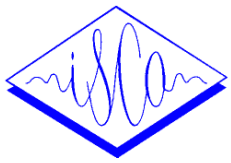

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

QUERY-RESPONSE RELATIONSHIPS IN THE OASIS SPEECH-RECOGNITION SYSTEM

B.L. Zeigler and B. Mazor

GTE Laboratories Incorporated
40 Sylvan Road
Waltham, MA 02254, USA

ABSTRACT

We have recently developed a speech-recognition system for automation of telephone-service orders, and tested the system through four months of regular use by GTE customers. In this paper, we use quantitative data (response content) to assess the effectiveness of our structured transaction model, and evaluate the extent to which natural queries yielded responses in the predicted linguistic form. Our results showed the structured model to be a successful approach to transaction automation. The results also show that customer responses to our designed queries were both predictable in form and quite limited in variety.

1. INTRODUCTION

The focus of the OASIS program at GTE Laboratories has been on developing speech interactive systems for applications in the domain of telephone-service order [1], [2]. The OASIS (for "Operations Automation Speech-Interactive Systems") program has included the development of speech recognition, knowledge processing, dialogs, speech generation, and system control to support the automation of applications within the domain. The first application targeted by OASIS was telephone-service-disconnect orders.

The OASIS disconnect order comprises seven major information elements: phone number to be disconnected, customer identification number, date the service is to be terminated, address for billing, referral service, phone number where the customer can be reached, and an order summary. A typical interaction spans an average of 15-20 queries, and lasts 3-4 minutes. During the transaction, users can say command words, either to request repetition of a query or to transfer to a service representative.

We have recently completed the implementation of the disconnect system and field-tested the system through four months of regular use by GTE customers. During the field test, customers who reached OASIS had no prior knowledge that they would be using an automated process to place their orders.

The OASIS field test provided the opportunity for a technology proof of our system, as well as a valuable opportunity to observe spoken behavior in the context of placing actual service orders. We used real-time observations, audio, and data to evaluate the performance of individual system functions and of the system as a whole. In general, the results were very encouraging - about 2/3 of the disconnect-order calls were successfully handled by the automated process. In this paper, we focus on the OASIS dialog and present results of our analyses of customers' spoken responses during OASIS interactions.

2. APPROACH

Dialogs for applications targeted by OASIS support the handling of a significant amount of information through a lengthy transaction. Our approach to dialog design includes characterization of the application in terms of its information elements, specification of the transaction flow, design of appropriate language structures, and construction of system speech. Dialog flow and language are designed in consideration of system recognition and knowledge capabilities.

OASIS transactions are constructed as a structured, progressive flow of information elements acquired through individual subdialogs [2]. System queries within the sub-dialogs follow the style of discourse in the sense that explicit direction regarding the expected content, timing, and manner of speech is generally avoided in favor of natural queries

with implicit or indirect cues. Query language and recognition vocabularies are specified synergistically and with consideration to expected relationships between queries and responses. Thus, particular challenges for the dialog developer are the textual specification of system language and the creation of associated speech that will elicit responses consistent with the predicted vocabularies.

As part of the system performance evaluation, we analyzed an audio sample of customers' responses to queries. In the following, we use quantitative data related to the content of responses to evaluate our dialog design. In particular, we assess the effectiveness of the structured transaction model, and we evaluate the extent to which natural queries yielded responses in the predicted linguistic form.

3. CONTENT ANALYSIS

Our data sample includes 3,685 response tokens from 230 service-disconnect calls placed by different customers on 21 days during the trial. Prior to analysis, we grouped responses according to the type and pragmatic context of the information sought: *digit strings* - area codes, phone numbers, and social-security number used for customer ID; *dates* - month and day-of-month; *affirmative/negative responses* - choice, verification of choice, verification of fact, and possession.

For each context group and subgroup, we assigned the responses to four categories, which we defined as follows:

Canonical responses provide only the information sought and in a construct that is considered natural within the language. For example, in response to the question *What is the area code?*, Canonical responses include *214* and *it's 214*. In response to *What day in March do you want your service disconnected?*, Canonical responses include *the 5th, March 5th, today,* etc.

Embedded responses are canonical responses that also include extraneous speech. Canonical responses with filled pauses are also classed as embedded. Examples of Embedded responses are *I'm sorry....the 13th* and *Umh...813.*

Anticipatory responses provide the information sought by the query, as well as other information. In most cases, the "other" information anticipates information sought by the subsequent query, with the target information either explicitly or implicitly stated. Alternatively, the "other" information can be a rephrasing of the target information. For example, in response to the question *Do you have the number with you?*, Anticipatory responses include *Yes, 9894521* and *4749812.* In response to the question *In what month do you want your service disconnected?*, Anticipatory responses include *March 17th, Immediately,* and *This month....December.*

Off-target responses do not provide the information sought by the eliciting query. Lack of response is included in the Off-target class.

4. RESULTS

4.1 Digit Strings

In response to digit-string queries, the large majority (986/1056 or 93%) of utterances were Canonical responses. Of those, 97% were pure digit strings, 1% included echoes of the query (*The area code is 813*), and 2% included numeric representations within the digit-string utterance (*6 2 6 seventy-six eighty*).

Only about 2% of the digit-string responses were classed as Embedded, 2% Anticipatory, and 3% Off-target. About half of the Anticipatory responses were due to inclusion of the 7-digit number with the area-code utterance, while the other half were due to inclusion of either the area code or an extension number with the 7-digit response. There were no Anticipatory responses among the social security numbers.

In addition to verbal responses, we observed DTMF responses to digit queries at the beginning of the transaction. We attributed this tendency to the customers' use of DTMF for call routing just prior to reaching OASIS.

4.2 Dates

The majority (179/206, 87%) of month responses were Canonical, and essentially all (99%) of these responses included only the name of the month. Responses not classified as Canonical were more often Anticipatory (8%) than either Embedded (~3%) or Off-target (~1%). Most (14/17) of the Anticipatory

utterances included both month and day-of-month information (*December 4th, now*) .

Like the responses to the month query, the majority of day-of-month responses were Canonical (175/212, 82%). Unlike the month responses, however, the day-of-month responses included a range of Canonical forms, including day-of month alone (*5th*, 36%), day-of-month preceded by "the" (29%), digits (14%), month-day (*February 5th*, 11%), variants on *today* (5%), and miscellaneous others (5%).

Only about 3% of the day-of-month responses were classified as Embedded, while 6% were Anticipatory, and 9% Off-target. Most of the Anticipatory utterances included day-of-week information in addition to day-of-month (e.g., *Sunday the 5th*); the exception was a re-phrasing (*the 13th, February 13th*). Off-target responses included non-lexical utterances, filled pauses, time-outs, and ambiguous day-of-week responses (*Tuesday*).

Although both digit-strings and dates show a high rate of Canonical response, the rate was clearly lower for dates. The difference between dates and digit strings is mainly due to the relatively higher rates of Anticipatory responses (Month) and Off-target responses (Day-of-month) present for dates.

4.3 Affirmative-Negative Responses

Overall, the majority (1797/1946, 92%) of affirmative-negative responses were Canonical. Of those, a substantial number (91%) were just the word *yes* or the word *no* spoken alone. Only about 1% of the responses were classified as Embedded, 3% Anticipatory, and 3% Off-target.

Analyses according to pragmatic category revealed interesting differences among affirmative/negative response types. While 96% of responses to choice and verification queries were Canonical, only 85% of possession queries elicited Canonical answers. The difference is primarily due to relatively high numbers of Off-target (7%) and Anticipatory (5%) responses to Possession queries. Off-target possession responses were related to unavailability of the target information (*I'll have to look it up...*) or semantic confusion (*Depends on what date..*) The Anticipatory answers reflect the tendency to provide the information sought in response to the possession question (e.g., the query *Do you have a new mailing address?* answered by *It's 1809 Main Street...*)

Like possession queries, verification-of-choice queries, such as *Do you mean March 16th?*, tended to elicit Anticipatory answers (5%). For verification-of-choice, the Anticipatory responses included either correction or confirmation of the to-be-verified information (*No, March 5th* ; *March 15th*).

4.4 Commands

About 2% of all utterances were *repeat* or *operator* requests. Of these, 81% were Canonical, 4% Embedded, and 15% Anticipatory. The Anticipatory command utterances were all in the form of a response to a query followed by an *operator* request.

5.0 DISCUSSION

Our challenge as developers was to create a high-performance automated version of a long and fairly complex transaction. Because the system would be used by customers, accurate and timely completion of orders was essential. We also needed to balance the tradeoffs among performance, naturalness, efficiency, development complexity and cost. The structured transaction model provided such a balanced solution. In addition, we designed the OASIS query language with particular emphasis on brevity, semantics, and prosodics (in line with suggestions of others, e.g., [4]). The resulting speech is quickly paced, semantically structured to focus on the target information, and prosodically designed to introduce turn cues only at appropriate points.

The present work had two primary objectives: (1) to assess the effectiveness of using a structured model to represent conversational transactions, and (2) to evaluate the extent to which naturally formed queries elicited responses in a predicted linguistic form.

With regard to the transaction model, our results show that customers responded cooperatively through the interaction, rarely anticipated future information, and provided on-target answers to our queries. When present, Anticipatory responses were limited to the contexts where semantically or pragmatically-related elements are generically difficult to separate. (e.g., the Month and Day-of-month components of the date.) Thus, the results, as well as our perception of the

naturalness of interactions with OASIS, clearly demonstrate the power of our structured model.

With regard to the predictability of responses, the measure of interest is the extent to which natural queries elicited responses classified as Canonical. Our overall results showed 92% of responses to be in the Canonical form, and similar results were observed in the analyses for individual information types. Thus, the great majority of responses were in accordance with our predictions concerning the appropriate linguistic and pragmatic forms. As a result, because the recognition capabilities for the service-disconnect system were developed in concert with these predictions, the vast majority of responses were compatible with OASIS recognition vocabularies.

The range of Canonical forms observed was generally rather narrow; moreover, responses were frequently observed in their simplest form. Digit strings, names of months and affirmative-negative responses tended to occur in a single form, while day-of-month responses appeared in 4-5 different forms. The narrow range of response types was also apparent in response categories other than Canonical. For example, Embedded responses were typically Canonical answers preceded by filled pauses. Thus, the customer responses during the trial were both predictable in form and quite limited in variety. Taken together, our results show that the OASIS disconnect system provided 94% lexical coverage with a vocabulary size of about 70 words.

Our findings regarding vocabulary size contrast sharply with the estimate proposed in [3] for the same type of transaction. That estimate, which was based on human-human conversations and not on an actual development, suggests that lexical coverage similar to that of OASIS would require 600 words - an order of magnitude higher than we achieved. The difference illustrates the degree of potential inaccuracy associated with methods that rely exclusively on natural conversations to develop estimates of recognition complexity and other parameters for automated processes.

The success of our approach is evident not only in the extent to which responses are predictable and canonical, but also in the low rates of disfluency that we observed. For example, filled-pauses were components of the relatively infrequent Embedded responses. Other disfluencies (mostly self-repairs) were observed in less than 1% of customer turns. Our results for disfluencies agree with Oviatt's finding that structured formats elicit far fewer disfluencies than do unconstrained formats, and that automated transactions elicit fewer disfluencies than are found in natural discourse [5]. The present data extend Oviatt's Wizard-based results for structured tasks to the context of placing real service orders with an actual speech recognition system.

In addition to disfluencies, we examined interruptions, by reviewing a subset of 45 calls for which two-way audio was available. Like disfluencies, interruptions occurred on only about 1% of the turns. The low interruption rate was consistent with our expectation that the likelihood of interruption would be substantially lower in an infrequently exercised transaction than in a highly practiced one (e.g., a book or movie ordering service).

Finally, our data regarding response forms, disfluencies, and interruptions show no evidence to support the claim that dialogs with a high degree of naturalness may "derail" by eliciting unconstrained, conversational speech [3]. In contrast, our results suggest that structured dialogs based on natural queries produce highly predictable, naturally constrained speech.

6.0 REFERENCES

[1] B. Mazor & B.L. Zeigler, "Dialog Design for a Speech-Interactive Automation System," *Speech Communication*, in press, 1995.

[2] B. Mazor, J. Braun, B. Zeigler, S. Lerner, M. Feng, & H. Zhou. "OASIS: A Speech Recognition System for Telephone Service Orders," *Proc. of ICSLP*, Sept. 1994.

[3] D. Karis & K.M. Dobroth. "Psychological and Human-factors Issues in the Design of Speech Recognition Systems," *Applied Speech Technology*, CRC. London 1995.

[4] C. A. Kamm. "User Interfaces for Voice Applications." *Voice Communication between Humans and Machines*, National Academy of Sciences. Washington 1994.

[5] S. Oviatt. "Predicting Spoken Disfluencies during Human-Computer Interaction," *Computer Speech and Language*, 1995,9.