Estimation method of user satisfaction using N-gram-based dialog history model for spoken dialog system

Sunao Hara, Norihide Kitaoka, Kazuya Takeda

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
{naoh, kitaoka, kazuya.takeda}@nagoya-u.jp

Abstract

In this paper, we propose an estimation method of user satisfaction for a spoken dialog system using an N-gram-based dialog history model. We have collected a large amount of spoken dialog data accompanied by usability evaluation scores by users in real environments. The database is made by a field-test in which naive users used a client-server music retrieval system with a spoken dialog interface on their own PCs. An N-gram model is trained from the sequences that consist of users' dialog acts and/or the system's dialog acts for each one of six user satisfaction levels: from 1 to 5 and ϕ (task not completed). Then, the satisfaction level is estimated based on the N-gram likelihood. Experiments were conducted on the large real data and the results show that our proposed method achieved good classification performance; the classification accuracy was 94.7% in the experiment on a classification into dialogs with task completion and those without task completion. Even if the classifier detected all of the task incomplete dialog correctly, our proposed method achieved the false detection rate of only 6%.

1. Introduction

Estimating performance is a central issue in designing spoken dialog systems. Speech recognition accuracy is the most important and commonly used measure of the performance of speech recognition systems (Dybkjar et al., 2005). On the other hand, user satisfaction is one of the most important metrics for measuring the performance of integrated systems such as SDSs (Gibbon et al., 2000). An automatic performance assessment based on objective features is generally preferable to a subjective assessment from a cost standpoint. PARADISE, which is a predictive model of system performance or user satisfaction as a function of objective metrics, was proposed as a general framework for characterizing user satisfaction with spoken dialog systems and used it for evaluations (Walker et al., 1997). However, user satisfaction is not a simple function of speech recognition accuracy because the impact of a recognition error on dialog quality reflects its context. It is assumed that detecting the problematic dialogs through the assessment of dialog context is useful approach for estimating the user satisfaction.

There have been a number of studies focused on detecting problematic dialogs in Interactive Voice Responses (IVRs) installed in call centers. Walker et al. (Walker et al., 2002) proposed a problematic dialog predictor based on the *SLU-success* feature, which encodes whether the spoken language understanding (SLU) component captured the meaning of each exchange correctly. They reported binary classification accuracy of 93% using whole dialog and accuracy of 86% even if using the first two exchanges. Kim (Kim, 2007) focused on enabling on-line prediction. He proposed an N-gram-based call quality monitoring system and achieved problematic call detection accuracy of 83% after five turns. However, he used only user utterances in the modeling. Herm et al. (Herm et al., 2008) proposed a combined model of the system log with an emotion recognition result, and they reported 79% classification accuracy of problematic/non-problematic calls after only the first four turns.

The aim of this study is to construct an estimation model of user satisfaction for spoken dialog systems based on the real-world data. From the users' point of view, they can observe only the system output (their speech prompts or responses), not the system internal states. Therefore, it is reasonable that the system outputs are heavily related to the user's impression, which directly affects user satisfaction. In this paper, we propose an estimation method of user satisfaction for a spoken dialog system using an N-gram-based dialog history model. To take the domain knowledge into account, the estimation model consisting of domain-specific concepts is effective. To generalize and to make model accurately, utterances are encoded to dialog acts' level. That is, the N-gram model is trained from users' and/or the system's dialog act sequences for each user satisfaction level.

The rest of this paper consists of four sections. In Section 2, we outline the field test and the data collection of the spoken dialog corpus. In Section 3, we present formulations of the dialog data and its N-gram modeling. In Section 4, we build N-gram models for estimating user satisfaction from their dialog act sequences and evaluate them. In Section 5, we summarize the paper.

2. Spoken dialog corpus

2.1. *MusicNavi2*: music search voice interactive system

Data collection was performed through field trials with the *Musicnavi2* music retrieval system (Hara et al., 2008), with which the user can look for and play music files on PCs through a spoken dialog. The client system can be downloaded and installed on PCs and works with a server program connected through the Internet. *Musicnavi2* can upload the input speech and the system behavior log to the server with user identification so that the server can collect a huge amount of speech data automatically to make the database. The speech interface on the client was im-

	System's prompt / response and user's utterance	Act symbols
USR:	*Hello.*	USR_CMD_HELLO
SYS:	Hello.	SYS_INFO_GREETING
USR:	*"SIMON AND GARFUNKEL".*	USR_REQUEST_BYARTIST
SYS:	Do you want to retrieve songs by "Simon and Garfunkel"?	SYS_CONFIRM_KEYWORD
USR:	*Yes.*	USR_ANSWER_YES
SYS:	Now, retrieving the songs by "SIMON AND GARFUNKEL".	SYS_INFO_SEARCHBYARTIST
SYS:	60 songs were found.	SYS_INFO_SEARCHSUCCESS
SYS:	"I AM A ROCK".	SYS_INFO_SONGTITLE
SYS:	"BRIDGE OVER TROUBLED WATER".	SYS_INFO_SONGTITLE
USR:	*That song, please.*	USR_CMD_THESONG
SYS:	Now, playing the song "BRIDGE OVER TROUBLED WATER" by "SIMON AND GARFUNKEL." (The system plays the song.)	SYS_PLAY_SONG
USR:	*Stop.*	USR_CMD_STOP
SYS:	OK, the song is stopped.	SYS_INFO_STOPPED
:	:	:

Figure 1: Example of dialog and its corresponding encoded symbols.

plemented using a grammar-driven speech recognition interface with limited vocabulary, which consists of player control words, song titles, artist names, and album names of the music files stored on the user PC. Julius 3.5.3(Lee et al., 2001) is used as the speech recognition engine. An example of a dialog with the system was shown in Fig. 1.

2.2. Field test in real user environments

In the field test, all the naive subjects were given a task, that is, they used the system until they listened to at least five songs by performing at least twenty Q&A dialogs, or using the system for over forty minutes. After that, they filled out an on-line questionnaire concerning the overall impression of the system and user profile. The questionnaires obtained the items listed in Table 1. In the questionnaire, users selected their satisfaction level (Q4a and Q4b) with the spoken dialog system on a scale of 1-5 (1: extremely unsatisfied, 2: unsatisfied, 3: acceptable, 4: satisfied, and 5: extremely satisfied). Users also selected their understanding level (from Q1a to Q1d) on a scale of 1-5 (1: not understood at all, 2: not understood, 3: almost understood, 4: understood, 5: understood well), and selected their impression of the quality of the dialog (Q2a and Q2b) on a scale of 1-5 (1: very bad, 2: bad, 3: acceptable, 4: good, 5: very good). Subjective Word Error Rate (Q3) is the answer to the question, "How often do you think the system failed to understand your speech?"

These experimental data were maintained as a *Musicnavi2* database consisting of large-scale spoken dialogs with subjective usability evaluation results in real user environments. A total of 1,359 users participated in this experiment, and the sum of their usage time was about 488 hours. While raw recorded data contained a lot of unnecessary data, the data was automatically segmented by *Musicnavi2* using speech power level and zero-cross count, and we obtained about 29 hours of speech segments, corresponding to about sixty thousand utterances.

Table 1: Items collected through the questionnaire

Age	
Gender	
Marital Status	
Address	47 prefectures
Job	14 classes
Experience	8 Boolean variables
Noise Source	4 Boolean variables
Microphone Type	text
Laud-speaker Type	text
Understanding	4 metrics of 5 classes
how to use microphone (**Q1a**)	
how to use Musicnavi2 (**Q1b**)	
what words to say (**Q1c**)	
the timing of speech (**Q1d**)	
Quality of Dialog	2 metrics of 5 classes
length of words and dialog (**Q2a**)	
naturalness of dialog (**Q2b**)	
Subjective WER (**Q3**)	integer (0 to 100)
Satisfaction	2 metrics of 5 classes
as spoken dialog system (**Q4a**)	
as music retrieval system (**Q4b**)	
Good Impression	text
Bad Impression	text

2.3. Overview of *Musicnavi2* database

We used 449 subjects consisting of 278 males and 171 females who completed the tasks from the database. They were classified according to their satisfaction levels as a spoken dialog system (Q4a), and labeled 1, 2, 3, 4 or 5. Additionally, we used 69 subjects who could not complete the tasks because of dialog failures and they were labeled ϕ; note that their profiles were unknown. Therefore, we used a total of 518 subjects classified into six classes.

Due to the nature of the task and the architecture of the system, most of the utterances were isolated-word utterances

Table 2: Overview of the database by the subjects classified into six classes according to satisfaction level. "Utt./Song" is the average of utterances per song played.

Class	ϕ	1	2	3	4	5
# of						
subjects	69	38	102	107	155	47
utterances	52.2	134.5	119.7	114.9	106.5	98.4
play songs	.485	18.6	22.4	22.4	25.2	28.7
WER [%]	70.5	54.1	51.0	46.8	41.2	35.3
Utt./Song	107	7.21	5.34	5.12	4.22	3.43
Q1a	—	4.50	4.30	4.58	4.66	4.80
Q1b	—	3.89	3.93	4.23	4.32	4.62
Q1c	—	3.03	3.35	3.87	4.23	4.57
Q1d	—	3.00	3.28	3.55	4.01	4.51
Q2a	—	2.11	3.27	3.54	3.95	4.57
Q2b	—	1.39	2.46	2.93	3.32	4.13
Q3	—	68.4	46.8	42.4	29.9	25.4

of an artist name, an album name, a song title or a short command sentence. On the other hand, the task vocabulary of *Musicnavi2* often contains uncommon phonetic contexts hardly ever seen in general Japanese texts such as newspaper articles because foreign words or even newly created words are used in the song titles, the album names and artist names.

2.4. Preliminary analysis

The average of each of several performance metrics by classes are shown in Table 2. WER was the objective metric for the recognition performance, and the average of utterances per song played was the objective metric for efficiency. These objective metrics showed inversed tendencies from the satisfaction level.

Table 3 shows the matrix of Spearman's rank-order correlations between the metrics of questionnaires (Table 1). As shown in the table, satisfaction for SDS (Q4a) has a larger correlation with Quality of dialog (Q2a and Q2b) than the correlation with Subjective recognition error (Q3). Usage of the system (Q1a and Q1b) has little correlation with the satisfaction, but usage of speech input, especially the linguistic aspect of speech (Q1c and Q1d) has a relatively high correlation. These results suggested the impact of the dialog efficiency and the dialog naturalness for the satisfaction of the SDS.

3. N-gram model of dialog act sequence

In the previous section, we showed that the dialog naturalness affected the user satisfaction. Therefore, we try to estimate the satisfaction level based on dialog naturalness. Linguistic naturalness of a dialog is related to the sequence of utterances. Use of the N-gram model is a good method for evaluating it. Although word-level information is informative, a more generalized form such as dialog act is better for accurate N-gram estimation. In this section, we define the dialog act sequences and model the sequences by N-gram.

3.1. Encoding the utterances to the dialog acts

We encoded system utterances and their actions to 21 system act symbols, and encoded user utterances and their actions to 19 user act symbols. In this study, we use automatically collected features to define the user and the system dialog acts. Therefore, not manual transcriptions but automatic speech recognition results were used and thus user utterances were encoded to 'user act' symbols automatically. The user act symbols were implemented in the recognition word vocabulary of *Musicnavi2* as non-terminal symbols in the grammar, thus they were easily mapped to dialog acts combining user acts obtained from the speech recognition results. Also, the 'system act' symbols were implemented as the words in the system prompts or responses, and a dialog act consisted of a sequence of system acts.

Fig. 1 shows an example of a dialog and its corresponding encoded symbols.

3.2. Training the N-gram model

A dialog act sequence was created for every user by arranging both the system action symbols and the user action symbols in time order. A dialog act sequence $\mathbf{x}$ is denoted as follows:

$$\mathbf{x} = \{x_1, \ldots, x_t, \ldots, x_T\} \tag{1}$$

where t is the dialog turn number.

Then, we assume that user satisfaction is affected by the current dialog act and some previous acts and their arrangement. Therefore, we model the dialog act sequence $\mathbf{x}$ by using the N-gram model $\mathcal{M}_s$ for each satisfaction level s:

$$\mathcal{M} = \{\mathcal{M}_s; s = \phi, 1, 2, 3, 4, 5\} \tag{2}$$

where ϕ denotes the 'failure' satisfaction level when the user cannot complete the task as mentioned in the previous section. The probability of the dialog act sequence $\mathbf{x}$ when given the satisfaction level s, which is a likelihood, is approximated by N-gram probability as follows:

$$P(\mathbf{x}|\mathcal{M}_s) = \prod_{t=1}^{T} P(x_t|x_{t-1}, \ldots, x_{t-N-1}, \mathcal{M}_s). \tag{3}$$

N-gram models were trained with the Witten-Bell discounting method by using SRILM toolkit (Stolcke, 2002).

4. Experiments to estimate user satisfaction

The proposed model was evaluated through experiments to estimate user satisfaction from a dialog act sequence for each user. A leave-one-out cross validation was performed using the data from 518 subjects. The dialog act sequence of one subject was used for testing, and the remaining dialog act sequences of 517 subjects were used for training the model for each test. In this study, we trained between 1-gram to 8-gram models for every user satisfaction level. Moreover, we compare the models trained by the dialog sequences in three conditions: using only the system dialog acts (`SYS`), using only the user dialog acts (`USR`), and using both the system and the user dialog acts (`SYSUSR`).

Table 3: Matrix of Spearman's rank-order correlations between the metrics of the questionnaires

	Understanding				*Quality of dialog*		*Recognition Error*	*Satisfaction*	
	Q1a	Q1b	Q1c	Q1d	Q2a	Q2b	Q3	Q4a	Q4b
Q1a	1.000								
Q1b	0.540	1.000							
Q1c	0.365	0.421	1.000						
Q1d	0.290	0.354	0.655	1.000					
Q2a	0.172	0.237	0.412	0.369	1.000				
Q2b	0.114	0.122	0.309	0.355	0.537	1.000			
Q3	-0.084	-0.093	-0.332	-0.290	-0.257	-0.333	1.000		
Q4a	0.183	0.221	0.445	0.398	0.525	0.551	-0.471	1.000	
Q4b	0.075	0.113	0.358	0.289	0.476	0.387	-0.384	0.539	1.000

Table 4: Confusion matrix of user satisfaction estimation result by 3-gram model of SYS condition.

		Estimated $\hat{\mathcal{M}}$					
		ϕ	1	2	3	4	5
Actual $\mathcal{M}$	ϕ	**43**	5	7	5	6	3
	1	0	**7**	8	9	11	3
	2	1	8	**31**	16	35	11
	3	0	9	22	**23**	45	8
	4	0	8	34	29	**66**	18
	5	0	4	5	6	24	**8**

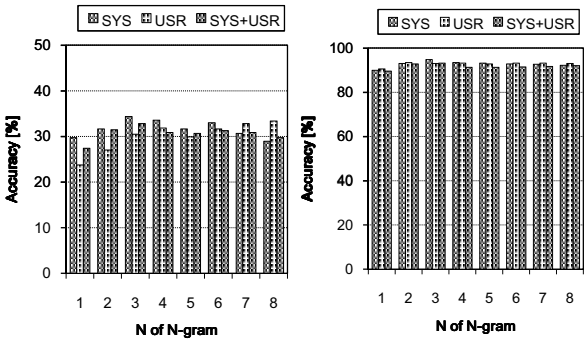


Figure 2: Classification accuracies of 6-class (ϕ, 1, 2, 3, 4, 5) experiment (left) and recalculated accuracies as 2-class (ϕ and the others) experiment (right)

4.1. Experiment to classify user satisfaction into six classes

Assuming a uniform *prior* $P(\mathcal{M})$, a maximum likelihood method was applied to classify user satisfaction into six classes and it is shown in the following equation:

$$\hat{\mathcal{M}} = \underset{\mathcal{M}_s}{\operatorname{argmax}} \{P(\mathbf{x}|\mathcal{M}_s)\} \tag{4}$$

where $\mathbf{x}$ is the input sequence and $\hat{\mathcal{M}}$ is the estimation result.

The left of Fig. 2 shows the classification accuracy. The result of the 3-gram model in the SYS condition achieved the highest accuracy of 34.4%. Table 4 shows the confusion matrix of the test result by the 3-gram model. Even the exact correct estimation was difficult, but components near the diagonal ones tended to have large value and this indicated that there are few large mistakes. Furthermore, the classification into incomplete tasks ($s = \phi$) and complete tasks (s = 1, 2, 3, 4 or 5) was performed well. The right of Fig. 2 shows the recalculated result that treated incorrect answers as correct answers even if a complete task was confused as the wrong level of complete task, that is, discrimination of comlete and incomplete tasks. As shown in the figure, the result of the 3-gram model in SYS condition also achieved the highest accuracy of 94.7%.

The result also showed that the classification accuracy was lower if using a user's utterances (in the condition USR and SYSUSR). It implies that the recognition errors caused the decrease of the model accuracy.

4.2. Experiment on estimating user satisfaction as binary classification problem

To investigate the classification performance in more detail, a binary classification method was adopted. In this paper, we focused on the performance of two classifiers, that is, 1) whether the user completed a task ($s \neq \phi$) or not ($s = \phi$), and 2) whether the user was satisfied ($s = 5$) or not ($s = 1$). An a posteriori odds (Jaynes, 2003) classifier was used and its equation was:

$$\hat{\mathcal{M}} = \begin{cases} \mathcal{M}_j & \text{if } \frac{P(\mathcal{M}_j|\mathbf{x})}{P(\mathcal{M}_i|\mathbf{x})} > 1, \\ \mathcal{M}_i & \text{otherwise}. \end{cases} \tag{5}$$

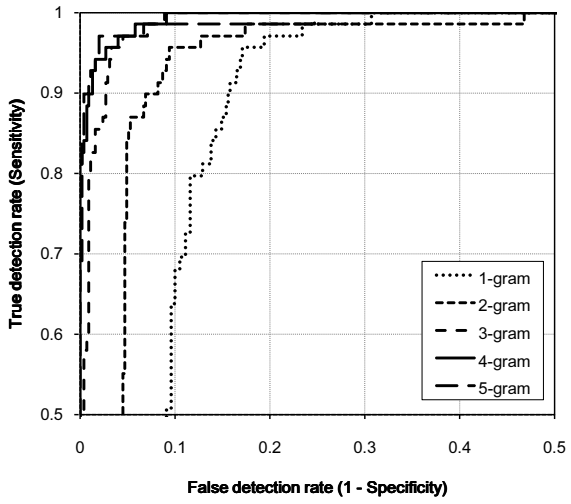
Applying Bayes' rule to Equation (5), we get:

$$\frac{P(\mathcal{M}_j|\mathbf{x})}{P(\mathcal{M}_i|\mathbf{x})} = \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_i)} \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} = \frac{1}{\alpha} \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} \tag{6}$$

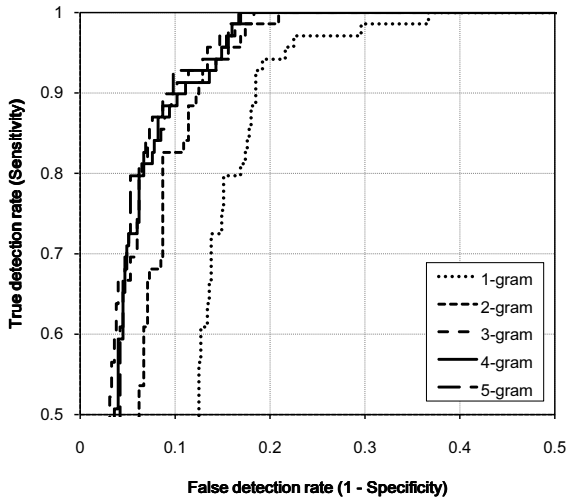
where α is an inverse of *a priori* odds. Finally, we defined the classifier as the following equation:

$$\hat{\mathcal{M}} = \begin{cases} \mathcal{M}_j & \text{if } \frac{P(\mathbf{x}|\mathcal{M}_j)}{P(\mathbf{x}|\mathcal{M}_i)} > \alpha, \\ \mathcal{M}_i & \text{otherwise}. \end{cases} \tag{7}$$

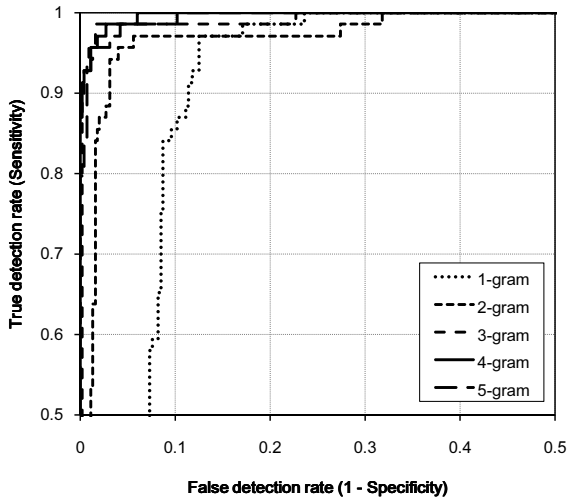
We changed the parameter α and evaluated system performance by depicting a Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC). In this study, classifier 1 was detecting that the test user was "task



(a) SYS condition

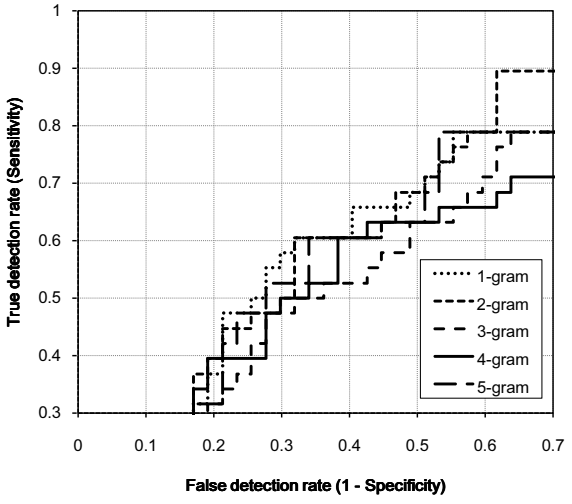


(b) USR condition

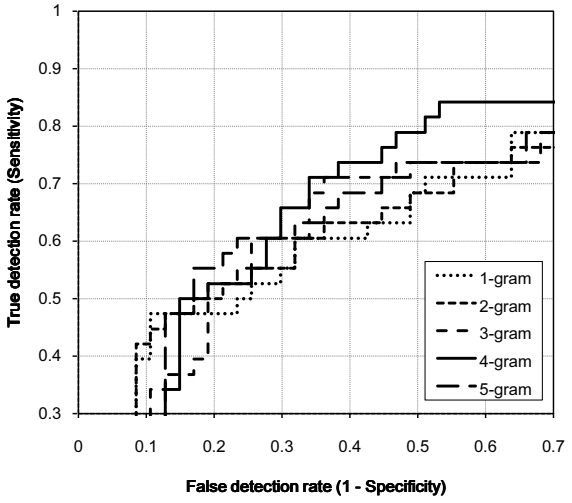


(c) SYSUSR condition

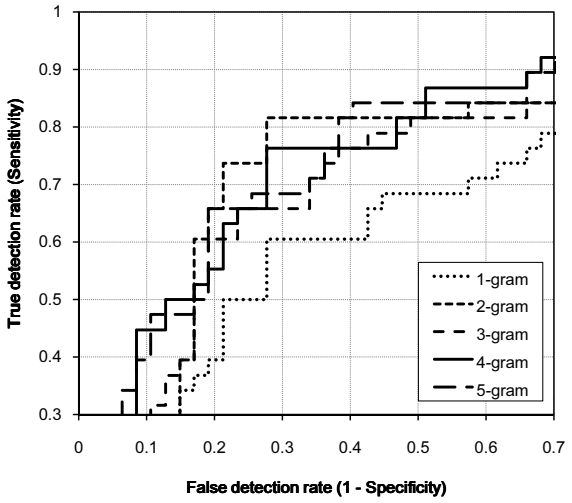
Figure 3: ROC curve for the test classified into whether the user completed the task ($s \neq \phi$) or not ($s = \phi$).



(a) SYS condition



(b) USR condition



(c) SYSUSR condition

Figure 4: ROC curve for the test classified into whether the user was satisfied ($s = 5$) or not ($s = 1$).

Table 5: Area Under the Curve (AUC) of the detection of task incomplete users

	SYS	USR	SYSUSR
1-gram	0.901	0.873	0.927
2-gram	0.948	0.929	0.977
3-gram	0.989	0.954	0.993
4-gram	0.995	0.952	0.997
5-gram	0.993	0.954	0.995
6-gram	0.989	0.951	0.995
7-gram	0.988	0.946	0.995
8-gram	0.987	0.936	0.994

Table 6: Area Under the Curve (AUC) of detection of unsatisfied users

	SYS	USR	SYSUSR
1-gram	0.611	0.638	0.619
2-gram	0.628	0.644	0.724
3-gram	0.591	0.651	0.704
4-gram	0.583	0.681	0.739
5-gram	0.629	0.662	0.739
6-gram	0.632	0.639	0.761
7-gram	0.604	0.633	0.765
8-gram	0.592	0.622	0.756

incomplete", and classifier 2 was detecting that the test user was "unsatisfied".

Fig. 3 shows the result of the classification of whether the task is completed or not. As shown in the figure, we obtained very good performance on the classification of task completion. Even if the classifier detected all of the task incomplete dialog correctly, in other words, the true detection rate was 100%, our proposed method achieved the false detection rate of only 6%. The AUC value, shown in Table 5, also indicated the highest score of 0.997 by the 4-gram model in the SYSUSR condition. Note that it seems to decrease the performance in the USR condition more than in other conditions, which is the same tendency as in the case of the experiment classified into six classes.

Fig. 4 shows the result of the classifier of whether users are satisfied or not. The figure shows the difficulty of classification of whether users are satisfied or not. However, the 2-gram model was more effective than the 1-gram model for seeing the AUC value as shown in Table 6. This result suggests the importance of considering the dialog history. Moreover, the performance of the model in the condition SYSUSR was higher than the model in the condition SYS or USR, and this fact suggested that the interaction between user and system affected user satisfaction.

5. Conclusion

An N-gram model for estimating the user satisfaction with spoken dialog systems was studied based on field trials of a voice-navigated music retrieval system. We proposed an estimation method based on N-gram models of user and system dialog act sequences. Experimental results showed good classification performance, especially the classification of whether the user could complete the task or not.

The proposed model's effectiveness was experimentally confirmed, but several future works remain. It is necessary to clarify the important dialog act contexts affecting user satisfaction through the analysis of the N-gram, and to research relationships between word error rate and estimation performance. Some keywords must be very important to estimate the satisfaction; thus we will investigate the word-dialog act hybrid estimation method. Prosodic features are also important and thus we will adopt such features.

6. References

Laila Dybkjar, Niels Ole Bernsen, and Wolfgang Minker. 2005. Overview of evaluation and usability. In *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, chapter 13, pages 221–246. Springer.

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors. 2000. *Handbook of multimodal and spoken dialogue systems*. Kluwer Academic Publishers, Boston.

Sunao Hara, Chiyomi Miyajima, Katsunobu Itou, Norihide Kitaoka, and Kazuya Takeda. 2008. Data collection and usability study of a PC-based speech application in various user environments. In *Proceedings of Oriental COCOSDA 2008*, pages 39–44, November.

Ota Herm, Alexander Shmitt, and Jackson Liscombe. 2008. When calls go wrong: How to detect problematic calss based on log-files and emotions? In *Proceedings of INTERSPEECH 2008*, pages 463–466, September.

E. T. Jaynes. 2003. Model comparison. In G. Larry Bretthorst, editor, *Probability Theory: The Logic of Science*, chapter 20, pages 601–614. Cambridge University Press, Cambrige.

Woosung Kim. 2007. Online call quality monitoring for automating agent-based call centers. In *Proceedings of INTERSPEECH 2007*, pages 130–133, August.

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. 2001. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1691–1694, September.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, pages 901–904, October.

Marilyn Walker, Diane Litman, Candace Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL 97*.

Marilyn A. Walker, Irene Langkilde-Geary, Helen Wright Hastie, Jerry Wright, and Allen Gorin. 2002. Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, 16:293–319, May.