End-to-end modeling for selection of utterance constructional units via system internal states

Koki Tanaka, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

Abstract In order to make conversational agents or robots conduct human-like behaviors, it is important to design a model of the system internal states. In this paper, we address a model of favorable impression to the dialogue partner. The favorable impression is modeled to change according to user's dialogue behaviors and also affect following dialogue behaviors of the system, specifically selection of utterance constructional units. For this modeling, we propose a hierarchical structure of logistic regression models. First, from the user's dialogue behaviors, the model estimates the level of user's favorable impression to the system and also the level of the user's interest in the current topic. Then, based on the above results, the model predicts the system's favorable impression to the user. Finally, the model determines selection of utterance constructional units in the next system turn. We train each of the logistic regression models individually with a small amount of annotated data of favorable impression. Afterward, the entire multi-layer network is fine-tuned with a larger amount of dialogue behavior data. An experimental result shows that the proposed method achieves higher accuracy on the selection of the utterance constructional units, compared with methods that do not take into account the system internal states.

1 Introduction

It is important for spoken dialogue systems to introduce internal states in order to realize human-like dialogue. By taking into account both input user utterances and system internal states, spoken dialogue systems are expected to generate more human-like natural utterances. Emotion has been considered as an internal state for spoken dialogue systems and virtual agents [13, 2, 3].

All authors
Graduate School of Informatics, Kyoto University, Kyoto, Japan.
e-mail: [tanaka][inoue][shizuka][takanasi][kawahara]@sap.ist.i.kyoto-u.ac.jp

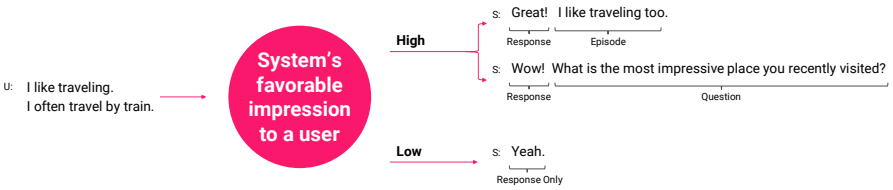

Fig. 1 Main idea of the proposed system that selects the next system utterance based on the system's favorable impression toward the user (U: user, S: system)

We address *favorable impression* to a user as an internal state of the system. We set up a speed-dating dialogue task where a male user talks with a female conversational robot about their profiles. In human-human speed-dating dialogue, their behaviors and attitudes sometimes reflect the degree of favorable impression to their interlocutors [9, 12]. In this study, to express the degree of favorable impression, we propose a dialogue system that selects utterance constructional units, inspired by a series of studies on the discourse analysis [17]. The utterance constructional units contain three parts: *response*, *episode*, and *question*. *Response* is a reaction to the user's utterance, such as feedbacks and answers to questions. *Episode* corresponds to information given by the system such as self-disclosure. *Question* is made by the system toward the user to elaborate the current topic or change the topic. Figure 1 illustrates the main idea of our proposed system. For example, when the degree of favorable impression to the user is high, the system tends to select multiple units such as the combination of *response* and *episode*, or another combination of *response* and *question*, to be more talkative. On the other hand, when the degree is low, the system would select only *response*.

We realize selection of utterance constructional units by a hierarchical structure of logistic regression models. The input is a set of features based on the user's dialogue behaviors. The output is a selection of the utterance constructional units of the next system turn. In the intermediate layer of the hierarchical structure, the degree of favorable impression is represented as an internal state. The proposed model predicts the favorable impression to the user and then the utterance constructional units step by step, where each step is realized with a logistic regression model. We train each logistic regression model with annotated labels of the favorable impression to the user. However, it is difficult to obtain a large number of training labels for the internal states. On the other hand, it is easier to get a large amount of data for the input and output behaviors because these are actual behaviors that can be objectively defined and observed in dialogue corpora. In this paper, we also propose an efficient model training to leverage the benefits of making use of internal states. At first, we pre-train each logistic regression model with a small number of training labels of the internal states. We then fine-tune the whole neural network with a larger amount of data of the input and output behaviors in an end-to-end manner. The pre-training captures the internal states, and the end-to-end fine-tuning scales up the amount of training data, which is vital for robust training. This study contributes to realizing

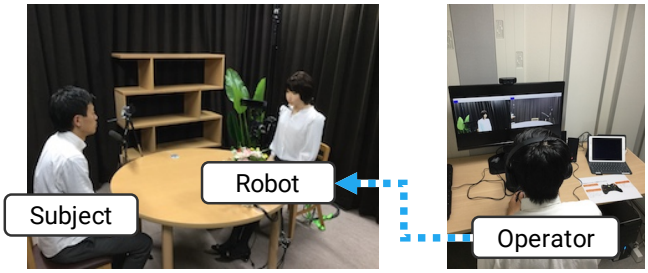

Fig. 2 Snapshot of data collection in WoZ setting

dialogue systems that model internal states and also efficient model training where the amount of training data for the internal states is limited.

2 Speed-dating human-robot dialogue corpus

In this section, we explain the dialogue data used in this study. We recorded a set of speed-dating dialogues where a male human subject talked with a female humanoid robot that was operated by another female subject. Right after the recording, we took a survey to obtain training labels of the internal states. We also manually annotated the utterance constructional units on the recorded dialogue data.

2.1 Dialogue data collection

We have collected a series of speed-dating dialogues between a male subject and a female humanoid robot named ERICA [7, 10]. ERICA was operated by another human subject, called an operator, who was in a remote room. When the operator spoke, the voice was directly played with a speaker placed on ERICA, and the lip and head motion of ERICA was automatically generated [8, 14]. The operator also controlled ERICA's behaviors such as eye-gaze, head nodding, and arm gestures. The snapshot of this data collection is shown in Figure 2. We recorded 18 dialogue sessions which lasted 10 minutes and 55 seconds on average. The human subjects were 18 male university students (both undergraduate and graduated students). The ERICA's operators were 4 actresses whose ages ranged from 20s to 30s. Whereas each human subject participated in only one dialogue session, each ERICA's operator participated in several sessions. They are all native Japanese speakers. We used multimodal sensors that consisted of microphones, a microphone array, RGB cameras, and Kinect v2. We manually annotated utterances, backchannels, laughing, fillers, dialogue turns, and dialogue acts using recommended standards [5].

The dialogue scenarios and instructions are as follows. Since they met each other for the first time, they had to exchange their personal information to know well each other. In advance, we gave the participants a list of conversational topics that are likely to be talked about in first-encounter dialogues, such as hobbies, occupation, and hometown. We then instructed the participants to make a conversation based on the topic list. In the actual dialogue, participants often talked about the topics on the list such as favorite movies, sports, food, and recent trips. For the ERICA's operator, we instructed how to select the utterance constructional units together with the concept of the favorable impression. We asked the operator to select the utterance constructional units based on the degree of her favorable impression to the subject, but we also told that she did not necessarily need to follow this to keep the dialogue natural. We also told that the operator did not need to entertain the subject and the degree of her favorable impression to the subject could be not only positive but also negative.

After each dialogue session, we asked the operator to answer a survey. After the operator listed dialogue topics that they talked about, she rated the following items for each topic on the 7-point scale.

1. Operator's favorable impression to the subject
2. Subject's favorable impression to ERICA estimated by the operator
3. Operator's interest in the topic
4. Subject's interest in the topic estimated by the operator

The favorable impression is represented in one-dimension, positive and negative, as we regard it as a specific indicator in first-encounter dialogue. Although we conducted a similar survey to the male subjects, we used only the survey result from the operators. The reason is that the male subject was a different person on each dialogue session while the operators' survey should be consistent among sessions.

2.2 Analysis

First, we segmented all utterances by dialogue turns. In total, the number of turns of the operators was 899. Then, we manually annotated a set of utterance constructional units for each turn. This annotation was made by one annotator. The distribution of the patterns of utterance constructional units is reported in Table 1. As we see from the table, the majority of the patterns of utterance constructional units was response only (472 samples). Notably, the operators occasionally gave their episode and asked back questions, but the cases having both an episode and a question was very rare (8 samples). We hypothesize that the operators reflected their favorable impression to the subjects on the utterance constructional units.

We analyzed the survey results from the operators on the following items: (1) operator's favorable impression to the user, (2) subject's favorable impression to ERICA estimated by the operator, (3) operator's interest in each dialogue topic, and (4) subject's interest in each dialogue topic estimated by the operator. The distribu-

Table 1 Distribution of the pattern of utterance constructional units

Utterance constructional units			
Response	*Episode*	*Question*	Frequency
✓	-	-	472
✓	✓	-	177
✓	-	✓	86
-	✓	-	69
-	-	✓	53
✓	✓	✓	8
others			34
Total			899

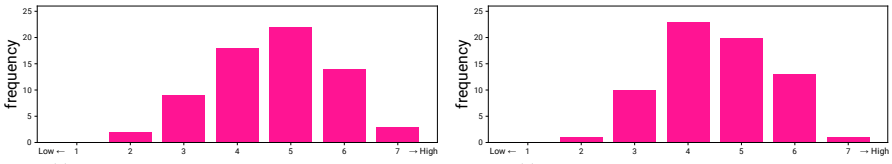

Fig. 3 Distribution of favorable impression reported by ERICA's operators

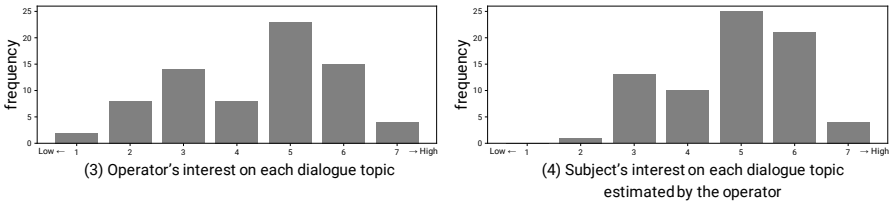

Fig. 4 Distribution of interest reported by ERICA's operators

tions of the four items are plotted in Figure 3 and Figure 4. The number of dialogue topics was 74 in total. The distributions of interest tended to be more varied than those of favorable impression. This result suggests that the degree of interest more depends on the dialogue topics. On the other hand, this result also suggests that the favorable impression is more stable and gradually changes during the dialogue.

3 Problem formulation

The task of this study is to select the utterance constructional units of the next system turn based on observed behavior features of the user. The problem formulation is illustrated in Figure 5. The input feature vector is based on both the speaking and listening behaviors of the user. The speaking behavior feature is extracted during the preceding user turn, referred as $\boldsymbol{o}_s$. The listening behavior feature is computed during the last system turn, referred as $\boldsymbol{o}_l$. We concatenate the behavior feature vectors

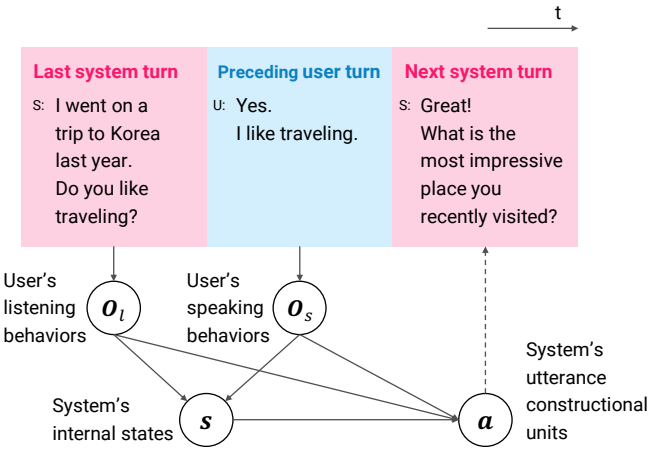

Fig. 5 Problem formulation for considering internal states to select the system next action

as:

$$\boldsymbol{o} := (\boldsymbol{o}_s\ ,\ \boldsymbol{o}_l)\ . \tag{1}$$

The detail of the feature set is explained in Section 5. The output is the pattern of the utterance constructional units that consists of three elements: *response*, *episode*, and *question*. We refer the output as a system action $\boldsymbol{a}$. In this study, we take into account the internal states such as the system's favorable impression to the user. We define the internal states as a vector $\boldsymbol{s}$. In summary, the problem in this study is to predict the next system action $\boldsymbol{a}$ from the observation behaviors $\boldsymbol{o}$ by considering the internal states $\boldsymbol{s}$. This is a typical formulation in conventional studies on spoken dialogue systems where the internal states $\boldsymbol{s}$ correspond to dialogue states of slot filling. In the case of conventional studies such as task-oriented dialogues, the dialogue states were defined clearly and objectively, which makes it easy to collect a large number of training labels for statistical dialogue models such as Markov decision process (MDP) and partially observable Markov decision process (POMDP) [20]. In the current study on the first-encounter dialogue, however, the internal states correspond to states such as favorable impression. These states are ambiguous and subjective, which makes it difficult to prepare a sufficient number of training labels of them. Therefore, we propose efficient end-to-end training by facilitating a small number of labels of the internal states.

Since the distribution of the utterance constructional units is skewed as shown in Table 1, we do not directly select the utterance constructional units. Instead, we divide this problem into the following two sub-tasks. These sub-tasks can be defined as a taxonomy depicted in Figure 6. The first task is to decide whether the system's turn consists of a response only or have other units (an episode and/or a question). If the decision is the latter case, the system triggers the second task which is to decide whether the system generates an episode or a question. Since we could observe only

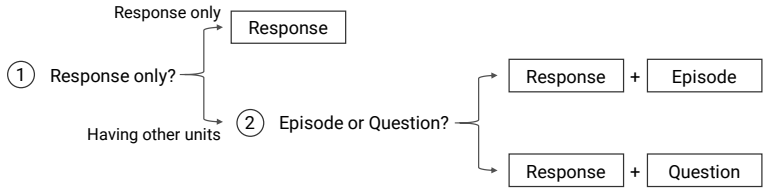

Fig. 6 Taxonomy for selection of the utterance constructional units. The numbers (1 and 2) in the figure correspond to classification tasks.

Table 2 Definition of labels of the utterance constructional units for each task (p: positive sample, n: negative sample, -: not used)

Utterance constructional units				Task	
Response	*Episode*	*Question*	Freq.	1	2
✓	-	-	472	p	-
✓	✓	-	177	n	p
✓	-	✓	86	n	n
-	✓	-	69	n	p
-	-	✓	53	n	n
✓	✓	✓	8	n	-
others			34		
Total			899		

a few samples where all three utterance constructional units were used at the same time, we do not consider this rare case in the current formulation. In this study, we make the selection model for each task independently, but we combine them to decide the pattern of the utterance constructional units finally. The distribution and definition of labels of the utterance constructional units are summarized in Table 2. The first task corresponds to the selection between the majority pattern and the others. The second task focuses on the remainder steps.

4 End-to-end modeling using a small number of labels of internal states

We take into account the internal states such as favorable impression to the user in order to select the utterance constructional units of the next system turn. However, the number of training labels of the internal states is limited. Actually, in the current study, we could obtain the labels of favorable impression and interest only on each topic, whereas we have to generate the system's action for every turn. This is a universal problem in modeling internal states. On the other hand, we can easily obtain the labels of behaviors such as the observation $\boldsymbol{o}$ and the action $\boldsymbol{a}$ because these behaviors can be objectively observed.

We propose efficient end-to-end modeling for the selection of the utterance constructional units by using a small number of labels of the favorable impression and

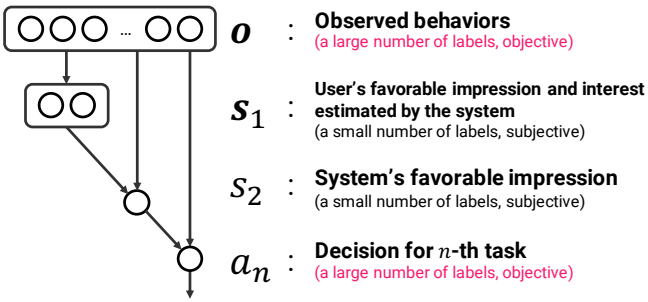

Fig. 7 Proposed model considering internal states as hidden layers of the network

the interest. The proposed model is based on hierarchical neural networks where the internal states are represented as hidden layers. Figure 7 depicts an overview of the proposed model. First, we train each layer one by one. For example, we train a prediction model for the user's favorable impression to the system based on the observed behaviors of the user ($\boldsymbol{o}$). This pre-training is done with a small number of labels of the internal states. After we train each layer, the entire network is fine-tuned with a much larger number of data sets of the observation $\boldsymbol{o}$ and the output system action $\boldsymbol{a}$.

4.1 Network architecture

The proposed hierarchical neural network estimates the internal states step by step. The network architecture is depicted in Figure 7. We observe the input feature vector $\boldsymbol{o}$. The dimension of the input vector is D_o. Although it is possible to directly predict the system's favorable impression to the user from the observation, we first estimate the user's favorable impression to the system and the interest on the current topic as:

$$\boldsymbol{s}_1 = \sigma(A_1 \boldsymbol{o}^T + \boldsymbol{b}_1^T) \ , \tag{2}$$

where $\boldsymbol{s}_1$ is a two-dimension vector corresponding to the values of the user's favorable impression and interest estimated by the system. A_1 and $\boldsymbol{b}_1$ are network parameters whose sizes are $2 \times D_o$ and 2, respectively. $\sigma()$ is the sigmoid function and T represents the transpose. Next, we predict the system's favorable impression to the user from both the user's favorable impression to the system and the interest estimated in the previous step and also from the observation (referred as $\boldsymbol{s}_{1'} = (\boldsymbol{s}_1, \boldsymbol{o})$):

$$s_2 = \sigma(A_2 \boldsymbol{s}_{1'}^T + b_2) \ , \tag{3}$$

where s_2 is a scalar corresponding to the value of the system's favorable impression to the user. A_2 and b_2 are network parameters whose sizes are $1 \times (2 + D_o)$ and 1,

respectively. Finally, we calculate the probability for each task of the selection of the utterance constructional units in the same manner as:

$$a_n = \sigma(A_3 \boldsymbol{s}_{2'}^T + b_3) \ , \tag{4}$$

where $\boldsymbol{s}_{2'}$ is a concatenated vector consisting of the predicted system's favorable impression to the user and the observation as $\boldsymbol{s}_{2'} = (s_2, \boldsymbol{o})$, and a_n is the probability for the n-th task which was defined as a binary classification defined in Section 3. A_3 and b_3 are network parameters whose sizes are $1 \times (1 + D_o)$ and 1, respectively. In this study, we solve the two tasks individually. We train the above model for each task, and the set of the output scalar values make a final system action $\boldsymbol{a}$.

4.2 Model training

The model training consists of two steps: pre-training and fine-tuning. First, we train each layer step by step as pre-training. Since we have labels of the internal states on each topic, we assume the internal states are unchanged during the same topic. This limitation also means that it is difficult to scale up the number of labels of the internal states. Therefore, we fine-tune the entire network with a larger number of labels of the observation $\boldsymbol{o}$ and the system action $\boldsymbol{a}$ through back-propagation. To keep the effect of the pre-training, we add the square error between the model parameters by the pre-training and those after the fine-tuning to the loss function as:

$$E'(W) = E(W) + SE(W, W_{pre}) \ , \tag{5}$$

where $E(W)$ is the loss function of the output layer of the network, and $SE(W, W_{pre})$ is the square error between the model parameters by the pre-training (W_{pre}) and those after the fine-tuning (W). Specifically, we summed squared Frobenius norm of the difference of each model parameter to calculate $SE(W, W_{pre})$.

5 Feature set

In order to implement the proposed system, we need to define the observation vector $\boldsymbol{o} = (\boldsymbol{o}_s, \boldsymbol{o}_l)$. We use both features of speaking and listening behaviors of the user. The features are chosen based on previous studies on emotion and interest recognition [15, 19, 1, 16, 18]. We manually annotated the following features and used them in the experiment, although the majority of these features can be computed automatically.

The features of speaking behaviors $\boldsymbol{o}_s$ are calculated from the preceding user turn and listed below.

- Turn duration

- Pause duration between the end of the last system turn and the beginning of the preceding user turn
- Voice activity ratio
- Global voice activity ratio from the beginning of the dialogue until the end of the preceding user turn
- Speech rate
- Intensity (mean, range)
- F0 (mean, range)
- Length of *episode* (if the turn contains *episode* otherwise zero)
- Laughter frequency
- Filler frequency (short phrases that fill a pause within a turn, such as "*uh*")
- Pattern of utterance constructional units

We used the Praat [4] software to extract intensity and F0 from the user utterances. We approximated the length of *episode* as the number of long utterance units (LUUs) [6]. The LUUs are defined to approximate semantic units so that we intended to capture the substantial volume of the episode. The pattern of the utterance constructional units of the user turn is represented as binary vectors where each dimension corresponds to the occurrence of each element of the utterance construction unit. The dimension of the vector $\boldsymbol{o}_s$ is 18.

The features of listening behaviors $\boldsymbol{o}_l$ are calculated from the last system turn and listed below.

- Backchannel frequency (such as "*yeah*")
- Laughter frequency

The dimension of the vector $\boldsymbol{o}_l$ is 2. We squeezed the feature set to these because this is the first step of the study. In future work, we will consider the use of additional listening behaviors such as eye gaze and head nodding.

6 Experimental evaluation

We evaluated the proposed method with the first-encounter dialogue corpus described in Section 2. Five-fold cross validation was conducted to calculate an average precision, recall, and F1 score. We implemented the neural network model with TensorFlow 1.7.0.We used Adam [11] as the optimization method and empirically set the learning rate at 10^{-2} for the first task and 10^{-6} for the second task. We prepared three compared models. The first model is to directly predict the utterance constructional units from the observation with a one-layer neural network, which is equivalent to a logistic regression model, referred as *baseline*. The second model has the same architecture as the proposed model in that it is a multi-layer neural network, but the pre-training is not conducted. Instead, the network parameters are initialized with random values, referred as *w/o. pre-training*. The third model is same as the proposed model, but the fine-tuning is not conducted, referred as *w/o. fine-tuning*.

Table 3 Prediction result on the first task (response only or having other units)

model	Precision	Recall	F1
baseline	0.667	0.658	0.662
w/o. pre-training	0.628	0.643	0.635
w/o. fine-tuning	0.708	0.586	0.641
proposed	0.679	0.674	0.677

Table 4 Prediction result on the second task (episode or question)

model	Precision	Recall	F1
baseline	0.617	0.748	0.676
w/o. pre-training	0.649	0.740	0.692
w/o. fine-tuning	0.664	0.784	0.719
proposed	0.666	0.788	0.722

As shown in Figure 6, we solve two different tasks for the selection of the utterance constructional units: (1) response only or having other units, (2) generate an episode or a question. The ratios of positive samples (chance levels) in the whole data set are 0.527 and 0.605 on the first and second tasks, respectively. The results of the two prediction tasks are reported in Table 3 and Table 4. Overall, the proposed method outperformed the *baseline* model and the *w/o. pre-training* model in both tasks. This shows that modeling and pre-training the internal states is effective in the proposed model. Furthermore, the combination of the pre-training and the fine-tuning improves the model performance. The fine-tuning makes it possible to train with a larger number of labels, which is an advantage of the use of hierarchical neural networks.

7 Conclusions

We have proposed a model that selects the utterance constructional units from the observed user behaviors by taking into account internal states such as favorable impression to interlocutors. The utterance constructional units were defined as the combination of three components: *response*, *episode*, and *question*. The proposed model is a hierarchical neural network that represents the internal states as hidden layers. The number of training labels of the internal states is limited so that we pre-trained each layer with a small number of labels one layer by one layer. Afterward, we fine-tuned the whole network with a larger number of training data of behaviors that can be objectively measured. This approach will be useful for systems with internal states that can have a small number of training data. We evaluated the system with the speed-dating dialogue corpus, and showed the proposed model achieved better prediction performance than the compared methods that did not take into account the system internal states. Although we dealt with the task for the three utterance constructional units, the proposed approach is not limited to this task.

In future work, we will implement the proposed system in a live spoken dialogue system to evaluate in real applications. It is needed to implement response generation using the prediction result of the utterance constructional units. Additionally, we plan to use multi-modal behaviors such as eye-gaze and head nodding.

Acknowledgements This work was supported by JST ERATO Grant Number JPMJER1401, Japan. The authors would like to thank Professor Graham Wilcock for his insightful advice.

References

1. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review **43**(2), 155–177 (2015)
2. Bates, J.: The role of emotion in believable agents. Communications of the ACM **37**(7), 122–125 (1994)
3. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the emotion dynamics of a multimodal conversational agent. In: ADS, pp. 154–165 (2004)
4. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International **5**(9), 341–345 (2001)
5. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., et al.: Towards an ISO standard for dialogue act annotation. In: LREC, pp. 2548–2555 (2010)
6. Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., Yoshida, N.: Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme. In: LREC, pp. 1483–1486 (2010)
7. Inoue, K., Milhorat, P., Lala, D., Zhao, T., Kawahara, T.: Talking with erica, an autonomous android. In: SIGDIAL, pp. 212–215 (2016)
8. Ishi, C.T., Ishiguro, H., Hagita, N.: Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In: IROS, pp. 2377–2382 (2012)
9. Jurafsky, D., Ranganath, R., McFarland, D.: Extracting social meaning: Identifying interactional style in spoken conversation. In: NAACL, pp. 638–646 (2009)
10. Kawahara, T.: Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS (2018)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
12. Pentland, A.S.: Honest signals: How they shape our world. MIT press (2010)
13. Picard, R.W.: Affective computing, vol. 252. MIT press Cambridge (1997)
14. Sakai, K., Ishi, C.T., Minato, T., Ishiguro, H.: Online speech-driven head motion generating system and evaluation on a tele-operated robot. In: ROMAN, pp. 529–534 (2015)
15. Schuller, B., Köhler, N., Müller, R., Rigoll, G.: Recognition of interest in human conversational speech. In: ICSLP, pp. 793–796 (2006)
16. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al.: The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Interspeech, pp. 148–152 (2013)
17. Sinclair, J.M., Coulthard, M.: Towards an analysis of discourse: The English used by teachers and pupils. Oxford University Press (1975)
18. Wang, W.Y., Biadsy, F., Rosenberg, A., Hirschberg, J.: Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. Computer Speech & Language **27**(1), 168–189 (2013)
19. Wu, C.H., Lin, J.C., Wei, W.L.: Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA transactions on signal and information processing **3**, 1–18 (2014)
20. Young, S., Gašić, M., Thomson, B., Williams, J.D.: Pomdp-based statistical spoken dialog systems: A review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)