Spoken Dialog Systems for Automated Survey Interviewing

Michael Johnston[1], Patrick Ehlen[2], Frederick G. Conrad[3], Michael F. Schober[4],
Christopher Antoun[3], Stefanie Fail[4], Andrew Hupp[3], Lucas Vickers[4],
Huiying Yan[3], Chan Zhang[3]
AT&T Labs Research, Florham Park, NJ, USA[1], AT&T, San Francisco, CA, USA[2]
Survey Research Center, University of Michigan, Ann Arbor, USA[3]
The New School, New York, NY, USA[4]
johnston@research.att.com, ehlen@research.att.com,
fconrad@umich.edu, schober@newschool.edu,
antoun@umich.edu, stefaniefail@gmail.com, ahupp@umich.edu,
lucasvickers@gmail.com, yanhuier@umich.edu, chanzh@umich.edu

Abstract

We explore the plausibility of using automated spoken dialog systems (SDS) for administering survey interviews. Because the goals of a survey dialog system differ from more traditional information-seeking and transactional applications, different measures of task accuracy and success may be warranted. We report a large-scale experimental evaluation of an SDS that administered survey interviews with questions drawn from government and social scientific surveys. We compare two dialog confirmation strategies: (1) a traditional strategy of explicit confirmation on low-confidence recognition; and (2) no confirmation. With explicit confirmation, the small percentage of residual errors had little to no impact on survey data measurement. Even without confirmation, while there are significantly more errors, impact on the substantive conclusions of the survey is still very limited.

1 Introduction

Survey interviews play a critical role in the operation of government and commerce. Large-scale social scientific surveys provide key indicators of the success or failure of economic and social policies, driving critical policy and funding decisions. Market research surveys are key in evaluating products and services for business.

Survey interviews are typically conducted either via telephone or face-to-face by skilled human interviewers. But ongoing changes in communication technology threaten the viability of these methods. As people migrate from landline telephony to mobile-only (Ehlen and Ehlen 2007) and Voice-over-IP (Fuchs 2008) as primary modes of communication, they undermine the effectiveness of traditional survey sampling techniques that rely on random selection of numbers within a dial code. Telephone respondents were once reachable at a fixed geographic location in a largely predictable conversational environment. Now they are increasingly mobile, and more apt to prefer asynchronous communication. Thus it is imperative to understand how these changing behaviors affect survey results.

The work described here is part of a larger research project (see Schober et al. 2012; Conrad et al. 2013) that investigates the viability of four different modes for administering a survey interview over a smartphone: automated voice, human voice, automated SMS text, and human SMS text. Here we focus specifically on the automated voice mode and explore the use of a spoken dialog system for survey administration.

Spoken dialog systems are widely used in telephony applications such as customer service, information access, and transaction fulfillment. They are also now common in virtual assistant applications for smartphones and mobile devices. But survey designers seeking automation have mostly eschewed spoken dialog in favor of textual web surveys or touchtone DTMF response systems. A preliminary comparison of spoken dialog and touchtone survey systems is available in Bloom (2008), and Stent et al. (2007) offer an evaluation of a spoken dialog system for academic course ratings. The work presented here describes the first large-scale investigation into spoken dialog technology as a viable means of administering the kinds of surveys that produce official statistics and social scientific data.

Survey interview designers should be interested in using spoken dialog systems for several reasons. The most obvious reason is to curtail the error and bias that human interviewers are known to introduce to survey results data. Decades of research and investment led to "standardized interviewing techniques" to reduce this error (Fowler and Mangione 1990), and limit a survey

Proceedings of the SIGDIAL 2013 Conference, pages 329–333,
Metz, France, 22-24 August 2013.

interviewer's ability to offer help or clarification in ways that might affect results. Automated dialog systems can be thought of as the ultimate in standardization, as they can be designed to provide exactly the same interaction possibilities to all respondents. In effect, everyone can be interviewed by the same "interviewer." Or, if survey designers want to allow clarification in an interview, an automated spoken dialog system can ensure that the same possibilities are available to all respondents (Schober and Conrad 1997).

Unlike systems that use human interviewers, there is marginal additional cost per interview after the initial investment of building a system. This offers significant potential for cost savings in large cross-sectional samples or repeated panel surveys, such as the U.S. Current Population Survey or the American Community Survey. Repeated data collection allows refinement and retraining of speech models to improve performance. Spoken dialog system surveys can be administered on demand at any time of day, allowing a better fit with respondents' circumstances and schedules. Compared to asynchronous text-based interviews like web or paper-and-pencil surveys, spoken dialog systems can capture richer verbal paradata (Couper 2009) or process data like pauses, disfluencies and prosody (Ehlen et al. 2007). Finally, survey tasks fit nicely within the limitations of current recognition and dialog technology, since they tend to have a purposefully structured and controlled interaction flow and generally require only a limited number of responses to each question.

While spoken dialog systems have the potential to remove data error that is introduced by variation in human interviewer behaviors, they also introduce risks to survey data quality due to speech recognition and understanding error. Numerous strategies for mitigating error have been explored in research on dialog systems (Bohus and Rudnicky 2005, Litman et al. 2006). One approach is to use either an explicit or implicit confirmation of the user's input. Following previous research showing that explicit confirmation is less confusing for users (Shin et al. 2002), we adopt an explicit confirmation strategy, which is also more in keeping with standardized interview techniques.

The effects of speech recognition and understanding errors may be different in a survey dialog system than in most current spoken dialog applications. One consideration is speaker initiative, and the stake of the user in the interaction. In systems for customer service, information access, or transactions, the user generally initiates contact with the system and seeks to accomplish a task where the system's recognition accuracy will affect success of the user's own goal. But in a survey dialog, the system initiates contact, and most respondents do not have a stake in whether the designers of the survey system succeed at collecting high quality data from them.

This is a key point where a survey interviewing system might differ from traditional SDS: From the survey researchers' perspective, the critical question is not whether individual users achieve some goal, but rather the extent to which individual errors in system recognition and understanding affect the distribution of responses across the population sample, affecting the quality of the estimates produced. If recognition errors do not affect the substantive conclusions based on the survey data, then survey researchers should be able to tolerate the imprecision of recognition error. This situation makes survey system evaluation rather different from how one would expect to evaluate the task success of a traditional SDS, like a customer service system.

In Section 2, we characterize the content of the survey items, describe the dialog strategy, and provide examples of interaction. Section 3 describes the technical architecture of the survey dialog system. We provide experimental evaluation in Section 4, and conclusions in Section 5.

2 Survey interview dialogs

After an initial question assessing whether the respondent is in an environment where it is safe for them to talk, our system administers a series of 32 questions drawn from major U.S. social surveys, including the Behavioral Risk Factor Surveillance System (BRFSS), National Survey of Drug Use and Health (NSDUH), General Social Survey (GSS), and the Pew Internet and American Life Project. The sample transcribed dialogs in Appendix 1 illustrate various features of interaction with the system. Question types include Yes/No, categorical (where users pick from a specified set of response options), and numerical questions. Some categorical items are grouped into battery questions with the same response options for all the items.

The system supports explicit requests to repeat the question or ask for help, and mimics a "standardized interviewing" style of interaction that trained interviewers would use to repeat or clarify a question when the answer is rejected or requires confirmation. Thresholds set on acoustic and language confidence scores are used to de-

cide whether to reject, explicitly confirm, or accept a response. The final question in the dialog in Appendix A ("Thinking about …") illustrates the importance of confirmation in ensuring the correct survey response is recorded. In this case, the system misrecognized "None" for "Nine," but this was caught by the explicit confirmation prompt. Two terms are introduced in the final example that we will return to in the evaluation. *First hypothesis* indicates the speech recognition and semantic result produced by the system the first time the question is asked. *Last hypothesis* indicates the speech recognition and semantic result that the system produced the last time the question was asked within the segment.

3 System Architecture

The survey dialog system is directly integrated with a custom-built survey data case collection management system (PAMSS). When a survey case is administered, the case management system makes an HTTP request to a voice gateway, which initiates a call to the respondent. When the respondent answers, it bridges the call to a spoken dialog system running within the AT&T Watson^{SM} speech platform. The system uses prerecorded prompts for survey questions and reprompts. Confirmations for numeric responses combine prompts with TTS output.

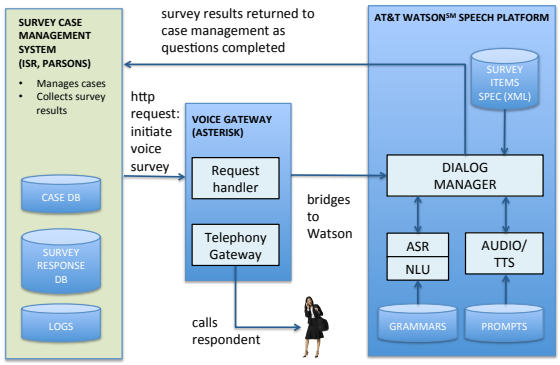

Figure 1: Survey Dialog System Architecture

Users' spoken inputs are recognized using state-specific grammars for each question. Data were not initially available for training statistical models, so SRGS (Hunt and McGlashan 2004) grammars were built for each answer. These were tuned in an initial pilot phase. The grammars included standard responses for the question, along with common paraphrases and framing words from the question. In the Watson platform, a dialog manager (built in Python) is integrated with ASR and TTS engines. Questions to be administered are represented in a declarative format in a survey item specification along with references to the appropriate prompts and grammars. The dialog manager interprets this specification to administer the survey and control the interaction flow. As the user responds to questions, the answers are posted back to the survey case management system.

4 Experimental Evaluation

We evaluated the survey dialog system as part of the first phase of a larger experiment comparing different survey interaction modes (Schober et al. 2012). In this phase, 642 subjects were recruited from Craigslist, Facebook, Google Ads, and Amazon Mechanical Turk. A web-based screener application verified respondents to be over 21 and collected their zip code. Of these, 158 respondents were randomly assigned to the automated voice condition. A $20 iTunes gift card was given as an incentive after completion of a post-interview web questionnaire. This included multiple-choice questions examining user satisfaction with their experience. In total there were 8,228 spoken inputs over the 158 respondent dialogs. These responses were transcribed, coded, and annotated for semantic content.

The questions we sought to answer were: What is the performance of a spoken dialog system on a typical survey task? What impact does speech recognition and concept error have on overall survey estimates? Does an automated survey system benefit from implementing a traditional confirmation strategy, where responses with low confidence scores are verified with confirmation dialog? We also examine the impact of dialog length and confirmation prompts on a qualitative measure of user satisfaction.

4.1 ASR and concept accuracy

We evaluated overall word, sentence, and concept accuracy for all 8,228 spoken utterances to the system, shown in the first row of Table 1.

Accuracy:	Word	Sentence	Concept
All	80%	78.2%	90.3%
First	81.2%	78%	88.9%
Last	88.5%	85.4%	**95.6%**

Table 1: System Performance

An input is "concept accurate" if the semantic value assigned by the system exactly matches that assigned by the annotator. *First* shows the performance on the first response made by a user to each question before any confirmation dialog. *Last* shows performance on the last time each question was asked. Concept accuracy on *last* responses is 95.6%, showing that the confirma-

tion strategy resulted in a 60% relative reduction in error compared to the first response.

4.2 Impact of Errors on Survey Estimates

Recognition error is undoubtedly a key factor in overall user experience. But unlike dialog systems for information access, search, and transactions, the most important factor in a survey dialog system is the impact of errors on the quality of the estimates derived from the survey. To examine the impact of the residual 4.4% concept error on overall survey error, we compared answer distributions derived from the system hypothesis for the last response versus the annotation of the last response using paired t-tests.

For the 18 categorical questions, we conducted t-tests comparing the counts for each response option of each question. For all 18 questions (a total of 77 response categories) none of the differences were statistically significant ($p<0.05$). For the 14 numerical questions, for only one ("Number of times shopping in a grocery store in the last month") did the interpretations differ significantly (Annotated: 7.8 times, Hypothesis: 7.6 times, $p=0.04$).[1] This is strong evidence that speech recognition errors in this system did not have a major effect on survey estimates.

How much survey error would have occurred without the dialog strategy? To test this, we compared the annotated last response to the system hypothesis for the first response, simulating an interaction without confirmation dialog, and thus lower recognition accuracy—see Table 1 (This is not a perfect simulation, as we have no independent evidence on whether the first or final response is true). There would indeed have been more survey error without dialog, although the overall level was still surprisingly low. For the 18 categorical questions, 14 of the 77 response categories show significant differences ($p<0.05$). For the 14 numerical questions, two showed significant differences.

4.3 User Satisfaction

One of the post-interview questionnaire items provided a qualitative measure of user satisfaction: "*Overall, how satisfied were you with the interview?*" The results were: *Very satisfied* (47.3%), *Somewhat satisfied* (41.8%), *Somewhat dissatisfied* (7.1%), and *Very dissatisfied* (0.6%). We examined the impact of various dialog features that seemed on intuitive grounds plausibly connected with satisfaction: average number of turns per question, average number of clarification prompts per session, and average number of no input response prompts. We conducted a series of logistic regressions with one variable controlled at a time to see the extent to which each of these features affected satisfaction. A Chi-squared test was used to measure significance. All three features were significant predictors when comparing *Somewhat/Very Dissatisfied* to *Very/Somewhat satisfied* (Table 2).

Feature	Odds ratio	SE	p
# turns per Q	10.411	0.787	0.003
# clarifications	1.043	0.033	0.024
# no input	2.001	0.176	<0.001

Table 2: User satisfaction regression

5 Conclusion

Our results demonstrate the viability of conducting survey interviews of the sort from which important national statistics are derived with spoken dialog systems. In our system, the speech recognition errors (with an overall concept recognition rate of 95.6%) did not substantially affect the error of the survey estimates; for only one of 32 questions was there a significant difference in the survey estimate determined by the automated spoken dialog system compared to the annotated result. Of course, we don't know whether these results generalize to dialog systems with other features, different questions, or different respondents; much remains to be learned.

Nonetheless, our results provide some guidance for improving respondent satisfaction and minimizing survey error in future development of survey dialog systems. For example, for numerical questions, which generally involve larger numbers of response options, recognition errors may be reduced by adopting the strategy of asking the respondent to select among categories representing ranges (e.g. "none", "1 to 5 times", "6 to 10 times"). Recognition performance could be improved by tuning confirmation strategies, e.g. applying a tighter confidence threshold for numerical vs. categorical questions. In a broad scale application of a repeated spoken dialog survey, greater amounts of data could be available for training statistical models for the responses, for improved recognition accuracy and further reduced concept error. Finally, it is worth exploring the trade-offs for survey error and respondent satisfaction between adding potentially frustrating confirmation dialog and accepting lower-confidence recognition for subsequent human annotation and processing.

[1] If we treat the two interpretations as independent samples, the response distributions did not differ significantly at all.

Acknowledgments: NSF #SES-1025645 and SES-1026225 to Conrad and Schober.

References

Jonathan Bloom. 2008. The Speech IVR as a Survey Interviewing Methodology. In Conrad and Schober (eds.), Envisioning the Survey Interview of the Future. Wiley, New York.

Dan Bohus and Alex Rudnicky. 2005. Sorry, I didn't Catch That: An Investigation of Non-Understanding Errors and Recovery Strategies. Proceedings of SIGdial-2005, Lisbon, Portugal.

Frederick G. Conrad, Michael F. Schober, Chan Zhang, Huiying Yan, Lucas Vickers, Michael Johnston, Andrew L. Hupp, Lloyd Hemingway, Stefanie Fail, Patrick Ehlen, and Chris Antoun. 2013. Mode Choice on an iPhone Increases Survey Data Quality. 68th Annual Conference of the American Association for Public Opinion Research (AAPOR), Boston, MA.

Mick P. Couper, 2009. The Role of Paradata in Measuring and Reducing Measurement Error in Surveys. NCRM Network for Methodological Innovation 2009: The Use of Paradata in UK Social Surveys.

John Ehlen and Patrick Ehlen. 2007. Cellular-Only Substitution in the United States as Lifestyle Adoption. Public Opinion Quarterly: Special Issue Vol 71 (5), pp. 717-733.

Patrick Ehlen, Michael Schober, and Frederick G. Conrad. 2007. Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces. Discourse Processes 44:3, pp. 245–265.

Floyd J. Fowler and Thomas W. Mangione 1990. Standardized Survey Interviewing; Minimizing Interviewer Related Error. Sage Publications, CA.

Marek Fuchs, 2008. Mobile Web Surveys: A Preliminary Discussion of Methodological Implications. In Conrad and Schober (eds.), Envisioning the Survey Interview of the Future. Wiley, New York.

Andrew Hunt and Scott McGlashan. 2004. Speech Recognition Grammar Specification Version 1.0. W3C Recommendation 16 March 2004. http://www.w3.org/TR/speech-grammar/.

Diane Litman, Julia Hirschberg, and M. G. J. Swerts. 2006. Characterizing and Predicting Corrections in Spoken Dialogue Systems. Computational Linguistics 32:3, pp. 417-438.

Michael F. Schober and Frederick G. Conrad. 1997. Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly, 61, pp. 576-602.

Michael F. Schober, Frederick G. Conrad, Chris Antoun, Carroll, Patrick Ehlen, Stefanie Fail, Andrew L. Hupp, Michael Johnston, Courtney Kellner, Kelly Nichols, Leif Percifield, Lucas Vickers, Huiying Yan, and Chan Zhang. 2012. Disclosure and Quality of Answers in Text and Voice Interviews on iPhones. 67th Annual Conference of the American Association for Public Opinion Research (AAPOR) Orlando, FL.

Jongho Shin, Shrikanth Narayanan, Laurie Gerber, Abe Kazemzadeh, and Dani Byrd. 2002. Analysis of User Behavior under Error Conditions in Spoken Dialogs. Proceedings of ICSLP-2002.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: The Rate-a-Course system. Proceedings of the IEEE/ACL 2006 Workshop on Spoken Language Technology. pp. 210-213.

Appendix – Sample Survey Dialogs

I: Have you smoked at least 100 cigarettes in your entire life? Yes or no.
R: No
I: Thanks

I: Now I'd like to ask you some questions specific to your diet. For the next group of questions, I'll ask you how much you favor or oppose engaging in certain behaviors. After I say the behavior, please say one of the following: strongly favor, somewhat favor, neither favor nor oppose, somewhat oppose or strongly oppose. The first behavior...avoiding fast food.
R: Strongly oppose
I: You said strongly oppose, right?
R: Yes
I: Got it
I: The next behavior: Maintaining a healthy diet
R: Repeat
I: The next behavior: Maintaining a healthy diet
R: Strongly agree
I: Sorry, I didn't hear you. When thinking about 'maintaining a healthy diet', please say one of the following: strongly favor, somewhat favor, neither favor nor oppose, somewhat oppose or strongly oppose.
R: Somewhat favor
I: You said somewhat favor, right?
R: Yes
I: Thanks
…
I: Thinking about the time since your eighteenth birthday (including the recent past that you've already told us about), how many male partners have you had sex with?
R: None (*First hypothesis*: "Nine")
I: I think you said '9'. Is that right? Yes or No.
R: No
I: Thinking about the time since your eighteenth birthday (including the recent past that you've already told us about), how many male partners have you had sex with?
R: Zero (*Last hypothesis*: "Zero")
I: Thanks