mean of the evaluation scores for each scene in the same group by 4 students. Finally, we denote

$$MES_g = \left\{ MES_{g1}, MES_{g2} \cdots\cdots MES_{g9} \right\} \left(1 \le g \le 4 \right)$$

$$MES = \left\{ MES_1, MES_2, MES_3, MES_4 \right\}$$

From the above mentioned, we know that *MES* is a 36-dimensional vector.

5. Correlation Analysis

In this section, we study the correlation among the verbal, nonverbal cues and *MES*.There are several correlation coefficients, often denoted p or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a

Table 1. Correlation between MSE and verbal , nonverbal features

		r	p
MES	Verbal Cues		
	NW	**0.52**	**0.021**
	NI	**0.57**	**0.002**
	NF	0.18	0.043
	Nonverbal Cues		
	GSL	**0.58**	**0.014**
	OL	0.34	0.033
	TL	0.38	0.017
	CGL	0.29	0.038
	NT	**0.54**	**0.007**

nonlinear function of the other). As Table 1 showed, we compute the Pearson correlation coefficient, hence reporting the r and p values.

Number of words(*NW*) had a correlation of 0.52 with *MES*. We can also find that Group Speaking Length(*GSL*) had a correlation of 0.58 with *MES*. This indicates enough speech is needed to guarantee the listener to understand the outline of the story .Number of filler (*NF*) had a correlation of 0.18 implies that number of filler had a not significant correlation with the effect of narrative.

Number of Interjections(*NI*) had a correlation of 0.57 and Number of nodding Times(*NT*) had a correlation 0.54 with *MES* . This implies that the feedback from listener can help narrators to grasp the conversation situation and improve the interactiveness between narrators and listeners.

6. Conclusion and Future Work

In this work, we extracted verbal and nonverbal features of participants from Cartoon Narrative Task, we made a correlation analysis between these features and the evaluation of narrative effect.Our result showed that the group speaking length and feedback of listeners (number of interjections and nodding times) have significant correlations with the third party evaluation of narrative effect in face-to-face conversation. As future work, we will increase the number of verbal and nonverbal features and analyze them by using machine learning techniques.

References

[1] Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review.Image Vis. Comput. 27(12), 1775–1787 (2009)

[2] Shogo Okada. Context-based conversational hand gesture classification in narrative interaction: ICMI '13,Pages 303-310

[3] H.H. Clark, T.B. Carlson, Hearers and speech acts, Language 58 (2) (1982) 332–373.

[4] R. Gifford, Personality and nonverbal behavior: a complex conundrum, in: V. Manusov, M.L. Patterson (Eds.), The SAGE Handbook of Nonverbal Communication, Sage, Beverly Hills, CA, 2006. [5] R. Stiefelhagen, Tracking focus of attention in meetings, in: Proc. of the Int.Conf. on Multimodal Interfaces (ICMI), Pittsburgh, PA, 2002.

[6] D. McNeill. Hand and Mind: What Gestures Reveal about Thought. Psychology/cognitive science. University of Chicago Press, 1996.

[7] http://chasen.naist.jp/hiki/ChaSen/