Is Voice Quality Enough? - Study on How the Situation and User's Awareness Influence the Utterance Features

Shinya Yamada, Toshihiko Itoh, Kenji Araki

Department of Information Science and Technology
Hokkaido University, Sapporo, Japan
yamaya@media.eng.hokudai.ac.jp, t-itoh@media.eng.hokudai.ac.jp,
araki@media.eng.hokudai.ac.jp

Abstract

This paper presents the characteristic differences of linguistic and acoustic features observed in different spoken dialogue situations and with different dialogue partners: human-human vs. human-machine interactions. And it also presents influences of awareness of users on those characteristics. We compare the linguistic and acoustic features of the user's speech to a spoken dialogue system and to a human operator in several goal setting and destination database searching tasks for a car navigation system. Because it is not clear enough whether different dialogue situations and different dialogue partners cause any differences of linguistic or acoustic features on one's utterances in a speech interface system , we have performed experiments in several dialogue situations[4]. However, in these experiments the conditions such as voice quality and awareness of users such as impressions on the partner and prejudices against a system have not been considered. And so we collected a set of spoken dialogues in new dialogue situations. To investigate influence of voice quality, we also prepare recorded voice for response of dialogue partners and compared the influences of voice (natural voice, synthetic voice and recorded voice). We also made users answer questionnaire before and after the experiments and investigated characteristic differences caused by awareness of users. Additionally, in order to confirm the usefulness of the results of all experiments, we actually applied acoustic features of users' utterances and identified the utterances made to a system.

1. Introduction

In recent years, some spoken dialogue systems have been developed and used as practical applications as car-navigation systems or robots. However, their usability contains some problems such as inability to recognize spontaneous speech, difficulties in dealing with spoken corrections, and understanding the changes of dialogue turns. These take place because speech recognition and interpretation systems are easily affected by the user's speech style, so it is important to analyze user's behavior in different circumstances in order to improve the system performance.

In the study of user's behavior, Amalberti et al.[1] compared behavior of two groups (those who talk with a computer and those who talk with an operator) using WOZ-system. They reported that with a computer, the users used simpler expressions, used fewer words per dialogue, and regarded the system as a tool. On the contrary, the users talked to the operator in a cooperative manner. In the study of user's behavior with the circumstance considered, Itoh et al.[2] prepared two dialogue partners (a spoken dialogue system and a human operator) and concurrent task (a car-driving task) as dialogue situations. They combined the partners and situations, and compared linguistic and acoustic features of user's utterances under each condition. As a result, they reported that some acoustic and linguistic features were affected by whether the dialogue partner is a human or a machine, while some acoustic features alone were affected by a concurrent task. Itou et al.[3] also investigated dialogue characteristics in different communication modes. As a result, they also reported that some acoustic and linguistic features were affected by whether the dialogue partner is a human or machine. In their studies, however, some conditions were different in each situation. We researched the influence of difference of dialogue partner, response ability and speech recognition rate with other conditions identical to theirs. As a result, we reported that response ability mainly affects linguistic features and recognition rate affects both linguistic and acoustic features[4]. However, the factors such as voice quality or rhythm of dialogue are not considered and it is not clear what kind of awareness affects user's utterances.

In this study, in order to investigate the influences of voice quality and awareness of users, we recorded new dialogues considering voice quality and added them to the previous experimental dialogue data. We compared all of dialogue data and researched the influence of voice quality. Additionally, by using all dialogue data of previous and this experiment, we investigated how the awareness of users such as prejudice and impressions vary users' utterances.

2. Change of Linguistic and Acoustic Features

2.1. Dialogue Data Collection

For investigating linguistic and acoustic properties and variability in different situational contexts in a task-oriented spoken dialogue, we collected user's utterances through a series of experiments. In the study we performed before, the experiments were prepared in two patterns; one is that we have a dialogue partner whose speech recognition rate is 100% (EXP1) and the other is that we have a dialogue partner whose speech recognition rate is about 80% (EXP2). This is because speech recognition rate of human is nearly 100% while in the case of current spoken dialogue systems it ranges from 70% to 85%. Synthetic voice was used as system and operator voices in both experiments in order to restrict the partner's response ability. In our new study, we prepared situations where recorded voice was used as response of system

in these two patterns (EXP3). The dialogue task simulates voice control of a car navigation system, where one should perform goal setting by entering several goal names along a driving route. 12 kinds of driving route setting scenario are prepared. Each scenario includes three goals and one of them (a hotel, a coffee shop and so on) must meet conditions of scenario and be looked up in the destination database. The subjects should memorize all the goals before each task and convey those names to the dialogue partner during the task.

In our new experiment (EXP3), we prepared two new dialogue situations which are shown in table 1 as "PR+DT" and "WR+DT". And in order to compare under equal condition of speech recognition rate, we recorded dialogues in these two dialogue patterns to each situation where the recognition rate is 100% or about 80%. As for the dialogue partners, we prepared two patterns in EXP3. One is a spoken dialogue system and the other is a human operator whose response ability is limited to the extent a spoken dialogue system can do. Hereinafter, it is called Pseudo-system.

Table 1: *Dialogue situations*

Partner		Operator			Machine	
Voice		Natural	Synthetic	Recorded	Synthetic	Recorded
Driving task(DT)	No	O	PS	-	WS	-
	Yes	O+DT	PS+DT	PR+DT	WS+DT	WR+DT

O:Operator P:Pseudo-system W:WOZ-system
S:Synthetic voice R:Recorded voice

In car-driving task, subjects have to drive a car simulator at 100km/h along an oval-course. For achieving more realistic driving, subjects are asked before experiment to start over again if they fail driving.

For realization of constant speech recognition rate, a human listens to users' utterance in all the situations and make a wrong response on purpose based on the constant probability. And the response content of all partners is the same to prevent the utterances from being affected by the content.

WOZ-system is used as a substitute for a spoken dialogue system. An operator for a role of Wizard listens to user's utterance and chooses an adequate response prepared previously. The response is conveyed to the user in synthetic or recorded voice and in this new study, we used recorded voice and recorded dialogue. The users were explained that this system could accept any type of utterance and they could say whatever was associated with goal setting. By using WOZ-system, the speech recognition performance becomes controllable.

The Pseudo-system was operated by the same person as the one who operates WOZ-system and by using the same application as WOZ-system. Thus, the only different point between Pseudo-system and WOZ-system is that the users know if the dialogue partner is a human or a machine.

The subjects were 12 male students and at the beginning of both experiments, they were made to practice driving in order to decrease the influence of their experience. And to keep results less affected, the order of dialogue situations and scenarios was decided at random.

2.2. Results of the Previous Study[4]

From our previous study, we achieved following results. The dialogue partner, especially partner's response ability, affects mainly linguistic features of users' utterances. If the ability is poor, users tend to speak more briefly. And psychological load affects only acoustic features. With psychological load for example driving, users' voice become loud and high. The partner's speech recognition rate affects both linguistic and acoustic features of users' utterances. If the rate becomes low, users tend to speak loudly and intonation of their utterances becomes strong in the case of synthetic voice. On the other hand, they tend to speak politely and intonation of their utterances becomes strong in the case of natural voice.

2.3. Changes of Features Caused by the Different Dialogue Partners

Table 2: *Linguistic Features of EXP3*

Dialogue Situation	PR+DT		WR+DT	
Recognition Rate	80%	100%	80%	100%
Tasks	24	24	24	24
Utterances	253	238	257	239
Utterances/Task	10.54	9.92	10.71	9.96
Words/Utt	3.34	3.53	3.27	3.68
Filled pauses/Task	1.25	1.96	1.21	2.46
Keywords/Utt	1.75	1.83	1.73	1.87
New-keywords/Utt	1.72	1.80	1.66	1.86
Omitted verbs	128	123	140	120

Table 3: *Acoustic Features of EXP3*

Dialogue Situation	PR+DT		WR+DT	
Recognition Rate	80%	100%	80%	100%
Start pause(sec)	0.66	0.84	0.71	0.55
Duration(1)(sec)	1.30	1.42	1.38	1.49
Duration(2)(sec)	1.02	1.05	1.00	1.09
Speaking rate(1)	7.62	7.79	7.39	7.54
Speaking rete(2)	9.77	10.00	9.83	9.99
RMS mean	1471	1161	1543	1221
RMS max	4018	3012	4051	3382
Pitch mean	140.6	136.0	140.0	134.7
Pitch min	61.3	61.6	63.1	64.0
Pitch max	299.8	308.4	268.9	285.3
Pitch S.D.	22.1	22.0	21.4	19.4

Table 2 and table 3 show the statistics of linguistic and acoustic features, which were separately calculated for each of four different dialogue situations. At the beginning, we compared the features between dialogue situations or dialogue partners in EXP3. According to the results of the T and F statistical tests, no difference was found. The study which we performed before reported that the factors such as poor quality and bad rhythm of synthetic voice make users feel as if they speak to a machine and these factors affect users' utterances. The results of this research show that the equal response capability makes no difference in user's utterances. This fact suggests that voice quality is not very important for users to feel a system to be a human and it is no matter to user's utterances whether the partner is a human or a system. There are two factors left to affect user's utterances, which are rhythm elements of a dialogue and prosody of utterances. We think the flexible response ability is important to realize natural talk which includes two factors mentioned above.

Next, we compared the features between EXP1 and EXP3 (100%) or between EXP2 and EXP3 (80%) in each condition. By this comparison, we could investigate influence of voice used for response. According to the results of T and F statistical tests, fol-

lowing characteristics were found (by comparing only recorded voice with natural voice): decrease of the use of filled pause in case of both partners ($p < 0.01$); increase of start pause in case of both partners ($p < 0.05$);

From these results, it is found that the users tend rather to wait for the end of partner's response than to insist on their turn of utterance in the case of recorded voice. It is possible that recorded voice impresses the users that the partner can respond only in restricted expression and the user's turn is clear. We think the reason of such result is because the impression makes users feel that insistence on their turn is not necessary.

According to comparisons of recorded voice with natural voice and recorded voice with synthetic voice: decrease of the use of verbs in the case of both partners ($p < 0.01$); decrease of the words number in the case of both partners ($p < 0.1$);

These results indicate that in the case of a system using recorded voice, user speaks more briefly than in the case of a human or a system using synthetic voice. This may be because of the imbalance coming from the fact that quality of recorded voice is better but its rhythm is poor what makes users feel uncomfortable and gives an impression of speaking to a machine. Alternatively, it may take place because the recorded voice itself (which never consists of a part of sentence but the complete sentence) gives impression that the system has an inability to respond flexibly and users come to speak in a brief way. Thus, to realize a natural talk, we think that it is important to improve the rhythm of a dialogue.

2.4. Changes of Features Caused by the Different Recognition Rate

In the EXP3, where recorded voice is used as partner's responses, we investigate what influence the partner's recognition rate has on user's utterances. According to the result of the T or F statistical tests, the following facts became clear: there is no change of linguistic features; RMS mean increases in the case of low recognition rate;

As for linguistic features, no change was found but we observed a tendency that users were inclined to decrease keywords per utterance in the case of low speech recognition rate. This tendency was found in the study which we did before and was observed also in current result. As for acoustic features, if speech recognition rate becomes low, RMS tends to increase, which was observed in the case of synthetic voice in the previous study. Thus, it seems that changes of linguistic and acoustic features in the case of recorded voice are the same those in the case of synthetic voice.

3. Influences of Prejudices against a System on Utterances

We have reported that the partner's response abilities affect user's utterances. We think that user's awareness also relates with their utterances and it is possible that the types of awareness such as prejudices also varied user's utterances. Consequently, we made users answer a questionnaire about their awareness and investigated its influence on their utterances.

Table 4: *Distribution of User*

Evaluation	Low	Middle	High
intelligence	20	13	2
recognition ability	19	12	4
tempo	24	8	3

In the EXP1, EXP2 and EXP3, we made the users answer questionnaire about their prejudices against a system - the prejudices which they used to have before experiments. The questionnaire items are about intelligence, speech recognition ability and the tempo of conversation of a system. We made users rate the items from one to seven and told them to give rate 5 to average human. We classified users into three groups based on the value which they answered. One is a group called "Low" which consists of those who answered from one to three. The second is a group "Middle" which consists of those who answered four or five and the last is a group "High" which consists of those who answer six or seven.

Table 4 shows the number of users in each group, which includes all users of each experiment. From the number of member in each group we can see that users generally think that some abilities of a system are the same level or poorer than ordinary people. And it is also found that users tend to think that tempo of conversation is worse than capabilities to understand speech such as intelligence and recognition ability.

Next, to investigate how the prejudice against a system affects users' utterances, we sum up only utterances to the system in each group and compared the features of those utterances between groups. Because in this study the number of users in group "High" is not enough to investigate, we compared groups "Low" and "Middle". As for an item of intelligence, according to the results of the T or F tests, following characteristic were found for the group "Low" in comparison to the group "Middle": decrease of keyword per utterance ($p < 0.01$); decrease of new-keyword per utterance ($p < 0.01$); decrease of the use of verbs ($p < 0.01$); increasing tendency to convey the needed information in shorter portions - users tended to separate words or use shorter utterances ($p < 0.01$);

From these results, it is found that if users think that a partner lacks intelligence, they tend to omit the verbs and to speak in fewer keywords and briefly, which indicates that they change their way of speaking according to the intelligence level of the partner. Additionally, a tendency that the users speak to a system more slowly also observed. This result means that users try to speak in order to make the system understand their utterance easily if they think it has poor intelligence.

As for the ability of speech recognition, we also performed the T or F tests as already stated. As the result, no change of utterance features was found but a tendency that speaking rate becomes low was observed. To make a reason of this tendency clear, we researched the utterances of EXP1 and those of EXP2 separately and it was found that only the users of the group of EXP2 decreased their speaking rate ($p < 0.1$). This fact indicates that users speak slowly if the partner actually mishears, which we think is because users try to make it easy to understand their utterances.

Following characteristics about tempo of conversation are found from results of T or F tests: decreasing tendency to convey the needed information in shorter portions - users tended to separate words or use shorter utterances ($p < 0.1$); decrease of the use of verbs ($p < 0.01$);

We think the reason is that if users think a tempo of conversation is bad in human-machine dialogue, they omit the verbs and not separate words within an utterance or use shorter utterances because they try to achieve a task as few utterances as possible.

4. Classification of Users' Utterance based on Utterance Features

We researched users' utterance features in various situations and influence of users' awareness on their utterances, and As a result, it is clear that both dialogue situations and users' awareness affect utterance features. We think that there are two ways of using these features. The one way is that by putting the factors of a system to those of a human which cause the differences of features between utterances to a human and to a system, we bring dialogue with a system close to that with a human. And the other way is that we utilize the fact that users cannot speak naturally to a system like to a human. As an example of the latter, we actually tried to identify the utterances made to a system using utterance features.

Table 5: *Used Features for Identification*

time	utterance time
	voiced utterance time
RMS	RMS mean
	RMS max
	RMS S.D.
Pitch	Pitch mean
	Pitch min
	Pitch max
	Pitch S.D.

Table 6: *Performance of Identifying Utterances Made to a System*

Utt to O	EXP1				EXP2			
Utt to W	EXP1		EXP2		EXP1		EXP2	
target	O	W	O	W	O	W	O	W
Precision	0.63	0.56	0.76	0.85	0.79	0.68	0.61	0.66
Recall	0.55	0.64	0.78	0.83	0.80	0.66	0.68	0.59
F-measure	0.59	0.60	0.77	0.84	0.80	0.67	0.64	0.62

At the beginning, we sum up utterances based on each dialogue partner in EXP1 and EXP2. We used only the utterances to Operator and to WOZ-system and acoustic features of those utterances which were used for identification (features shown in table 5). We also identified the utterances to a system in each combination of the partners, which include combinations of Operator (EXP1)-WOZ (EXP1), Operator (EXP1)-WOZ (EXP2), Operator (EXP2)-WOZ (EXP1) and Operator (EXP2)-WOZ (EXP2). We use all utterances described above as learning data or test data and carried out a 10 fold cross validations. In this experiment, we used "C4.5" as a machine learning tool.

Table 6 shows results of partner identification. From these results, it is to some extent possible to distinguish the utterances to an operator from those to a system, and especially in the case of combination where the recognition rate is different, the performance of identification becomes higher. This fact proves that difference of speech recognition rate changes users' utterances. And when these two combinations with different recognition rate are compared to each other, it is found that the performance of identification is higher in the case of comparison of Operator (EXP1)-WOZ (EXP2) than that in the case of comparison of Operator (EXP2)-WOZ (EXP1). This probably takes place because in the case of low recognition rate, users tend to change their utterances to a system more drastically than those to a human. From these facts, we think those utterance features are useful for various applications. Although in this experiment, we applied "C4.5" as a machine learning tool and used only acoustic features, we expect that better tool and use of linguistic or other features will achieve higher performance in identification.

5. Conclusions

In this study, we investigated the influence of voice quality and awareness of users such as impressions and prejudices in various situations on the linguistic and acoustic features. The voice quality included natural voice, synthetic voice, and recorded voice. We analyzed the overall characteristics of the linguistic and acoustic features in terms of the intra-utterance and the whole utterance statistics. As a result, in the case of the system whose voice quality is good but rhythm of the dialogue is bad, users tend to speak in more machine-friendly way. From this fact, we found that improvement of voice quality is not enough to realize natural talk and the factors such as prosody and tempo of conversation are important to realize it. Additionally, we also found that the prejudices against a system change user utterances. We also investigated usefulness of utterance features by identifying the dialogue partner using them. As a result, we could get high performance of identification process.

The features of utterances in all kinds of situations are very useful and it is possible to use them in many applications. Thus, we have to discover efficient usages of these utterance features. In our future work, we are going to identify several states of users using utterance features and other information.

6. References

[1] R.Amalbeti, N.Carbonnell and P.Falzon, "User representations of computer systems in human-computer speech interaction", Int.J. Man-Machine Studies, pp.547-566, 1993.

[2] T.Itoh, A.Kai, T.Konishi, and Y.Itoh, "Linguistic and acoustic changes of user's utterances caused by different dialogue situations", International Conference on Spoken Language Processing, pp.545-548, 2002.

[3] K.Itou, K.Fujimura, N.Kawaguchi, K.Takeda, and F.Itakura, Dialogue characteristics in different communication modes, Special Workshop in Maui Lectures by Masters in Speech Processing, 2004.

[4] S.Yamada, T.Itoh and K.Araki, "Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate", Proceeding of InterSpeech2005, pp.3393-3396, 2005.

[5] S.Oviatt, G-A.Levow, M.MacEachern and K.Kuhn, Modeling hyperarticulate speech during human-computer error resolution, International Conference on Spoken Language Processing., pp.801-804, 1996.

[6] S.Oviatt, M.MacEachern and G-A.Levow, Predicting hyperarticulate speech during human-computer error resolution, Speech Communication, Vol.24, pp.87-110, 1998.

[7] M.Swerts, D.Litman and J.Hirschberg, Correction in spoken dialogue systems, International Conference on Spoken Language Processing, pp.615-618, 2000.

[8] J.D.Lee, T.L.Brown B. Caven, S. Haake, K.Schmidt, "Does a speech-based interface for an inveigle computer distract drivers?", Proc. World Congress on Intelligent Transport System., 2000.