Selecting Type of Response for Chat-like Spoken Dialogue Systems Based on Acoustic Features of User Utterances

Kengo Ohta, Ryota Nishimura, Norihide Kitaoka

Abstract— This paper describes a method of automatically selecting types of responses in conversational dialog systems, such as back-channel responses, changing the topic, or expanding the topic, using acoustic features extracted from user utterances. These features include spectral information described by MFCCs and LSPs, pitch information expressed by F0, loudness, etc. A corpus of dialogues between elderly people and an interviewer was constructed, and the results of evaluation experiments showed that our method achieved an F-measure of 49.3% in a speech segment identification task. Moreover, further improvement was achieved by utilizing the delta coefficients of each feature.

Keywords— Spoken dialogue system, Chat-like conversation, Response selection, Machine learning, Acoustic features.

I. Introduction

Recent advances in speech recognition technology have expanded the range of applications for spoken dialogue interfaces. For example, task-oriented spoken dialogue systems such as personal assistants (e.g. Apple's Siri[1], Microsoft's Cortana[2], and Google's Now[3]), which are designed to fulfill a user’s requests, are now widely used on a daily basis. On the other hand, non-task-oriented spoken dialogue systems such as conversation robots [1] (also known as chatbots) are expected to be future applications such as cognitive training or increasing opportunities of communication for elderly people. Additionally, we consider that the chat-like interface will be important for communication with humanoid robots in future. Based on a common recognition of such issues, a balanced corpus of everyday conversation are also built for analysis of the turn-taking [2]. The primary aim of such non-task-oriented conversation systems is for users to enjoy the conversation itself, thus it is more important for chatbots to be able to prolong a natural conversation as long as possible than to satisfy a user's specific demands.

In this study, we propose a method of selecting the type of system response in a non-task-oriented conversational dialogue system in a manner which is likely to continue a conversation. Our method employs a support vector machine (SVM) [3] classifier which uses acoustic features extracted from previous user utterances to select the appropriate type of system response. The acoustics of user utterances are described using spectral and prosodic features such as MFCC-based features, loudness, pitch-related features, etc.

This paper is organized as follows. We first discuss some related studies in Section II.

In Section III, we describe the development of our speech corpus and explain our response selection method. One of our goals was to develop a dialogue system for reminiscence therapy for the elderly, so we used the utterances of elderly people in this study.

We then evaluate the proposed method in Section IV and conclude the paper in Section V.

II. Related Work

Examples of research on chat-like dialog systems which do not use speech (i.e., which use text only) include Ritter et al. [4] in which a model for chat-like dialogue was created using a massive Twitter corpus of 1.3 million tweets. Vinyals et al. [5] built a language model for chat-like dialogue systems using a neural network to analyze movie subtitles and IT helpdesk transcripts. The system responds to users using common sense learned from a large corpus. This approach has been improved by introducing the objective function based on Maximum Mutual Information [6] or by introducing reinforcement learning [7]. In order to increase the duration of chat-like automated dialogue, in this study we take into consideration the atmosphere or mood of the dialogue, which is indispensable, using acoustic information from user speech. However, on a theoretical basis this study is actually an extension of the studies mentioned above.

As an example of a chat-like spoken dialogue system, Nishimura et al. [8] have been operating a spoken dialogue system called Takemaru-kun at a community center for over six years, and Lee et al. [9] have been operating a speech-oriented,

[1] https://www.apple.com/ios/siri/

[2] https://www.microsoft.com/en-us/windows/

[3] https://www.google.com/search/about/

Kengo Ohta is with the Dept. of Creative Technology Engineering, Anan National College of Technology, 265 Aoki Minobayashi, Anan, Tokushima, Japan (e-mail: kengo@anan-nct.ac.jp).

Ryota Nishimura is with the Grad. School of Technology, Industrial and Social Sciences, Tokushima University, 2-1 Josanjima, Tokushima, Japan (e-mail: ryota@nishimura.name).

Norihide Kitaoka is with the Grad. School of Technology, Industrial and Social Sciences, Tokushima University, 2-1 Josanjima, Tokushima, Japan (e-mail: kitaoka@is.tokushima-u.ac.jp).

digital information kiosk called Mei-chan for campus guidance. Although these systems can produce chat-like dialogue, system responses are chosen using template matching, so a suitable response type does not need to be selected.

As examples of methods which select response types for spoken dialogue systems using acoustic features of user speech, Osuga et al. [10] proposed a method of discriminating whether or not speakers observe turn-taking based on prosodic features of the interlocutors speech, such as fundamental frequency (F0), power, and duration, as extracted from the users utterances. Kitaoka et al. [11] used decision trees to determine system timing for back-channel responses and turn-taking. Prosodic information such as pitch and power gradients at the end of user utterances are used as features of a decision tree, and linguistic information such as the part of speech of the last word and the identity of the last content word in the last utterance are also used as features. These studies have shown that acoustic information is useful when estimating the timing of turn-taking and back-channel responses.

TABLE I
LABELS FOR NINE TYPES OF RESPONSES

Label	**Response Type**
back	Back-channel response (neutral)
p-back	Back-channel response (positive)
n-back	Back-channel response (negative)
exp	Expand on the current topic
gin-up	Ginger/Liven up the conversation
change	Change the topic
smile	Smile
emp	Show empathy
non	Do nothing (Just waiting for the user’s next utterance)

In human-to-human dialogues, we not only control the timing of our utterances, but also have to decide which types of utterances are appropriate in order to have an enjoyable conversation. In this research, our goal is to extend the duration of dialogues by choosing the appropriate type of system response from among nine types of responses (e.g., backchannel, expand the topic, liven up the topic, etc.) based on acoustic information from user speech.

III. RESPONSE SELECTION BASED ON ACOUSTIC FEATURES

A. *Conversation Corpus*

As mentioned above, one of the goals of this research was to build a reminiscence therapy dialogue system for elderly people. Thus, in order to train and evaluate our classifier for response selection, we built a conversation corpus of dialogues between elderly people and an interviewer in cooperation with a nursing faculty. All of the dialogues were recorded in a low-noise environment. In each dialogue, an elderly person speaks freely in response to ten questions (e.g., Did you go somewhere recently?) asked by an interviewer. A total of 3,062 utterances from seven speakers were collected and manually classified. Here, each utterance is a unit of speech segmented by silences of 200 milliseconds or longer. As the result of a preliminary investigation, these utterances were classified into nine categories, as shown in Table I, and all of the utterances were annotated with these labels for the supervised training of our classifier. The number of speech segments of each type for each speaker (A-G) is shown in Table II.

TABLE II
NUMBER OF UTTERANCES OF EACH TYPE FOR EACH SPEAKER

Speaker	A	B	C	D	E	F	G
back	211	396	131	307	50	256	34
p-back	77	96	62	109	27	35	9
n-back	14	49	19	17	2	15	1
exp	46	33	27	13	11	9	5
gin-up	35	19	18	21	10	7	7
change	10	8	10	10	10	10	9
smile	41	38	35	19	9	27	3
emp	15	10	8	19	7	7	2
non	87	190	64	117	33	71	15
Total	536	839	374	632	159	437	85

TABLE III
FEATURE SET

Descriptors	**Functions**
PCM loudness	position- max./min.
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff. 1/2
F0 by Sub-Harmonic Sum.	lin. regression error Q/AF0
Envelope	quartile 1/2/3
Voicing Probability	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter DDP	percentile range 99-1
Shimmer local	

B. *Acoustic Features*

Successful response selection requires a method that is able to feel the mood of a dialogue, and thus it is related to recognition of the interlocutors emotional state. With this in mind, we utilized acoustic features which were based on a standard feature set defined in the INTERSPEECH 2010 Paralinguistic Challenge[12] for response selection. A total of 1,429 acoustic features were obtained using the following four steps. First, the low-level descriptors shown in Table III were extracted at 100 frames per second and smoothed by moving average low-pass filtering. Second, their first order regression coefficients were added. Third, the nineteen functions shown in Table III were applied for each descriptor. Finally, some zero information features were discarded and two single features, fundamental frequency (F0) of onset and turn duration, were added.

C. Extension of Feature Vectors

In addition to the previously mentioned acoustic features, we also investigated the effects of extending the feature vector by referring features of the preceding utterances (i.e., use of utterance history). Two kinds of extension methods were investigated. The first method is to simply add the features of the n preceding utterances (*n* = 1, 2, 4) to the feature vector. As a result of this extension, a 1,429 * (*n* + 1) length feature vector is obtained from each utterance. The second method is to add the delta coefficients of the features calculated using the users last utterance and the preceding n utterances (*n* = 2). As a result of this extension, 1,429 * 2 = 2,858 length feature vector is obtained for each utterance.

TABLE IV
EXPERIMENTAL SETUP

Classifier	Support Vector Machine
Kernel Function	RBF Kernel
Feature Set	1,429 Acoustic Features
# Classes	9 (back, p-back, n-back, exp, gin-up, change, smile, emp & non)
Evaluation	7-fold Cross Validation

D. Response Selection using Support Vector Machine Classifier

A support vector machine classifier with a radial basis function (RBF) kernel was trained to classify the utterances of our elderly subjects into one of the nine response types (shown in Table I), using the previously described feature vectors.

IV. EVALUATION EXPERIMENT

A. Experimental Set-up

In order to evaluate our proposed method, we conducted evaluation experiments using the conversation corpus described in Section III, A. In these experiments, we used 7-fold cross validation, where the data of six speakers was used for training data and the data of the one remaining speaker was used as evaluation data. The RBF kernel functions were used to carry out the classification. Our experimental setup is shown in Table IV.

We compared the three following methods: 1) using acoustic features extracted from only the last user utterance, 2) adding the features extracted from the preceding n utterances to the features selected in method 1, and 3) adding delta coefficients between the last n + 1 user utterances to the features selected in method 2.

B. Experimental Results

Classification results for each speaker were evaluated with respect to precision, recall, and F-measure. The results using acoustic features extracted from only the last user utterance (method 1) are shown in Table V. Although there were differences in difficulty of utterance classification among the speakers, an average F-measure of 49.3% was achieved. Note that we also compared results using training labels annotated by the interviewer and by a third party annotator in preliminary experiments. However, there was no significant difference in classification performance, suggesting that variation in annotation criteria among individuals has little effect on classification performance.

We then extended the feature vector by adding features extracted from the preceding n utterances (method 2). Results when adding the one, two and four utterances preceding the last utterance are shown in Tables VI-VIII. As we can see from these results, adding the features of more preceding utterances increasingly degrades classification performance. We hypothesize that this is because the linear increase in the length of the feature vector, when keeping the number of training samples constant, makes it difficult to obtain generalization ability during learning.

On the other hand, extending the feature vector by adding the delta coefficients between the last user utterance and the preceding two utterances (method 3) improved classification accuracy, as shown in Table IX. This suggests that utilizing information from previous utterances is effective.

The confusion matrix of classification results when using our best classifier, which was the extended feature vector using delta coefficients (method 3), is shown in Table X. We can see that most of the misclassified utterances were mistakenly classified as back-channel responses. This may be because the number of utterances with back-channel response labels is greater than the number of utterances with other response labels in the training data sets. In particular, the misclassification rates between four similar response classes (back, p-back, n-back, and empathy) were notably higher than misclassifications rates between other classes. This suggests that there is some ambiguity when people generate backchannel responses to the utterances of a conversation partner as to whether these responses are neutral, positive or negative meanings.

V. CONCLUSIONS

We proposed a method for selecting the type of response by spoken dialogue systems based on the acoustic features of a user’s previous utterances, in order to extend the length of conversations with non-task oriented systems. Based on our belief that a user’s emotional state is strong indicator of what kind of response a system should make, we used acoustic features conventionally used for emotion recognition from the OpenSMILE toolkit. The features were fed to an SVM classifier, which selected one of nine types of responses. Since our target application was a reminiscence therapy system for the elderly, we performed evaluation experiments using speech data from conversations between elderly people and an interviewer. The results showed the efficacy of the use of acoustic features. Moreover, our results suggest that while it is important to refer back to the user’s utterance history by using features of some of the users preceding utterances, it is also important to suppress the dimensionality of the feature vector in order to maintain the generalization ability of the classifier. In future studies, we plan to investigate the effects of changing the amount of training data, and of using acoustic and linguistic features such as the distributed representations of words [13]. In addition, we intend to explore the use of deep neural networks or long short-term memory to capture long-term context over several utterances.

ACKNOWLEDGMENT

This work was supported in part by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Margaritis, M. Montemerlo, J. Pineau, J. Schulte et al., “Towards personal service robots for the elderly,” in Workshop on Interactive Robots and Entertainment (WIRE 2000), vol. 25, pp. 184, 2000.

[2] Hanae Koiso, Tomoyuki Tsuchiya, Ryoko Watanabe, Daisuke Yokomori, Masao Aizawa, Yasuharu Den. “Survey of conversational behavior: Towards the design of a balanced corpus of everyday japanese conversation, ” Proc. of LREC, pp. 4434-4439, 2017.

[3] V. Vapnik, The nature of statistical learning theory. Springer science & business media, 2013.

[4] A. Ritter, C. Cherry, and B. Dolan, “Unsupervised modeling of twitter conversations,” Proc. of NAACL, pp. 172–180, 2010.

[5] O. Vinyals and Q. Le, “A neural conversational model,” arXiv preprint arXiv:1506.05869, 2015.

[6] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, “A diversity-promoting objective function for neural conversation models,” Proc. of NAACL-HLT, pp.110-119, 2016

[7] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, “Deep reinforcement learning for dialogue generation,” Proc. of EMNLP, pp.1192-1202, 2016.

[8] R. Nisimura, A. Lee, M. Yamada, and K. Shikano, “Operating a Public Spoken Guidance System in Real Environment,” Proc. of Interspeech, 2005.

[9] A. Lee, K. Oura, and K. Tokuda, “MMDAgent: fully open-source toolkit for voice interaction systems,” Proc. of ICASSP, pp. 8382–8385, 2013.

[10] T. Ohsuga, M. Nishida, Y. Horiuchi, and A. Ichikawa, “Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue.” Proc. of Interspeech, pp. 33–36, 2005.

[11] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, “Response timing detection using prosodic and linguistic information for humanfriendly spoken dialog systems,” Transactions of the Japanese Society for Artificial Intelligence, vol. 20, no. 3, pp. 220–228, 2005.

[12] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Muller, S. S. Narayanan ¨ et al., “The INTERSPEECH 2010 paralinguistic challenge,” Proc. of Interspeech, pp. 2795–2798, 2010.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Proc. of NIPS, pp.3111-3119, 2013.

Kengo Ohta received his B. S. and M. S. and Ph. D degrees from Toyohashi University of Technology. He is now a Lecturer in Anan National College of Technology.

Ryota Nishimura received his B. S. and M. S. and Ph. D degrees from Toyohashi University of Technology. He is now a Specially Appointed Researcher in Tokushima University.

Norihide Kitaoka received his B. S. and M. S. degrees from Kyoto University. In 1994, he joined DENSO CORPORATION. In 2000, he received his Ph. D degree from Toyohashi University of Technology (TUT).He joined TUT as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. He was an associate professor in Nagoya University from 2006 to 2014. Since 2014, he has been a professor in Tokushima University.

TABLE V
CLASSIFICATION RESULTS USING ACOUSTIC FEATURES EXTRACTED FROM THE LAST USER UTTERANCE

Speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**Average**
Precision	0.463	0.512	0.439	0.459	0.424	0.554	0.499	0.479
Recall	0.485	0.588	0.464	0.521	0.447	0.596	0.458	0.508
F-Measure	0.474	0.547	0.451	0.488	0.435	0.574	0.478	0.493

TABLE VI
CLASSIFICATION RESULTS USING EXTENDED FEATURE VECTOR (ADDING ONE PRECEDING UTTERANCE)

Speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**Average**
Precision	0.483	0.485	0.421	0.422	0.395	0.550	0.467	0.460
Recall	0.485	0.569	0.459	0.506	0.422	0.567	0.460	0.495
F-Measure	0.484	0.524	0.439	0.460	0.408	0.558	0.463	0.477

TABLE VII
CLASSIFICATION RESULTS USING EXTENDED FEATURE VECTOR (ADDING TWO PRECEDING UTTERANCES)

Speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**Average**
Precision	0.492	0.489	0.400	0.402	0.380	0.561	0.475	0.457
Recall	0.481	0.556	0.440	0.501	0.398	0.546	0.475	0.485
F-Measure	0.486	0.520	0.419	0.446	0.389	0.553	0.475	0.471

TABLE VIII
CLASSIFICATION RESULTS USING EXTENDED FEATURE VECTOR (ADDING FOUR PRECEDING UTTERANCES)

Speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**Average**
Precision	0.450	0.448	0.411	0.411	0.327	0.538	0.410	0.428
Recall	0.457	0.540	0.430	0.512	0.404	0.571	0.455	0.481
F-Measure	0.453	0.490	0.420	0.456	0.361	0.554	0.431	0.453