

Prototype of Conversation Support System for Activating Group Conversation in the Vehicle

Susumu Kono^(✉), Yohei Wakisaka, and Atsushi Ikeno

TOYOTA InfoTechnology Center Co., Ltd.,
6-6-20 Akasaka, Minato-ku, Tokyo 107-0052, Japan
{su-kono,yo-wakisaka,atsu-ikeno}@jp.toyota-itc.com

Abstract. This paper describes the results of our research on a prototype of the conversation support system capable of activating group conversation in a vehicle.

The goal of this system was to enable further activating a group conversation. Based on methodology used in existing technology of utterance analysis, we estimated the intentions and desires of each group member in a vehicle from their conversation, aiming to enhance the overall situation of the group members in the vehicle by providing appropriate reference information corresponding to the situation in a timely manner through a conversational agent system.

Manufacturing a prototype of the system, we verified both its operations involved in the test case and its capability to infer the intentions and desires of each group member and intervene to their conversation in an appropriate timing.

In this research, we have demonstrated that our method used for the prototype was appropriate to a practical use through. In future, we will focus on optimizing the logic and system functions of group situation estimation and the subsequent steps of providing the reference information.

Keywords: Conversation estimation · Group conversation · Utterance feature · Conversational agent · Conversation analysis

1 Introduction

The information retrieval service using speech recognition such as “Siri”¹ and “Google Voice Search”² is widely used in a smartphone in our lives, because it enables a user easily to retrieve user’s desired information by the user’s speaking keywords relevant to the desired information without inputting any characters. However, such existing services require user’s button operation with voice commands to initiate information retrieval and information provision.

Considering a user’s convenience, it is more desirable to recognize user’s intentions spontaneously from utterance contents without user’s button operation and commands, and then to intervene to their conversation at appropriate

¹ <http://www.apple.com/ios/siri/>.

² <http://www.google.com/insidesearch/features/voicesearch>.

timing. As other existing spontaneous conversation systems and corpus have been examined already in prior research [6–8], here we examined a feasibility of providing services to support a user with a spontaneous conversation.

What particularly motivated us toward this research is to make it feasible by using speech recognition and intention extraction techniques.

Our objective is to clarify the effectiveness of estimating intention of utterances from contents of a group conversation, which is also aimed to lead to an enhancement of group situation with appropriate reference information in a timely manner.

2 Related Work

2.1 Noise Reduction and Voice Separation

Noise reduction function based on a multiple microphone array combined with an adaptive postfiltering scheme has been developed and utilized. The noise reduction function is achieved by utilizing the directivity gain of the array and by reducing the residual noise through postfiltering of the received microphone signals [5,9]. The microphone array system is also helpful for voice separation speaker by speaker in the conversation by plural members.

2.2 Tension Extraction from Utterances

We define *utterance* as the smallest unit of speech of spoken language, that is a continuous piece of speech beginning and ending with a pause, *speech* as the vocal form of human communication, *conversation* as a form of interactive, spontaneous communication between two or more people, typically occurring in spoken communication, and *conversational agent* as a computer system intended to converse with humans.

In existing technology of utterance analysis, the utterance feature values like spectrum of utterance power levels have been used for estimation of member tension in previous research [1]. *Tension* is defined as mental appearances of physiological responses in this paper.

2.3 Intention Extraction from Utterances

We define *intention in spoken dialogue* as a plan or an expectation in a speaker's mind to do something that has been mentioned in their speech, and it can be estimated by comparing the text data between speech recognition results and spontaneous dialogue corpora.

Methods of intention extraction in spoken dialogue utterances have been established by prior research, and the accuracy of intention recognition has recently improved [2–4,10]. The intention of each utterance is extracted from converted text data with speech recognition based on the methods of previous research above, and topics in the conversation and desires of each member is estimated.

3 Methodology

We are studying to enhance the current voice utilizing services as mentioned in Sect. 1, and to develop the support service to the traveling group for the conversation in the vehicle. Then, as a test case, we applied this methodology of conversation support to a group of travelers who needed to make some decisions (e.g., destination, route) while seated in a car. We manufactured a prototype of our system to verify the relevant operations in the test case, and also to estimate the possibility of inferring the intentions and desires of each group member by measuring the utterance characteristics in the group conversations.

We assume to monitor the voice of the conversation by the traveling group in the vehicle continuously, and to estimate topics in the conversation and desires of each member by the extracted intention from converted text data with speech recognition. Thus, we assume to realize that a conversational agent gives a group advice and based on this estimation. The image of the support system and the conversational agent that are applied for conversation in the vehicle is shown in Fig. 1.

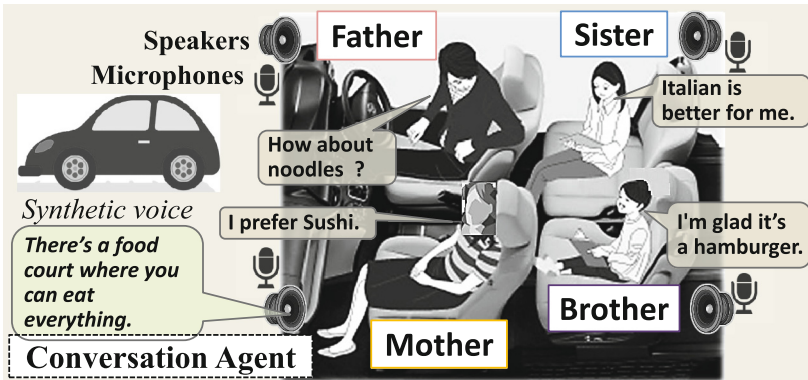


Fig. 1. The image of the applied conversation support system in the vehicle.

The methodology we use here, which is based on real-time reactions to spoken speech and its response, is not new, as noted in related work (see Sect. 2). However, our proposed method aims to go beyond previous work and recognize the group status in the conversation; it does not simply provide information to spoken speech.

3.1 Pre-processes of Utterances

We apply the synchronized microphone array system for the noise reduction and the voice separation in our prototype. For the conversation analysis, at the first step, noises such as the sound of the engine or the ventilation of the air-conditioner have to reduce from the input voice signal, and to recognize a voice

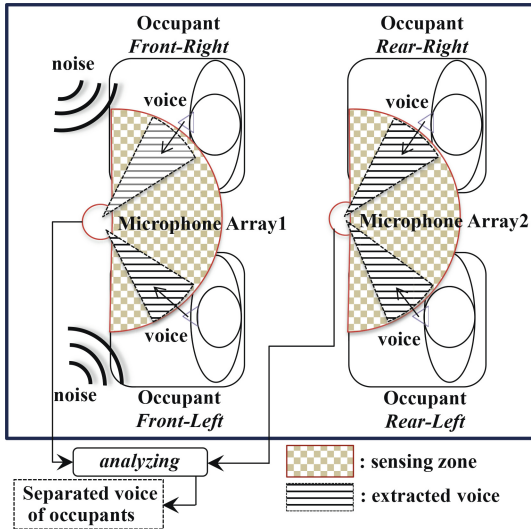


Fig. 2. The image of voice separation process.

speaker by speaker separately by using the microphone array, even in the case of the overlapping of utterances at the same time by plural members. The image of voice separation process is shown in Fig. 2.

3.2 Extraction of Intentions in Utterances

After pre-processes above, utterance sections are detected and each utterance is recognized. Based on the estimated intention by a conversation support system, we assume to choose the information which may be matched with group members from travel information database, and to be informed it to members by a synthetic voice in the vehicle. The flow to the intervention in the conversation is shown in Fig. 3.

After the speech recognition process, “subject (theme)” (e.g., the place to eat lunch, the gift to buy in the shopping mall), “category” (e.g., meal, shopping), “sentence style” (e.g., positive, negative, interrogative), “intention” (e.g., proposal, question, agreement, opposition), and “expected action” (e.g., decision on where to have lunch, searching the shop to buy gifts) are extracted by comparing the text data from the speech recognition results and spontaneous dialogue corpora data. For example, in the case of “Let’s talk about where to have lunch,” we have the following values: subject; lunch, category; meal, sentence style; positive, intention; proposal, and expected action; decision on where to have lunch.

The politeness level (e.g., polite words used toward elders or superiors in formal speech, or informal terms used toward friends in casual speech) and the existence of childish words (e.g., words often used by children) or instructional words (e.g., words obligatorily used for instruction by someone in a position of

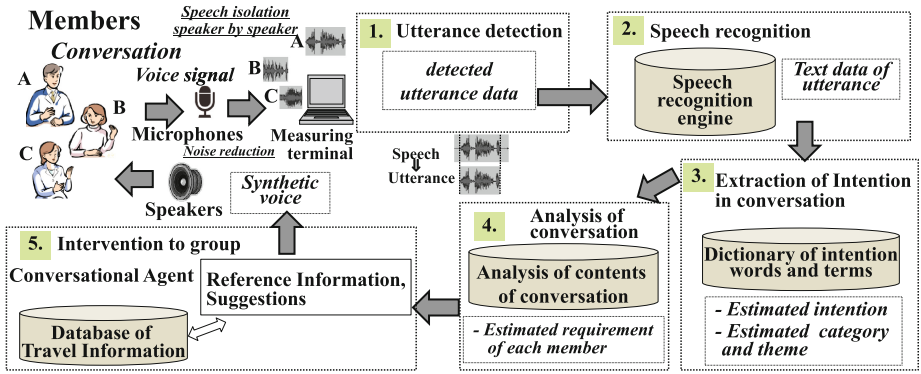



Fig. 3. Flow to intervention in the conversation.

authority) of each utterance are also estimated by comparing them with the dictionary of specific words and terms.

3.3 Extraction of Tensions in Utterances

The tension strength in each utterance is estimated by the utterance feature values (see Sect. 2.2). By combining the extracted intentions of utterances and the indicated tension strength of each utterance, we could identify specific phenomena, such as high tension in negative replies. The proposed method aims to recognize changes in group status through careful monitoring and estimation. The example of member’s status extraction from utterance data is shown in Table 1.

Table 1. Example of member’s status extraction from utterance data.

Input	Extracted items	
Voice signal of utterance 	Utterance feature value -tone -speed -power level -length -times -overlaps	Tension in utterance -in positive reply -in negative reply etc.
Text data of utterance Thank you.	Members’s status in conversation	-intension -theme, category -polite word -instruction word -childish word

3.4 Identification of Utterances in a Single Conversation

A series of utterances in a single conversation that have the same theme can be identified by checking the estimated theme of each utterance. Then, the theme and category of the whole conversation (e.g., the place to eat lunch, the gift to buy in the shopping mall) are estimated.

If it is recognized that the theme has clearly changed, this can be identified by the estimated theme of each utterance. Time course information and location information (e.g., GPS position) are also used for this identification, especially in cases where themes are unable to be identified clearly. For example, “long interval after previous utterances” is recognized as a time course information, and “place of destination in the conversation” is recognized as a location information. These are useful for identification. The example of identification of utterances in the conversation is shown in Table 2.

Table 2. Example of identification of utterances in conversation.

	Text	Intention	Category	Theme
Utterance-1	A: “Let’s talk about place of <u>lunch</u> .”	bringing up	meal/lunch	↑ Conversation-1 - Lunch - meal - Kamakura ↓
Utterance-2	B: “How about the <u>Italian</u> in <u>Kamakura</u> ?”	proposal	meal/Italian	
Utterance-3	C: “I prefer the local <u>seafood</u> restaurant rather than <u>Italian</u> .”	opposition /proposal	meal/seafood meal/Italian	
	⋮			
Utterance-7	A: “OK, it was <u>decided</u> . We <u>will go</u> to a <u>meal</u> there.”	finalization	meal	
Utterance-8	A: “By the way, how is the <u>weather forecast</u> tomorrow?”	question	weather (forecast)	↑ Conversation-2 - weather - tomorrow ↓
	⋮			

____ : “Keyword” for extraction of intentions / categories/ themes

3.5 Providing Information

Our proposed conversational agent system mechanically provides appropriate and timely reference information that takes into account the wishes of each member according to the topic of group discussion. The members wishes are estimated from the result of intention extraction mentioned above, and location information (e.g., GPS position) are also used for this identification. In this way, the system helps to bring every member’s opinion into the conversation, leading to greater satisfaction in the group conversation. In the case of conflict the desires of members, the conversational agent can choose based on the profile of members (e.g., precedence for senior/junior member) or the sequence of seated position.

4 Preliminary Experiment

The prototype system which has the basic function to analyze and intervention to the conversation was developed, and the operation of the prototype system was tested by monitors.

Table 3. Results of the preliminary experiment.

Conversation #	Member	Numbers of utterances	recognized utterances including target words (value, rate)	Extraction of utterances (value, rate)in conversation
1	A-1	4	4/4 100.0%	4/4 100.0 %
	A-2	3	3/3 100.0%	3/3 100.0 %
2	B-1	4	2/4 50.0%	2/4 50.0%
	B-2	3	2/3 66.7%	2/3 66.7%
3	C-1	4	4/4 100.0%	4/4 100.0 %
	C-2	3	3/3 100.0%	3/3 100.0 %
4	D-1	4	2/4 50.0%	2/4 50.0%
	D-2	3	2/3 66.7%	2/3 66.7%
Average	–	3.5	78.6%	78.6%

We implemented two kinds of test dialogues (approximately three minutes) with two-member groups (age: 29–51, all males) in November 2015. The theme of the test dialogues were a place of lunch in the downtown of Tokyo as decided beforehand, and conversation was began by expressing preferred food category of each member, and conversational agent intervene to the conversation and inform related restaurant information to each member. In the preliminary experiment, we required to monitors to speak loudly and not speedy, though the function of voice recognition we prepared does not have the availability to deal high speed speech or low voice at the current stage. The results of the preliminary experiment with our prototype system are shown in Table 3 and below;

1. We assume that we will apply the proposed model to group discussion in a car. However, we have not prepared the appropriate noise reduction system for a car driving certain speed as yet. Then, we implemented the test in the stopped car, which is hardly affected by noise at all. In the next step, we will prepare the appropriate noise reduction system, and implement the test in a car driving certain speed.
2. The proportion of speech recognition of words for extraction of intention was 78.6% on average. We assume that this is not a high proportion, but it can be used to extract the intention of almost all conversations.
3. The intention of utterance could be also extracted in 78.6% of all utterances. However, we could identify all utterances in conversation through the extraction of intention and time course information.
4. The extraction of the utterance feature values was quite successful, and we could calculate the strength of the tension using these utterance feature values. We confirmed that the strength of the tension was correctly estimated through human monitoring, excluding any utterances that did not have sufficient length or power to judge the tension.

5. The estimated intention of each utterance could also be used for comparisons with the calculated strength of the tension in conversation, and the strength of the tension could be linked to the intention of each utterance. Thus, we could confirm the possibility of identifying specific phenomena, such as high tension in negative replies.
6. We implement the simple function of informing, which is just picking up the restaurant information related with keywords in the utterance from database and just reading out the corresponded article. At the moment, it took about ten seconds until reading out after speaker uttered the keywords. We plan to further optimize this method by continued testing with many additional kinds of utterance data in future work.
7. The results of our preliminary test show that the basic concept of our proposed method, as outlined above in items 1–6, is generally appropriate.

The proposed system can obtain all utterance feature values and combine them with the extracted intention of each utterance. Thus, we can say that the inferring of the classification and status of groups by measuring the utterance characteristics of their users is possible, as shown by our preliminary experiment.

We also confirmed the possibility of measuring the utterance characteristics of group members as well as a method of providing suitable and appropriate information to group members using our system.

The result of the feasibility test shows basic availability of the noise reduction, voice separation speaker by speaker, speech recognition, extraction of intention and providing the related information timely was confirmed by our prototype system in the specific condition such as the conversation with the voice uttered clearly and loudly.

5 Conclusion

We demonstrated that the conversation support system to infer users' intentions and desires by monitoring utterance data in a group conversation, and to enhance the overall group situation through a conversational agent system by using estimated intentions and desires of each group member. The basic availability of our method to estimate intentions and provide the reference information in the group conversation was confirmed by testing in our prototype system.

In the next step, we will improve noise reduction and speech recognition systems, and perform a test with speedy speech by plural occupants in a car driving certain speed. Even regarding the contents of provided information, we will prepare to extend wider topics for conversation in the car trip, not only about restaurant and food information, but also the related information about historical and sightseeing spots and so on.

Then, we will further verify details of our method and system through continuous field tests, by collecting many additional kinds of utterance test data and

further clarifying the appropriate parameters for estimation and information provision using machine learning. In the final step for embodying the method, we will aim to evaluate the effectivity of intervention by conversational agent with comparison of the emotion of occupants before and after the conversation.

References

1. Ariga, M., Yano, Y., Doki, S., Okuma, S.: Mental tension detection in the speech based on physiological monitoring. In: IEEE International Conference on Systems, Man and Cybernetics, ISIC, pp. 2022–2027 (2007). *psychology* 51(3), 629 (1955)
2. Eckert, W., Levin, E., Pieraccini, R.: Automatic evaluation of spoken dialogue systems. In: TWLT13: Formal semantics and pragmatics of dialogue, pp. 99–110 (1998)
3. Hodjat, B., Amamiya, M.: Applying the adaptive agent oriented software architecture to the parsing of context sensitive grammars. *IEICE Trans. Inf. Syst.* **83**(5), 1142–1152 (2000)
4. Hodjat, B., Amamiya, M.: Introducing the adaptive agent oriented software architecture and its application in natural language user interfaces. In: Ciancarini, P., Wooldridge, M.J. (eds.) *AOSE 2000*. LNCS, vol. 1957, pp. 285–306. Springer, Heidelberg (2001)
5. Kaneda, Y., Ohga, J.: Adaptive microphone-array system for noise reduction. *IEEE Trans. Acoust. Speech Sig. Process.* **34**(6), 1391–1400 (1986)
6. Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous speech corpus of Japanese. In: *Proceedings of LREC2000 (Second International Conference on Language Resources and Evaluation)*, vol. 2, pp. 947–952 (2000)
7. Shriberg, E.: Spontaneous speech: how people really talk and why engineers should care. In: *INTERSPEECH*, pp. 1781–1784 (2005)
8. Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Van Ess-Dykema, C: Dialog act modeling for conversational speech. In: *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 98–105 (1998)
9. Zelinski, R.: A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In: *1988 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-1988*, IEEE, pp. 2578–2581 (1988)
10. Zhong, G., Hodjat, B., Helmy, T., Amamiya, M.: Software agent evolution in adaptive agent oriented software architecture. In: *IWPSE 1999 Proceedings* (1999)