Causal analysis of task completion errors in spoken music retrieval interactions

† Sunao Hara, ‡ Norihide Kitaoka, † Kazuya Takeda

† Graduate school of information science, Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, Japan
‡ Graduate school of information science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
† hara@is.naist.jp, ‡ {kitaoka, kazuya.takeda}@nagoya-u.jp

Abstract
In this paper, we analyze the causes of task completion errors in spoken dialog systems, using a decision tree with N-gram features of the dialog to detect task-incomplete dialogs. The dialog for a music retrieval task is described by a sequence of tags related to user and system utterances and behaviors. The dialogs are manually classified into two classes: completed and uncompleted music retrieval tasks. Differences in tag classification performance between the two classes are discussed. We then construct decision trees which can detect if a dialog finished with the task completed or not, using information gain criterion. Decision trees using N-grams of manual tags and automatic tags achieved 74.2% and 80.4% classification accuracy, respectively, while the tree using interaction parameters achieved an accuracy rate of 65.7%. We also discuss more details of the causality of task incompletion for spoken dialog systems using such trees.

Keywords: Spoken dialog, Task incompletion, Interaction parameters

1. Introduction
Estimating performance is a key issue in the design of spoken dialog systems, and the task completion rate is commonly used as a performance metric. Generally, the task completion rate is calculated based on manually labeled transcriptions of dialog data. However, automatic performance assessment (e.g. assessment based on objective features) would generally be preferable from a cost standpoint. In addition, if a spoken dialog system can estimate its own performance on-line and in real time, the system can change its dialog strategies, reducing the risk of problematic dialogs.

There have been a number of studies focused on detecting problematic dialogs in Interactive Voice Responses (IVRs) installed in call centers, e.g. (Walker et al., 2002; Kim, 2007; Herm et al., 2008; Engelbrecht and Möller, 2010). We have also proposed methods for predicting task completion errors during music retrieval tasks (Hara et al., 2010; Hara et al., 2011a; Hara et al., 2011b). These studies were aimed at the detection of problematic dialogs, however the causality of these problematic dialogs has still not been investigated sufficiently.

We based our study on the construction a model to estimate user satisfaction while using spoken dialog systems. From the users’ point of view, only the system’s output, such as speech prompts or responses, can be observed, not the system’s internal states. Therefore, a system’s outputs are assumed to strongly influence a user’s impression of the system, which directly affects user satisfaction. Previously, we proposed an estimation method of user satisfaction for a spoken dialog system using an N-gram-based dialog history model (Hara et al., 2011a). The N-gram model was trained with user and system utterance sequences, which were encoded according to the users’ and the system’s dialog utterances for each user satisfaction level. In this paper, we investigate why users could not finish their music retrieval dialog tasks. For this purpose, we analyzed corpus specifications, focusing on tag classification performance, which is the first step toward understanding overall system performance, and heavily related to our proposed method for detecting incomplete dialog problems. We then constructed decision trees for each of two features using our method; N-gram features and interaction parameter features. The trees were assumed to clarify the rank of importance of the component containing each feature.

The rest of this paper consists of four sections. In Section 2, we outline the field test used to collect spoken dialog data, and give some brief statistics regarding the data collection process. In Section 3, we present the specifications of the dialog data in terms of interaction parameters and N-gram features. In Section 4, we construct the decision trees and discuss their structures and components. In Section 5, we summarize the findings of our paper.
Corpus of Spoken dialog from music retrieval tasks

2.1 MusicNavi2: an interactive, spoken dialog music search system

Data collection was performed through field trials of the MusicNavi2 music retrieval system (Hara et al., 2010), with which users can search and play music stored on PCs through spoken dialogs. The client system can be downloaded and installed on PCs and works with a server program connected through the Internet. The MusicNavi2 client uploads the input speech and the system behavior log so that the server can automatically collect a huge amount of speech data for the database. The client’s speech interface used a grammar-driven speech recognition interface, whose the vocabulary consisting of music player control words, song titles, artist names, and the album names of the music files stored on the user’s PC. Julius (Kawahara et al., 2004) was used as the speech recognition engine. An example of a dialog using the system is shown in Figure 1. The system was used to collect speech data in a field test. Untrained subjects were instructed to use the system until they had listened to at least five songs by performing at least twenty Q&A dialogs, or until they had listened to at least five songs using the system for over forty minutes. These experimental data was preserved as a large-scale MusicNavi2 database consisting of spoken dialogs in real user environments, and included subjective usability evaluation results collected from users (Hara et al., 2008; Hara et al., 2010). A total of 1,359 users participated in this experiment, and the total usage time was about 488 hours. The raw recorded data contained many sections that were unnecessary for retrieval dialogs, so the data was automatically segmented by MusicNavi2 using speech value and zero-cross count. We then obtained about 29 hours of speech segments, corresponding to about sixty thousand utterances. Preliminary analysis of the database, e.g. recognition accuracy, SNR, speaking rate, etc. were presented in another paper (Hara et al., 2008). These analyses were focused on utterances of users. In this paper, we focused on the intermediate component of speech recognition systems, that is, the interactive dialogs. A total of 515 subjects were selected for analysis from the database. Each of them engaged in several dialogs when they used the system. We segmented and labeled each dialog from the database, and classified them into two classes: those in which the dialog resulted in completion of the music retrieval task (COMPLETE) and those which failed to complete it (INCOMPLETE). Class COMPLETE was composed of 3,720 dialogs, and class INCOMPLETE was composed of 3,828 dialogs.

Due to the nature of the task and the system architecture, most of the utterances were isolated-word utterances of an artist name, an album name, a song title, a short sentence including such proper names, or a short command sentence. It should be noted that the task vocabulary for MusicNavi2 often contains unusual phonetic contexts rarely seen in typical Japanese texts as newspaper articles because foreign words or even neologisms are used in music/song contexts.

2.2 Encoding utterances and behaviors as tags

We encoded system utterances and their actions into 19 system-related tags, and encoded user utterances and their actions into 19 user-related tags. We called them “utterance-and-behavior tags”. The system-related tags were derived from words in the system prompts or responses. User related tags, on the other hand, were obtained using two methods; automatic tags and manual tags. Since the user-related tags appeared in MusicNavi2’s recognition vocabulary as non-terminal symbols in the grammar, they were easily mapped to the tags from the speech recognition results. We called these tags automatic user-related tags. These tags clearly displayed the results of speech recognition errors. The manual user-related tags were obtained from manual transcription. These manual transcription-based tags were likely to be very helpful for tendency analysis in determining what kind of errors tended to severely affect task completion/incompletion.

We used not only utterances but also behaviors to help determine if a task was completed or not. The system tags contain events related to system behavior: playing a song (PLAY-SONG), ignoring the input because the input was noise (IGNORE-BYGMM), and ignoring the input because the trigger button was not pushed (IGNORE-BYNOTRIGGER). The user tags also contain user behaviors such as pressing the initialize button (PUSH-INITIALIZE) and input being classified as noise by the input interface (CATCHNOISE-TOOSHORT/-GMM/-ASR). Figure 1 shows an example of a dialog and its corresponding encoded tags. Tables 1 and 2 show the definitions of the 19 user tags and 19 system tags (behavior tags are noted in the tables with “*”). These tables also show the relative frequencies with which the tags appeared in the database. These tables are discussed in more detail in the next section.

3. Specification of corpora

3.1 Performance of tag classification task

The user tags were automatically encoded by MusicNavi2 as mentioned above. Its automatic encoding performance, in other words, its tag classification performance, was assumed to be related to the performance of the comple-

1366
Table 1: Relative frequency of user’s utterance-and-behavior tags in the corpus.

<table>
<thead>
<tr>
<th>Definition of User tags</th>
<th>Relative frequency of manual tags</th>
<th>Relative frequency of automatic tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>COMPLETE</td>
</tr>
<tr>
<td>REQUEST-BYARTIST</td>
<td>9.5 %</td>
<td>9.8 %</td>
</tr>
<tr>
<td>REQUEST-BYMUSIC</td>
<td>14.7 %</td>
<td>21.9 %</td>
</tr>
<tr>
<td>ANSWER-YES</td>
<td>7.3 %</td>
<td>10.8 %</td>
</tr>
<tr>
<td>ANSWER-NO</td>
<td>2.1 %</td>
<td>2.1 %</td>
</tr>
<tr>
<td>Cmd-TheSong</td>
<td>1.9 %</td>
<td>3.4 %</td>
</tr>
<tr>
<td>Cmd-NextSong</td>
<td>1.7 %</td>
<td>3.1 %</td>
</tr>
<tr>
<td>Cmd-PreviousSong</td>
<td>0.8 %</td>
<td>1.4 %</td>
</tr>
<tr>
<td>Cmd-HELLO</td>
<td>1.4 %</td>
<td>0.8 %</td>
</tr>
<tr>
<td>Cmd-MicTest</td>
<td>0.8 %</td>
<td>0.3 %</td>
</tr>
<tr>
<td>Cmd-Respeech</td>
<td>0.7 %</td>
<td>1.0 %</td>
</tr>
<tr>
<td>Cmd-BackHistory</td>
<td>0.2 %</td>
<td>0.2 %</td>
</tr>
<tr>
<td>Cmd-Stop</td>
<td>6.2 %</td>
<td>8.3 %</td>
</tr>
<tr>
<td>Cmd-Initialize</td>
<td>1.2 %</td>
<td>0.7 %</td>
</tr>
<tr>
<td>Cmd-EXIT</td>
<td>1.5 %</td>
<td>0.6 %</td>
</tr>
<tr>
<td>Push-InitializeButton</td>
<td>1.2 %</td>
<td>0.7 %</td>
</tr>
<tr>
<td>CatchNoise-TooShort</td>
<td>0.3 %</td>
<td>0.3 %</td>
</tr>
<tr>
<td>CatchNoise-Gmm</td>
<td>39.0 %</td>
<td>27.8 %</td>
</tr>
<tr>
<td>CatchNoise-ASR</td>
<td>0.0 %</td>
<td>0.0 %</td>
</tr>
<tr>
<td>Undefined</td>
<td>10.2 %</td>
<td>7.4 %</td>
</tr>
</tbody>
</table>

Table 2: Relative frequency of system’s utterance-and-behavior tags in the corpus.

<table>
<thead>
<tr>
<th>Definition of System tags</th>
<th>Relative frequency of automatic tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>PROMPT-ANYSong</td>
<td>1.4 %</td>
</tr>
<tr>
<td>PROMPT-Respeak</td>
<td>1.4 %</td>
</tr>
<tr>
<td>Suggest-SongTitle</td>
<td>12.0 %</td>
</tr>
<tr>
<td>Play-Song</td>
<td>9.3 %</td>
</tr>
<tr>
<td>Confirm-Exit</td>
<td>0.9 %</td>
</tr>
<tr>
<td>Confirm-Requested</td>
<td>10.0 %</td>
</tr>
<tr>
<td>Info-SearchFail</td>
<td>0.5 %</td>
</tr>
<tr>
<td>Info-SearchSuccess</td>
<td>2.2 %</td>
</tr>
<tr>
<td>Info-SearchByArtist</td>
<td>1.8 %</td>
</tr>
<tr>
<td>Reply-CmdHello</td>
<td>0.6 %</td>
</tr>
<tr>
<td>Reply-CmdMicTest</td>
<td>0.3 %</td>
</tr>
<tr>
<td>Reply-CmdRespeech</td>
<td>1.0 %</td>
</tr>
<tr>
<td>Reply-CmdBackHistory</td>
<td>0.4 %</td>
</tr>
<tr>
<td>Reply-CmdStop</td>
<td>3.9 %</td>
</tr>
<tr>
<td>Reply-CmdInitialize</td>
<td>1.5 %</td>
</tr>
<tr>
<td>Info-GoodBy</td>
<td>1.4 %</td>
</tr>
<tr>
<td>Ignore-ByGmm</td>
<td>12.6 %</td>
</tr>
<tr>
<td>Ignore-ByNoTrigger</td>
<td>38.7 %</td>
</tr>
<tr>
<td>Undefined</td>
<td>0.0 %</td>
</tr>
</tbody>
</table>
The N-gram, or N-order Markov chain, is a simple assumption for modeling sequential data, but is also a very useful modeling method (Kim, 2007; Engelbrecht and Möller, 2010; Hara et al., 2011a). We compared N-grams with N=1, 2, 3 and denoted them as 1-, 1-2 and 1-3 grams, respectively; e.g., a 1-3 gram represents the features constructed from the frequencies of 1-gram, 2-gram, and 3-gram. The number of unique N-grams in the entire corpus were 38, 1036, 7798, respectively. For this experiment, we calculate the number of unique N-gram tags limited by two or more occurrences, which we called N-gram tag cutoffs, for both the manual tag and the automatic tag corpora. The manual tag corpus contained 38, 780, and 4144 unique cutoff N-grams, and the automatic tag corpus contained 38, 939, and 4576 unique cutoff N-grams. These numbers correspond to the feature vector lengths, which are used to construct decision trees in a following section.

3.3. Interaction parameters

Interaction parameters were used for dialog quality evaluation (Schmitt et al., 2010; Engelbrecht and Möller, 2010; Möller, 2005), and one standardization has been established as the ITU-T Recommendation (ITU-T, 2005). In this study, we extracted 13 interaction parameters whose definitions and average parameters are shown in Table 5. Note that parameters REJ-GMM and REJ-TRG are the extended parameters for our MusicNavi2 system and its corpus. Since the parameters in Table 5 are automatically extracted from the system logs, they include errors caused by the speech recognition function of the system, as was the case with the automatic user tags.

4. Structure of decision trees

We trained decision trees to detect task-incompleteness using the WEKA toolkit (Hall et al., 2009). The trees were constructed from the frequencies of 1-gram, 2-gram, and 3-gram. The number of unique N-grams in the entire corpus were 38, 1036, 7798, respectively. For this experiment, we calculate the number of unique N-gram tags limited by two or more occurrences, which we called N-gram tag cutoffs, for both the manual tag and the automatic tag corpora. The manual tag corpus contained 38, 780, and 4144 unique cutoff N-grams, and the automatic tag corpus contained 38, 939, and 4576 unique cutoff N-grams. These numbers correspond to the feature vector lengths, which are used to construct decision trees in a following section.

Figure 1: Example of dialog and its corresponding encoded tags.
Two decision trees were constructed based on information gain, and its maximum depth was constrained to 5. Two decision trees were constructed, one for manual tags and one for automatic tags. Input features were vectors consisting of N-gram frequencies of automatic utterance-and-behavior tags for each dia-

---

We used the `REPtree` method of WEKA, which can more easily control the depth of the tree than `J48` method, with only a small decrease in classification performance.
Table 5: Interaction parameters of dialog data in corpus, their value and definitions.

<table>
<thead>
<tr>
<th>Abbreviation of parameters</th>
<th>Value of parameters</th>
<th>Definition of parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INCOMPLETE</td>
<td>COMPLETE</td>
</tr>
<tr>
<td>DD [sec]</td>
<td>62.4</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>Overall duration of dialog</td>
<td></td>
</tr>
<tr>
<td>STD [sec]</td>
<td>1.81</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>Average duration of system turn</td>
<td></td>
</tr>
<tr>
<td>UTD [sec]</td>
<td>1.75</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>Average duration of user turn</td>
<td></td>
</tr>
<tr>
<td>SRD [sec]</td>
<td>0.38</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Average delay of system response</td>
<td></td>
</tr>
<tr>
<td>URD [sec]</td>
<td>4.01</td>
<td>3.50</td>
</tr>
<tr>
<td></td>
<td>Average delay of user response</td>
<td></td>
</tr>
<tr>
<td>#TURN</td>
<td>23.6</td>
<td>26.3</td>
</tr>
<tr>
<td></td>
<td>Overall number of turns uttered in dialog</td>
<td></td>
</tr>
<tr>
<td>#STURN</td>
<td>12.7</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>Overall number of system turns uttered in dialog</td>
<td></td>
</tr>
<tr>
<td>#UTURN</td>
<td>10.9</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>Overall number of user turns uttered in dialog</td>
<td></td>
</tr>
<tr>
<td>#REJ-TRG</td>
<td>5.62</td>
<td>3.70</td>
</tr>
<tr>
<td></td>
<td>Overall number of trigger-activated rejections by user; i.e., count where user didn’t push trigger button of MusicNavi2</td>
<td></td>
</tr>
<tr>
<td>#REJ-GMM</td>
<td>1.74</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>Overall number of automatic rejections by GMM</td>
<td></td>
</tr>
<tr>
<td>#REJ-ASR</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Overall number of null results by Automatic Speech Recognition</td>
<td></td>
</tr>
<tr>
<td>#BARGE-IN</td>
<td>1.07</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>Overall number of user barge-in attempts in dialog</td>
<td></td>
</tr>
<tr>
<td>#CANCEL</td>
<td>0.07028</td>
<td>0.00030</td>
</tr>
<tr>
<td></td>
<td>Overall number of user cancels attempted in dialog; i.e., count where user pushed initialize button of MusicNavi2</td>
<td></td>
</tr>
</tbody>
</table>

4.1. N-gram feature of tag sequence

Figures 2(a) and 2(b) show the structures of the decision trees using 3-gram frequencies with manual tags and automatic tags, respectively. These trees achieved 74.2% and 80.4% discrimination accuracy by 10-fold cross validation, respectively.

These figures show that the automatic tags needed the 3-gram, but that it was not needed with the manual tags. This might be caused by differences in the complexity of the N-gram tag patterns in each corpus. Interestingly, the accuracy rate achieved using the automatic tags was the same as the rate using the manual tags. This was because the automatic tags included the misrecognition errors of the speech recognition system. In other words, although the automatic tag corpus contains only the system’s understanding of results and actions, the tags represent the actual interaction between the user and the system.

4.2. Interaction parameters

Figure 3 shows the structure of the tree using interaction parameters. This tree achieved 65.7% discrimination accuracy by 10-fold cross validation.

The results shown in the figure suggest the importance of the parameter “SRD” for task-incomplete detection. The “#CANCEL” parameter was used in the 2nd stage, however this was not important because users could restart their music retrieval dialog using the “cancel button” on the GUI of MusicNavi2. Regarding the usage of the parameters shown in Figure 3, there were 9 important parameters, namely; SRD, #CANCEL, URD, STD, #REJECT-TRG, DD, UTD, URD, and #STURN.

The N-gram features and interaction parameters could be concatenated, and the concatenated parameters could be used to construct a more accurate decision tree than if only individual features are used, but this is a topic for future work.

5. Summary

The causes of task completion errors in spoken dialog systems were analyzed using detection trees with N-gram features of the music retrieval dialog. Results using this method suggested that higher order N-grams are not particularly effective for increasing the detection rate, but also illustrated the importance of the last part of the dialog for detecting task completion.

These findings should be applied to methods for detection of task-incompletion dialogs or other problematic dialogs in future work. This study used only linguistic information, however, acoustic information must also be important. This is also a subject for future work.

6. Acknowledgements

This work was supported in part by the NEDO Grant for Industrial Technology Research Program (07A12203a) and KAKENHI (23500209).

7. References

Klaus-Peter Engelbrecht and Sebastian Möller. 2010. Sequential classifiers for the prediction of user judgments
Figure 2: Structure of decision tree based on N-gram features. Each branch contains a 1-gram, 2-gram or 3-gram. Figures in circles represent the chosen class.
Figure 3: Structure of decision tree based on interaction parameters.


