User Modeling by Using Bag-of-Behaviors for Building a Dialog System Sensitive to the Interlocutor's Internal State

Yuya Chiba, Takashi Nose, Akinori Ito
Graduate School of Engineering,
Tohoku University, Japan

Masashi Ito
Faculty of Engineering
Tohoku Institute of Technology, Japan

Abstract

When using spoken dialog systems in actual environments, users sometimes abandon the dialog without making any input utterance. To help these users before they give up, the system should know why they could not make an utterance. Thus, we have examined a method to estimate the state of a dialog user by capturing the user's non-verbal behavior even when the user's utterance is not observed. The proposed method is based on vector quantization of multi-modal features such as non-verbal speech, feature points of the face, and gaze. The histogram of the VQ code is used as a feature for determining the state. We call this feature "the Bag-of-Behaviors." According to the experimental results, we prove that the proposed method surpassed the results of conventional approaches and discriminated the target user's states with an accuracy of more than 70%.

1 Introduction

Spoken dialog systems have an advantage of being a natural interface since speech commands are less subject to the physical constraints imposed by devices. On the other hand, if the system accepts only a limited expression, the user need to learn how to use the system. If the user is not familiar with the system, he/she cannot even make an input utterance. Not all users are motivated to converse with the system in actual environments, and sometimes a user will abandon the dialog without making any input utterance. When the user has difficulty to make the utterance, conventional systems just repeat the prompt at fixed interval (Yankelovich, 1996) or taking the initiative in the dialog to complete the task (Chung, 2004; Bohus and Rudnicky, 2009). However, we think that the system has to cope with the user's implicit requests to help the user more adequately. To solve this problem, Chiba and Ito (2012) proposed a method to estimate two "user's states" by capturing their non-verbal cues. Here, the state A is when the user does not know what to input, and the state B is when the user is considering how to answer the system's prompt. These states have not been distinguished by the conventional dialog systems so far, but should be handled differently.

The researchers of spoken dialog systems have focused on the various internal states of users such as emotion (Forbes-Riley and Litman, 2011a; Metallinou et al., 2012), preference (Pargellis et al., 2004) and familiarity with the system (Jokinen and Kanto, 2004; Rosis et al., 2006) to build natural dialog system. In particular, the user's "uncertainty" is assumed to be the nearest user's states that we wish to study. Forbes-Riley and Litman (2011b) and Pon-Barry et al. (2005) introduced a framework for estimating the user's uncertainty to a tutor system.

The above-mentioned researches have a certain result by employing linguistic information for the estimation, but it remains difficult to assist a user who does not make any input utterance. By contrast, the method by Chiba and Ito (2012) estimated the target user's state by only using the user's non-verbal information. In their work, the user's multi-modal behaviors were defined empirically, and the labels of the behaviors were annotated manually. Based on this result, the present paper proposes the method that does not use manually-defined labels nor manual annotation. The multi-modal behaviors are determined automatically using the vector quantization, and the frequency distribution of the VQ code is used for estimation of the user's state. Because this approach expects to construct clusters of the speech events or behaviors of the user, we called it as Bag-of-Behaviors approach.

2 Data collection

The experimental data (video clips) were the same as those used in the experiment by Chiba et al. (Chiba and Ito, 2012; Chiba et al., 2012). The video clips contained the frontal image of the user

Proceedings of the SIGDIAL 2014 Conference, pages 74–78,
Philadelphia, U.S.A., 18-20 June 2014. ©2014 Association for Computational Linguistics

and their speech, which were recorded with a web camera and a lapel microphone, respectively. The task of the dialog was a question-and-answer task to ask users to answer common knowledge or a number they remembered in advance, such as "Please input your ID." 16 users (14 males and 2 females) participated in the dialog collection.

Recorded clips were divided into sessions, where one session included one interchange of the system's prompt and the user's response. The total number of sessions was 792. Then we employed evaluators to label each video clip as either state A, B or C, where state A and B were that described in the previous section, and state C is the state where the user had no problem answering the system. We took the majority vote of the evaluators' decisions to determine the final label of a clip. Fleiss' κ among the evaluators was 0.22 (fair agreement). Finally, we obtained 59, 195 and 538 sessions of state A, B and C, respectively.

3 Discrimination method by using Bag-of-Behaviors

In the work of Chiba et al. (2013), the user's state was determined using the labels of the multi-modal events such as fillers or face orientation, which were estimated from the low-level acoustic and visual features.

Here, inventory of multi-modal events was determined empirically. There were, however, two problems with this method. The first one was that the optimality of the inventory was not guaranteed. The second one is that it was difficult to estimate the events from the low-level features, which made the final decision more difficult. Therefore, we propose a new method for discriminating the user's state using automatically-determined events obtained by the vector quantization.

First, a codebook of the low-level features (which will be described in detail in the next section) is created using k-means++ algorithm (Arthur and Vassilvitskii, 2007). Let a low-level feature vector at time t of session s of the training data be $\boldsymbol{x}_t^{(s)}$. Then we perform the clustering of the low-level feature vectors for all of t and s, and create a codebook $\mathcal{C} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_K\}$, where $\boldsymbol{c}_k$ denotes the k-th centroid of the codebook.

Then the input feature vectors are quantized frame-by-frame using the codebook. When a session for evaluation s_E is given, we quantize the input low-level feature vectors $\boldsymbol{x}_1^{(s_E)}, \dots, \boldsymbol{x}_T^{(s_E)}$ into $q_1, \dots, q_T$, where

$$q_t = \arg\min_q ||\boldsymbol{x}_t^{(s_E)} - \boldsymbol{c}_q||. \tag{1}$$

Then we calculate the histogram $\boldsymbol{Q}_0(s_E) = (Q_1, \dots, Q_K)$ where

$$Q_k = \sum_{t=1}^{T} \delta(k, q_t) \tag{2}$$

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{3}$$

Then $\boldsymbol{Q}(s_E) = \boldsymbol{Q}_0(s_E)/||\boldsymbol{Q}_0(s_E)||$ is used as the feature of the discrimination. The similar features based on the vector quantization were used for image detection and scene analysis (Csurka et al., 2004; Jiang et al., 2007; Natarajan et al., 2012) and called "Bag-of-Features" or "Bag-of-Keypoints." In our research, each cluster of the low-level features is expected to represent some kind of user's behavior. Therefore, we call the proposed method the "Bag-of-Behaviors" approach.

After calculating the Bag-of-Behaviors, we employ an appropriate classifier to determine the user's state in the given session. In this research, the support vector machine (SVM) is used as a classifier.

4 The low-level features

In this section, we describe the acoustic and visual features employed as the low-level features.

The target user's states are assumed to have similar aspects to emotion. Collignon et al. (2008) suggested that emotion has a multi-modality nature. For example, Wöllmer et al. (2013) showed that the acoustic and visual features contributed to discriminate arousal and expectation, respectively. Several other researches also have reported that recognition accuracy of emotion was improved by combining multi-modal information (Lin et al., 2012; Wang and Venetsanopoulos, 2012; Paulmann and Pell, 2011; Metallinou et al., 2012). Therefore, we employed similar features as those used in these previous works, such as the spectral features and intonation of the speech, and facial feature points, etc.

4.1 Audio features

To represent spectral characteristics of the speech, MFCC was employed as an acoustic feature. We used a 39-dimension MFCC including the velocity and acceleration of the lower 12th-order coefficients and log power. In addition, a differential component of log $F0$ was used to represent the prosodic feature of the speech, and zero cross (ZC) was used to distinguish voiced and unvoiced segments. Therefore, total number of audio features was 3. The basic conditions for extracting each feature are shown in Table 1. Here, five frames (the current frame, the two previous frames and two following frames) were used to calculate the Δ and $\Delta\Delta$ components of MFCC and Δ component of log $F0$.

4.2 Face feature

Face feature (Chiba et al., 2013) was extracted by the Constraint Local Model (CLM) (Saragih et al., 2011) frame by frame. The coordinates of the points relative to the center of the face were used as the face features. The scale of the feature points was normalized by the size of the facial region. The number of feature points was 66 and the dimension of the feature was 132.

4.3 Gaze feature

The evaluators of the dialogs declared that movement of the user's eyes seems to express their internal state. The present paper used the Haar-like feature which has a fast calculation algorithm using the integral image to represent the brightness of the user's eye regions. This feature was extracted by applying filters comprehensively changed the size and location to the image (eye regions in our case). The eye regions were detected by the facial feature points. Because this feature had large dimensions, the principal component analysis (PCA) was conducted to reduce the dimensionality. Finally, gaze feature had 34 dimensions and the cumulative contribution rate was about 95%.

4.4 Feature synchronization

The audio features were calculated every 10 ms (see Table 1) while the visual features were extracted every 33 ms. Therefore, the features were synchronized by copying the visual features of the previous frame in every 10 ms.

5 Discrimination examination

5.1 Conditions of the Bag-of-Behaviors construction

We built the Bag-of-Behaviors under two conditions described below.

Let $\boldsymbol{x}_{at}^{(s)}, \boldsymbol{x}_{ft}^{(s)}$ and $\boldsymbol{x}_{et}^{(s)}$ represent the audio feature, face feature and gaze feature of the session s at time t, respectively.

Table 1: Conditions of audio feature extraction

	MFCC	log $F0$	ZC
Frame width	25.0 ms	17.0 ms	10.0 ms
Frame shift	10.0 ms	10.0 ms	10.0 ms

Table 2: Experimental conditions

# of sessions		State A(59), State B(195)
Codebook size	K	4, 8, 16, 32, 64
	K_a	4, 8, 16, 32, 64
	K_f	4, 8, 16, 32, 64
	K_e	4, 8, 16, 32, 64

In Condition (1), the three features are combined to single feature vector $\boldsymbol{x}_t^{(s)}$:

$$\boldsymbol{x}_t^{(s)} = (\boldsymbol{x}_{at}^{(s)}, \boldsymbol{x}_{ft}^{(s)}, \boldsymbol{x}_{et}^{(s)}) \tag{4}$$

Then, the low-level feature vectors $\boldsymbol{x}_t^{(s)}$ are clustered to construct one codebook $\mathcal{C}$ with size K. When an input session s_E is given, we calculate the combined feature vector $\boldsymbol{x}_t^{(s_E)}$, and generate the Bag-of-Behaviors $\boldsymbol{Q}(s_E)$. This method is a kind of the feature-level fusion method.

In Condition (2), the three features are used separately. First, we generate three codebooks $\mathcal{C}_a, \mathcal{C}_f$ and $\mathcal{C}_e$ using the audio, face and gaze features, respectively. Size of those codebooks were K_a, K_f and K_e. When an input session s_E is given, we generate three Bag-of-Behaviors feature vectors $\boldsymbol{Q}_a(s_E), \boldsymbol{Q}_f(s_E)$ and $\boldsymbol{Q}_e(s_E)$ using the three codebooks. Finally, we combine those features as

$$\boldsymbol{Q}(s_E) = (\boldsymbol{Q}_a(s_E), \boldsymbol{Q}_f(s_E), \boldsymbol{Q}_e(s_E)). \tag{5}$$

5.2 Experimental condition

We employed the SVM with RBF-kernel as a classifier. The experimental conditions are summarized in Table 2. The hyperparameters of the classifier were decided by grid-searching. Since the session of state C and the other states (state A and state B) were clearly distinguished by the duration of the session, we used only the session of state A and state B for the experiments. Hence, each experiment was a two-class discrimination task.

As explained, the experimental data were unbalanced. Since it is desirable that the system can discriminate the user's state without deviation, the harmonic mean H of the accuracy of the two states was used for measuring the performance. This is calculated by

$$H = \frac{2C_A C_B}{C_A + C_B}, \tag{6}$$

where C_A and C_B represent the discrimination accuracy of state A and state B, respectively. The experiments were conducted based on a 5-fold cross validation.

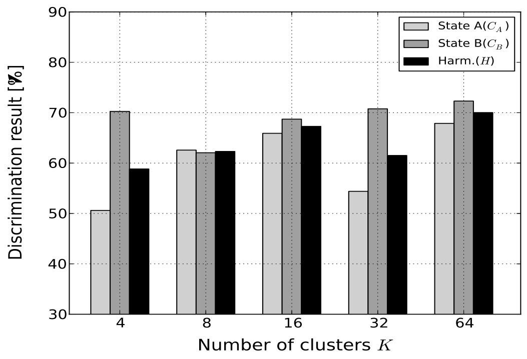

Figure 1: Discrimination results of condition (1)

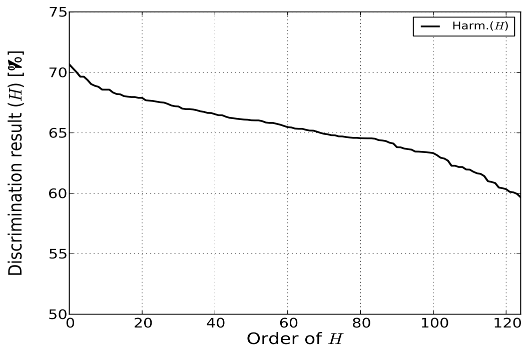

Figure 2: Discrimination results of condition (2) arranged in descending order

5.3 Experimental results

The results of condition (1) are shown in Figure 1. The figure shows the best H of each number of clusters. In condition (1), the best result ($H = 70.0\%$) was obtained when the number of clusters K was 64. Figure 2 shows the results of condition (2). In this figure, the results are shown in descending order of the harmonic mean for all combination of codebook size of the three codebooks (there were $5^3 = 125$ conditions). The best $H = 70.7\%$ was obtained when $K_a = 8, K_f = 8$ and $K_e = 64$.

The best results of the tested methods are summarized in Table 3. Here, "Baseline + NN" in the table denotes the result in Chiba et al. (2013), where the visual events and acoustic events were annotated manually, and the manual labels were used as input for a neural network for the classification. The gaze feature was not used in "Baseline + NN." We added the result when including the gaze feature, shown as "Baseline + Gaze + NN." As shown in Table 3, the performance of the method proposed in this paper surpassed the baseline methods. Therefore, the proposed method could not only automatically determine the inventory of the audio-visual events, but also achieved better discrimination accuracy. One of the reasons of the improvement is VQ can construct the clusters in proper quantities.

Table 3: Comparison of estimation methods

	State A	State B	Harm.
Baseline + NN	52.5	65.1	58.2
Baseline + Gaze + NN	64.5	59.5	61.9
Condition (1) + RBF-SVM	67.9	72.3	70.0
Condition (2) + RBF-SVM	67.7	73.8	70.7
Condition (2) + MKL-SVM	**68.0**	**76.4**	**72.0**

Comparing the two conditions of feature combination, H of condition (2) (denoted as "Condition (2) + RBF-SVM") was slightly higher than that of condition (1) (denoted as "Condition (1) + RBF-SVM"). This result was similar to Split-VQ (Pariwal and Atal, 1991) where a single feature vector split into subvectors and the input vector was quantized subvector by subvector.

We conducted additional experiments for condition (2) by using SVM with combined kernel trained by Multiple Kernel Learning (MKL) (Sonnenburg et al., 2006). The combined kernel is represented as a linear combination of several subkernels. The distinct kernel was employed for the speech, face feature and gaze feature, respectively. This paper used the RBF-kernel having the same width as the sub-kernels. The best result was shown as "Condition (2) + MKL-SVM" in Table 3. As shown in the table, the MKL-SVM showed the highest performance of 72.0 %. The weights of the audio, face and gaze feature were 0.246, 0.005 and 0.749, respectively. This result suggested that the contribution of the face feature was weaker than the other features.

6 Conclusion

In this paper, we proposed a method to estimate the state of the user of the dialog system by using non-verbal features. We proposed the Bag-of-Behaviors approach, in which the user's multmodal behavior was first classified by vector quantization, and then the histogram of the VQ code was used as a feature of the discrimination. We verified that the method could discriminate the target user's state with an accuracy of 70% or more.

One of the disadvantages of the current framework is that it requires to observe the session until just before the user's input utterance. This problem makes it difficult to apply this method to an actual system, because the system has to be able to evaluate the user's state successively in order to help the user at an appropriate timing. Therefore, we will examine a sequential estimation method by using the Bag-of-Behaviors in a future work.

References

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proc. the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.

Dan Bohus and Alexander I. Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361.

Yuya Chiba and Akinori Ito. 2012. Estimating a user's internal state before the first input utterance. *Advances in Human-Computer Interaction*, 2012:11, DOI:10.1155/2012/865362, 2012.

Yuya Chiba, Masashi Ito, and Akinori Ito. 2012. Effect of linguistic contents on human estimation of internal state of dialog system users. In *Proc. Feedback Behaviors in Dialog*, pages 11–14.

Yuya Chiba, Masashi Ito, and Akinori Ito. 2013. Estimation of user's state during a dialog turn with sequential multi-modal features. In *HCI International 2013-Posters' Extended Abstracts*, pages 572–576.

Grace Chung. 2004. Developing a flexible spoken dialog system using simulation. In *Proc. the 42nd Annual Meeting on Association for Computational Linguistics*, pages 63–70.

Olivier Collignon, Simon Girard, Frederic Gosselin, Sylvain Roy, Dave Saint-Amour, Maryse Lassonde, and Lepore Franco. 2008. Audio-visual integration of emotion expression. *Brain research*, 1242:126–135.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Proc. workshop on statistical learning in computer vision, ECCV*, pages 1–2.

Kate Forbes-Riley and Diane Litman. 2011a. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53:1115–1136.

Kate Forbes-Riley and Diane Litman. 2011b. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105–126.

Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proc. of the 6th ACM international conference on Image and video retrieval*, pages 494–501.

Kristiina Jokinen and Kari Kanto. 2004. User expertise modelling and adaptivity in a speech-based e-mail system. In *Proc. the 42nd Annual Meeting on Association for Computational Linguistics*, pages 88–95.

Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei. 2012. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Trans. Multimedia*, 14(1):142–156.

Angeliki Metallinou, Martin Wöllmer, Athanasios Katsamanis, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2012. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affective Computing*, 3(2):184–198.

Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, and Unsang Park, Rohit Prasad, and Premkumar Natarajan. 2012. Multimodal feature fusion for robust event detection in web videos. In *Proc. Computer Vision and Pattern Recognition*, pages 1298–1305.

Andrew Pargellis, Hong-Kwang Jeff Kuo, and Chin-Hui Lee. 2004. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42:329–351.

Kuldip Paliwal and Bishnu Atal. 1993. Efficient vector quantization of lpc parameters at 24 bits/frame. In *IEEE Trans. Speech and Audio Processing*, 1(1):3–14.

Silke Paulmann and Marc Pell. 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, 35(2):192–201.

Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. 2005. Responding to student uncertainty in spoken tutorial dialogue systems. *Int. J. Artif. Intell. Edu.*, 16:171–194.

Fiorella Rosis, Nicole Novielli, Valeria Carofiglio, Addolorata Cavalluzzi, and Berardina Carolis. 2006. User modeling and adaptation in health promotion dialogs with an animated character. *J. Biomedical Informatics*, 39:514–531.

Jason Saragih, Simon Lucey, and Jeffrey Cohn. 2011. Deformable model fitting by regularized landmark meanshift. *Int. J. Computer Vision*, 91(2):200–215.

Yongjin Wang and Anastasios Venetsanopoulos. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Trans. Multimedia*, 14(3):597–607.

Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Nicole Yankelovich. 1996. How do users know what to say? *Interactions*, 3(6):32–43.