SPOKEN LANGUAGE RECOGNITION OF THE YEAR 2000 CENSUS QUESTIONNAIRE FEASIBILITY TEST

Martin V. Appel, U.S. Bureau of the Census, Ronald Cole, Center for Spoken Language Understanding
U.S. Bureau of the Census, Washington D.C. 20233-0400

1.0 BACKGROUND

The U.S. Bureau of the Census (CB) is a large, general-purpose, statistical agency. It conducts the Census of Population and Housing (Decennial Census) in years ending in 0, the Economic and Agricultural Censuses in years ending in 2 and 7, and hundreds of establishment and household surveys on a biannual, annual, monthly, or weekly schedule. The Decennial Census is by far is its largest and most complex data collection activity. Until quite recently, almost all the Bureau's data collection activities utilized paper questionnaires. This is changing.

2.0 TECHNOLOGY DEFINITION

For Census Bureau purposes, the Voice Recognition Entry (VRE) is defined as an automated data capture technology which allows a respondent, speaking over a telephone, to reply to computer generated prompts. The VRE system functions as an interviewer. At a minimum the system must, answer a call, prompt the respondent, recognize the respondents vocal replies and store the collected information. Recognition of the vocabulary that controls navigation through the VREQ must be extremely accurate. Other data items if not properly recognized can be recorded for future data keying. Respondents must have the ability to "bail-out" to a human operator if they desire.

3.0 MOTIVATION

For the Year 2000 Census, the Census Bureau is exploring alternative data collection technologies to help reduce the cost of taking the census and to provide response options for people who otherwise may not be counted, especially those populations which were differentially undercounted in the 1990 Census, and to offer alternative modes of responses in the hopes of decreasing the population of those who choose not to return their questionnaires.[1]

While it is very likely that the primary data collection vehicle will still be the paper questionnaire, during certain periods of the data collection process other methodologies may be advantageous. Potentially the VREQ offers a number of advantages relative to a written questionnaire: 1) it is not necessary that a respondent be able to read or write, 2) responses can be interpreted by the system and repaired by the respondent, 3) Upon completion of the VREQ the data is available for processing. Example: when a respondent calls us to request a questionnaire or we call the respondent to request data, an VREQ would be superior to waiting for the mailed questionnaire or having the telephone interviewer key the data.

The Bureau has a desire to maximize the participation rates. Studies have shown that non-respondents many times have different characteristics then the population which returned the questionnaire. To counter this reluctance or reduced ability, the Bureau is searching for ways to make the completion of questionnaires as "friendly" as possible. To this end, we are assessing a number of new technologies. The use of the telephone by the respondent for direct data entry is one of the new methodologies under study.

By its very nature, the Bureau's traditional short form census questionnaire favors the introduction of telephone technology. It contains only a few questions which, if we can structure the dialogue properly, can be answered with a constrained vocabulary while still maintaining a natural interaction with the respondent, while eliciting responses equivalent to those that would have been entered on a paper questionnaire.

4.0 EVALUATION PROCESS

The evaluation of the VREQ collection methodology is following the Census Bureau recommended three step process; 1) initial technology assessment, 2) small scale feasibility testing, and 3) large scale operational testing.

The first step, an **initial technical assessment** (ITA), summarizes what is currently known about the a candidate technology from publications, formal and informal organizational reports, material provided by vendors, and other easily accessible sources. The ITA answers a series of standardized questions covering such topics as: range of potential survey and census uses; stage of development; difficulty of application setup; costs of initial investment; user training required; user acceptance; and effects on survey costs, coverage, response rates, estimates, and timeliness.

Promising technologies proceed to **small-scale feasibility testing** to answer questions unresolved from the ITA or to evaluate their appropriateness for particular applications. Feasibility tests may include small field trials, studies arranged with university laboratories, and cooperative activities with other Federal agencies.

The third step in the evaluation process is a **large-scale operational test** of the technology in production use. An approved research plan is required. At a minimum, the test should measure the technology's impact on: survey costs; response rates; data quality; timeliness; and survey estimates. These objectives generally require an experimental design.

5.0 VRE ASSESSMENT

An ITA of VRE was performed last year. As a result of this assessment, the evaluation committee recommended, "that this technology be subjected to a feasibility study. Specifically, a short answer survey should be identified which requires both numeric and nonnumeric responses, and a research plan developed and implemented."

The information which follows describes briefly the research plan developed and the feasibility test that the Bureau is conducting.

6.0 RESEARCH PLAN

To explore the potential of VREQ, the Census Bureau, through the Office of Naval Research (ONR), has commissioned the Oregon Graduate Institute Center for Spoken Language Understanding and Carnegie Mellon University to build prototype VREQ systems that model a subset of the decennial census short form questionnaire. One system's recognizers will use neural-network technology and the other will use hidden-Markov modeling.

6.1 Objective

The specific objective of this project is to determine the feasibility of using a VREQ to facilitate data collection and capture for the Year 2000 Census. Feasibility will be determined by constructing prototype systems that successfully acquire the following information from callers in both English and Spanish:

1. Last name, first name, middle initial
2. Gender
3. Marital status
4. Date of Birth
5. Spanish/Hispanic origin
6. Race
7. Home Telephone Number with Area code

6.2 Prototype Construction Goals

1. To determine the most effective dialogue structures for the task and to refine methods of dialogue design for speech-based census collection;
2. To collect and transcribe a national sample of the speech data (English and Spanish) needed to design, train and evaluate the prototype understanding systems;
3. To develop a semantic and dialogue model that handles spontaneous speech (English and Spanish) and derives the intended response, using speech data acquired from above;
4. To develop two prototype spoken language understanding systems (each handling English and Spanish) for census questions defined above.

6.2 Prototype Evaluation Goals

1. To evaluate the prototype VREQ systems and compare performance of the two systems;
2. To identify the additional research needed to produce a viable VREQ for use in the Year 2000 Census.

7.0 PROTOTYPE DEVELOPMENT

7.1 Protocol Development

The goal of this phase of the research was to create a dialogue that feels natural and expressive for the caller while effectively limiting word choice. The concept is that the system's prompt creates a dialogue in which the natural act for the caller will be an utterance interpretable by the system and equivalent to her/his paper questionnaire response. This effort involved an iterative approach consisting of the following steps:

1. For each desired response, design a set of reasonable prompts;
2. Collect speech data from groups of callers using protocols containing the different prompts;
3. Analyze the responses; eliminate "bad" prompts; refine most promising prompts; and
4. Iterate stages 2 and 3 until satisfied.

The early rounds of protocol development involved the collection, transcription and analysis of responses from approximately 500 different telephone callers using six different protocols.

In the first round ("round-1") three alternative protocols were evaluated, ranging from open-ended ("What is your birth date?") to structured ("What day, month and year were you born?") to highly structured ("Please say the year in which you were born." "Please say the month in which you were born.") After collecting the round-1 data a qualitative analysis to determine the variability of data acquired and the degree of coverage associated with each protocol and question was performed. This analysis led directly to revisions in the phrasing and structuring of protocol questions for the second round.

The round-2 protocol was evaluated in terms of categories of callers' responses to individual questions using behavioral codes that included categories such as concise responses, usable (but not concise) responses, unresponsive answers, and no response. This analysis was used to choose among alternative prompts for developing round-3 protocols. For example, we found that one of three prompts that asked the respondent to tell us their sex was clearly superior. The three candidate prompts were:

"What is your sex?"

"What is your sex, female or male?"
"Are you female or male?"

About 90 calls were collected for each protocol. The percentages of concise responses to these prompts were approximately 95, 100, and 91 percent respectively. Consequently, it was concluded that the second prompt is the most effective for this question. Similar analyses were performed for the other questions.

This process of protocol development led to the selection of two "final" protocols used to collect nearly 4,000 calls. With one exception (the race question), the two protocols differed by small wording changes (E.g., "What is your sex, male or female?", verses "What is your sex, female or male?") Each protocol was recorded by a male and female speaker, and also produced using a speech synthesizer with a male and a female voice. These eight conditions (two protocols by four voices) were presented equally often to callers. One of the two protocols is given in Appendix A.

7.2 Speech Corpus Development

Task specific data is needed for the training and testing of recognizers and to refine the protocols. Using the round-3 protocols and untrained recognizers for navigation, we collected 3985 completed telephone conversations in English.

The calls were collected from 12 regions across the U.S. Calls from each region used a unique toll-free telephone number so we were able to categorize by originated region, the mix of accents and dialects in the corpus. Table 1 shows the distribution of calls by the regions. Respondents were recruited by the U.S. Bureau of the Census from adult families and friends of their field staff.

Approximately 100 different responses to each prompt were labeled and coded for each region; a total of about 1,200 different responses to each prompt. Table 2 shows the behavioral code that was assigned to each utterance indicating the nature of the response given to the prompt.

Table 1: English Round-3 Data Collection Report 11/22/93 through 02/20/94 (13 weeks)

Regions	Calls Rcd.	Compl. Proto.	Comp Eval
New York	273	162	161
Los Angeles	582	388	386
Dallas	494	275	271
Chicago	490	232	229
Boston	728	424	419
Charlotte	629	399	397
Atlanta	439	299	297
Philadelphia	697	346	343
Denver	907	375	367
Kansas City	703	485	484
Detroit	504	313	308
Seattle	345	280	279
Headquarters	109	7	6
Total	6900	3985	3947

Table 2: Behavioral Code Definitions

Code	Description
aa1	Adequate answer - the target word(s).
aa2	Target word in a common, or expected phrase, like "I'm white."
aa3	Target word in a non-expected environment or no appearance of the target word, but a statement which, using natural language could be interrupted as the target word.
qa	Qualified answer - speaker expressed doubt - "married I think."
ial	Speaker answers, but the answer is off the wall, doesn't even answer the question. Maybe they misunderstood the question.
ia2	Hang ups, speaker lurking on the line.
rc	Speaker requests specific clarification of the question.
in	Speaker interrupts the prompt, so the beginning of the answer is cut off.
dk	Speaker says "I don't know" or indicates that he doesn't know the answer to the question.
rf	Speaker refuses to answer the question "I refuse to answer that."
o	Other respondent behavior - this usually is given to calls in which the caller has done something in addition to just answering. Like he/she may speak to the person in the background, or he/she may cough, etc.

7.3 System: Hardware and Configuration

The Census VREQ system is distributed over several platforms. It uses a digital (T-1) telephone line, providing 24 channels shared amongst 15 toll-free 800 numbers. The T-1 line is connected to three LINKON voice boards in a PC-class computer running Solaris. When recognition or text-to-speech was required as part of the protocol, processing was sent, over a LAN, to a DEC Alpha computer. The telephone interface was shared with other OGI/CSLU applications. DNIS *(dialed number identification sequence)* was used to start the appropriate application.

The typical Census VREQ dialogue cycle was:

1. Play instructions followed by a question according to the current dialogue state;
2. Record the caller's response;
3. Perform utterance detection to remove background noise;
4. Invoke the recognizer, with a grammar and vocabulary specific to the question being answered;
5. Repeat the question if the confidence is low;
6. Branch to a new state of the dialogue depending on the recognized response and the confidence of the match.

Prerecorded system prompts are stored on the disk of the PC, but a final summary of the caller's responses is synthesized in real time and pipelined over the network.

8.0 RECOGNITION BREAKDOWN

The system attempts to resolve difficulties whenever possible. Repair strategies currently supported by the system include:

- Repeating the question if low confidence;
- Confirming the response if medium confidence;
- Taking the best guess and continuing with the next question if the system fails to recognize a response on a second attempt.

If difficulties persist, the system must fail gracefully. Currently, the system provides a summary of information recorded at the end of the dialogue. The user is asked to indicate which information, if any, is incorrect. Because all responses are recorded, a human may be able to resolve errors at a later time. In the final working system, we anticipate supporting more elaborate kinds of human intervention, the specific details of which are still to be resolved.

9.0 PERFORMANCE EVALUATION

At this time, results are only available for the English language neural network VREQ.

In this section we evaluate the protocol design based on an analysis of the data collected in round-3. We evaluate the performance of the recognizers based on the development test deck. We will also discuss the wider issues of measuring the effectiveness of spoken language systems in real world applications.

9.1 Evaluating the Protocols

Were the protocols effective in obtaining the desired information? This analysis focused on three major factors: completion, content, and conciseness of the responses.

Completion. Of those callers who responded to the first prompt, only 2.2% failed to complete the protocol.[2]

Content. How many responses provided the requested information. Responses were combined for the natural and synthetic male and female voices and for the two protocols whenever the prompts differed only slightly.[3] Table 3 shows the percentage of responses that contain the desired information for each prompt. It can be seen that the percentage of informative responses ranged from 99.1% to 99.9%.

Conciseness. A detailed analysis was made of the distribution of informative responses. It can be seen that about 97% of the responses contained the desired word, either by itself (aa1; "Male") or in a common phrase (aa2; "I'm male"). About three percent of the informative responses were coded as aa3; that is, the response did not contain the exact word or phrase, but did provide the desired information. Subsequent analysis of this category revealed that well over half of the aa3 responses were concise, and could be recognized without natural language processing. For example, instead of the target word "white" the caller may have said "caucasian."

9.2 Recognition Performance

Word Recognition. Our strategy has been to focus on the questions in turn, building a vocabulary-dependent recognizer for each. Since most of the informative responses contain one of the desired target words, we want to first achieve acceptable performance on them before tackling the remaining responses.

Furthermore, rather than working on each question until we have achieved the best possible performance, we stop when the performance is acceptable and move on to the next question. It is important to have a complete interactive system with reasonable recognition rates as soon as possible so that we can evaluate the dialogue component in a live system.

To date, we have developed task-dependent recognizers for most questions. They were trained on the hand-transcribed portion of the corpus, using automatically located phoneme boundaries. Table 4 shows the system performance for the transcribed portion of the development set. Only responses containing a target word are considered.

The prototype developers report that the data collected for this task is noisier than other corpora they have collected. It is also regionally very diverse. It is felt that all recognizers except the numbers (day and year) are performing reasonably well for this stage of development. Work on the number recognizer is ongoing as of this writing.[4]

Table 3: Percentage of Responses Which Are Informative

	Calls	%	%	%	%
Task	Eval.	aa1	aa2	aa3	non-inform.
Day	2475	85.4	11.0	2.9	0.7
Ever married	2421	97.4	0.8	1.6	0.3
	Calls	%	%	%	%
Task	Eval.	aa1	aa2	aa3	non-inform.
First name	1440	89.5	3.5	6.6	0.3
Gender	2364	97.0	1.4	1.1	0.4
Hispanic	2432	98.8	0.5	0.4	0.3
Last name	1248	91.6	1.7	6.1	0.3
Marital status	2052	89.1	1.9	8.1	0.9
Middle initial	1231	95.0	3.7	1.1	0.2
Month	2404	92.3	1.7	5.8	0.1
Race	1230	92.6	3.2	3.7	0.6
Spell first name	1307	91.7	4.9	2.9	0.5
Spell last name	1231	90.3	7.5	1.9	0.4
Year	2341	95.0	1.6	2.9	0.6

Table 4: System Performance on Development Calls Which Contain One of the Target Words

Task	Test Calls	Percent Correct
Month born	252	96.8%
Race (grp 1)	291	98.3%
Marital status	236	98.7%
Gender	242	100.0%
Yes/no	797	99.1%
Day born	240	85.0%
Year born	245	79.0%

10.0 CONCLUSIONS

These preliminary research results suggest that VREQ can be designed to produce concise and informative responses for a census task. Callers who completed the protocol produced the desired information about 99% of the time. The system's recognition rates shown in Table 3 are very encouraging.

There is much work yet to be done to produce a robust and graceful VREQ for this task. The prototype system described was developed as part as a feasibility study. The next step will be to develop a production prototype. Some of the issues that need to be address are:

- Improved detection of non-responsive utterances.
- The capability to hand the call off to an operator. The system should transfer the caller to a human operator when it judges that it cannot handle the call.
- Extending the questionnaire to include all members of the household.
- Including coverage questions, (usually requiring yes/no answers) such as whether other members of the household are temporarily residing elsewhere.
- Coping with breakdown and repair.
- Improved understanding of spelling and names. We plan to use probabilities of names given the other information we have collected (for example, Bill is a less likely first name for females).
- More robust recognition in the presence of noise, due to noise "subtraction" and modeling.

Based on the initial results, we expect that feasibility can be demonstrated for this task, and that spoken language systems could be developed for widespread use in the Year 2000 Census, possibly for several languages. The task requirements, consisting of a small set of responses to most questions, provide an ideal match to the capabilities of current technology. It is likely that continued improvements in the technology will produce systems that met the requirements of the task and are acceptable to users.

The next step will be to determine user acceptance. When the prototypes are completed in June '94, the Bureau's Center for Survey Methods Research will be evaluating respondent use of the VREQ in a laboratory setting. If no "gross" problems are encountered, it is expected that the VREQ will be introduced into the 1995 Census Test, a part of the research and planning for the 2000 Census.

APPENDIX A

English Round-3 Protocol

- Thank you for calling the OGI census project. We appreciate your help. The goal of this study is to determine the feasibility of using a computerized questionnaire for the Year 2000 Census. This research is sponsored by the United States Census Bureau. The answers you give to the following questions will be kept confidential. Afterwards we will ask you some questions to help us evaluate this questionnaire. It will take approximately four minutes to complete. Please wait for the tone before answering each question.
- Please say your first name.
- Please spell your first name.
- Please say your last name.
- Please spell your last name.
- Please say your middle initial. If you have no middle initial, say "none".
- What is you sex, female or male?
- We will now ask about your marital status. Have you ever been married? Please say yes or no.
- *(if yes, then)* Which one of the following options best describes your current marital status: now married, widowed, divorced, or separated?
- We will now ask about your date of birth. What month were you born?
- What day of the month?
- What year?
- We will now ask about your origin. Are you of Spanish or Hispanic origin? Please say yes or no.
- *(if yes then)* Are you of Mexican, Mexican-American or Chicano origin? Please say yes or no.
- *(if no then)* Are you of Puerto Rican origin?
- *(if no then)* Are you of Cuban origin?
- *(if no then)* Please say what other Spanish or Hispanic group is your origin.
- Please spell that.
- We will now ask about your race. Are you: White, Black or Negro, American Indian, Eskimo, Aleut, or other?
- *(if American Indian, then)* What is the name of your tribe?
- Please spell that.

- *(if other, then)* Okay. Are you: Chinese, Japanese, Asian Indian, Korean, Vietnamese, or other?
- *(if other, then)* Okay. Are you: Filipino, Hawaiian, Samoan, Guamanian, or other?
- *(if other, then)* Please say the name of your race.
- Please spell that.
- Is that the name of an Asian or Pacific Islander race?
- Do you have a telephone at home? Please say yes or no.
- *(if yes, then)* Please say your home telephone number, area code first.
- Finally, we'd like some additional information to help us with our study. What is your native language?
- In what city and state did you spend most of your childhood?
- Are you a Census Bureau employee?
- **This concludes the questionnaire portion. We will now ask you some questions to help us evaluate this questionnaire.**
- Would you be willing to provide census information using a questionnaire of this type over the telephone?
- In this questionnaire, we asked about your name, sex, marital status, date of birth, origin, race and telephone number. Please tell us about any questions you found unclear or poorly worded.
- What, if anything, did you like about this questionnaire?
- What, if anything, do you suggest we do to improve this questionnaire?
- We would like to hear any further comments you may have. You may begin speaking at the tone. When you're through, if you would like a gift certificate to either Baskin Robbins, TCBY Yogurt, B. Dalton Books, McDonald's, or Blockbuster Video, please say which one and leave your mailing address. Thank you for your help.

REFERENCES:

A list of meaningful references is too extensive to include. If you would like a copy, please contact: Martin V. Appel, U.S. Bureau of the Census, Room 3000 FOB 4, Washington D.C. 20233, (Ph) 301-763-2364.

[1] 1990- 66% response rate for the short form mailout form. Total returns from occupied households was 74.9%

[2] To eliminate crank calls, wrong numbers, etc., we calculated the number of callers who completed the protocol after responding to the first prompt.

[3] Analyses of the behavioral codes showed small differences among these conditions.

[4] The alphabet recognizer has not been evaluated on this task yet.