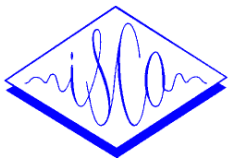

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A PROTOTYPE VOICE-RESPONSE QUESTIONNAIRE FOR THE U.S. CENSUS

Ronald Cole, David G. Novick, Mark Fanty,
Pieter Vermeulen, Stephen Sutton, Dan Burnett & Johan Schalkwyk

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
P.O. Box 91000, Portland, OR 97291-1000 USA

ABSTRACT

This paper describes a study conducted to determine the feasibility of using a spoken questionnaire to collect information for the Year 2000 Census in the USA. To refine the dialogue and to train recognizers, we collected complete protocols from over 4000 callers. For the responses labeled (about half), over 99 percent of the answers contain the desired information. The recognizers trained so far range in performance from 75 percent correct on year of birth to over 99 percent for marital status. We developed a prototype system that engages the callers in a dialogue to obtain the desired information, reviews the recognized information at the end of the call, and asks the caller to identify the response categories that are incorrect.

INTRODUCTION

We have conducted a study to determine the feasibility of using an automated spoken questionnaire to collect information for the Year 2000 Census in the United States of America. The goal of the study was to develop and evaluate a telephone questionnaire that automatically captures and recognizes the following information: (1) full name, (2) sex, (3) birth date, (4) marital status (now married, widowed, divorced, separated, never married—choose one), (5) Hispanic origin (yes or no); if Hispanic: Mexican, Mexican-American, Chicano, Puerto Rican, Cuban or other (specifiy), (6) race: White, Black or Negro, American Indian (specify tribe), Eskimo, Aleut, Chinese, Japanese, Filipino, Asian Indian, Hawaiian, Samoan, Korean, Guamanian, Vietnamese or other (specify).

After preliminary rounds of data collection to refine the selection and wording of the system prompts, a large, regionally diverse data collection effort resulted in approximately 4000 calls. This paper describes the effectiveness of the protocol in eliciting the desired information and it describes the spoken language system that resulted.

SYSTEM

Recognition

Signal Processing. The caller's response is transmitted over the digital phone line as a 8 kHz mu-law encoded digital signal. A seventh order Perceptual Linear Predictive (PLP) analysis [1] is performed every 6 msec using a 10 msec window.

Phonetic Classification. Each 6 msec frame of the signal is classified phonetically by a three layer neural network. To achieve maximum performance, a separate vocabulary-dependent network is trained for each response category, using a phoneme set particular to the expected pronunciations of words in that response category. This consists of the subset of standard phonemes which occur in the vocabulary, plus any additional context-dependent phonemes which were deemed necessary (e.g. [tw] for the [t] in "twenty" and "twelve"). The background noise and silence are modeled by a special phoneme [.pau].

For each frame of speech, the neural network is provided with 70 inputs, which consists of eight PLP coefficients and two voicing outputs from the frame to be classified and averaged over each of the following regions before and after the frame to be classified: 6 to 18 msec, 36 to 48 msec and 72 to 84 msec.

The two inputs that estimate voicing for each frame are provided by a separate three-layer neural network trained on voiced and voiceless speech frames from ten different languages. Although the voicing classifier is trained with the same PLP features described above, experiments have shown that including these features improves classification performance.

The outputs of the network fall in the range (0,1) because of the sigmoid transfer function, and, ideally, approximate the *a posteriori* probability of that phoneme given the input [2]. These values are divided by the prior probability of the phoneme in the training set [3].

Training the Classifiers. Training the neural network required phonetically segmented data. We used a semi-automatic procedure that involved hand transcription at the word level of about a quarter of the corpus and automatic generation of "forced" phonetic alignment of these transcriptions using a classifier trained on a different task. A new classifier was then trained on the automatically aligned census data and used to realign it. The process was repeated a couple of times until performance asymptoted.

An equal number of training samples (approximately 1000) was used for each phoneme class. As a consequence, rare phonemes were sampled more finely than common phonemes. Training examples for background noise and silence were chosen such that at least half occur close to phoneme boundaries. This balancing was needed to train for proper discrimination between the background class and unvoiced closures.

The neural network was trained using backpropagation and a combination of stochastic gradient descent (initial estimate of weights) and conjugate gradient optimization. Training was stopped once performance on cross-validation test data peaked. The number of hidden nodes was selected to maximize performance of the cross-validation test data and varies between 30 and 50 depending on the number of phoneme classes.

Word Representation and Viterbi Search. Every word in the vocabulary is represented as a probabilistic pronunciation graph, with the nodes representing phonemes. Each graph has a designated set of initial and final states. The sequence of possible words was represented as a probabilistic finite state network. Each node of this grammar has a list of allowed words. For each utterance, a Viterbi search finds the highest scoring sequence of words, obeying the constraints of the grammar and pronunciation graphs. The score for a word is the sum of the log probabilities from the neural network for the matching phoneme of each frame the word spans. In addition, phoneme segments which exceed expected duration limits are penalized on a frame-by-frame basis. As of this writing, all transitions in the pronunciation graphs for the regular vocabulary and grammar have "probabilities" of 1.0 and do not contribute to the score.

A word-dependent N-Best search is performed [4] so that the top two hypotheses can be retrieved.

Word Spotting. While the vast majority of the responses in the 4000-call corpus are succinct, there are enough responses with "extraneous" speech that it needed to be dealt with. In addition, there was a great deal of background noise in many of the calls.

However, because the majority of responses are succinct, our initial system has a simple word spotting approach in which all words and sounds not in the target set match a single garbage model. We use the approach described in [5], in which the output score for the garbage word is computed as the median value of the top N phoneme scores for each frame, where N varies with the task and is set empirically. The grammar for the responses (except numbers, which are more complex) is [garbage/silence] [target word] [garbage/silence].

Confidence. The recognizer always returns a target response. This means that if we are trying to recognize "male" or "female" and the caller just coughs or asks "Do I press a button or what?" then either "male" or "female" will be recognized. We assign a confidence score to these matches so they can be rejected. Having a continuous score, rather than a binary decision, gives the dialogue module more information so it can make an appropriate response.

An ideal confidence score will be low for all incorrect responses and high for all correct responses. So far, the best confidence score we have found is to take the difference of the average frame score for the top-scoring target word and the average frame score for the second highest scoring target word.

Dialogue. Our efforts have focused on designing a system for cooperative users. Even with cooperative users, unexpected dialogue situations may arise. Some responses will inevitably fall outside the preferred response set. The ability to cope with unexpected dialogue situations is essential to achieving good system robustness. The current system includes three aspects to dialogue repair:

1. Detecting breakdowns. The system is capable of identifying certain difficulties when they arise (e.g., low confidence for a recognized response).

2. Recovering from breakdowns. Repair strategies currently supported by the system include repeating the question if confidence is low; confirming the response if confidence is medium; and taking the best guess and continuing with the next question if the system fails to recognize a response on a second attempt.

3. Review, followed by confirmation or identification of errors. The dialogue concludes with a summary of the information recognized by the system. The caller is asked if the system's information is correct; if not, the caller is asked to identify the incorrect categories. A human operator may be able to resolve errors either during or following the call, as described in the next section.

Operator Monitoring

The system includes a graphical user interface that enables a human operator to monitor active calls or review any call at a later time. The caller's responses can be played at any time. Each recognition response is displayed as text, color coded to indicate the system's confidence. The operator can change the record to correct recognition errors.

At the end of each call, the system reviews each recognized response using pre-recorded speech; the caller's first and last name are spelled. After the review, the caller is asked if the information is correct. If not, the system signals the operator that the call has an error and asks the caller which of the questions have incorrect answers. The operator can then listen to this information, and correct errors in the appropriate fields of the record without reviewing the whole call.

PERFORMANCE EVALUATION

In this section we evaluate the effectiveness of the protocol, the performance of the recognizer, and discuss the wider issue of measuring the effectiveness of spoken language systems in real world applications.

Evaluating the Protocols

Our analyses of the protocols focused on three questions: (1) What percentage of callers completed the protocol? (2) What percentage of responses contained the desired information? (3) What percentage of responses could be recognized with a small vocabulary key-word spotter?

Completion. We eliminated spurious data such as wrong number and crank calls and then totaled the number of callers who completed the protocol. Only 2.2 percent failed to complete the protocol. (Note that the respondents were Census employees or friends and relatives of Census employees.)

Desired Information. Next we analyzed how many responses provided the requested information. The percentage of informative responses ranged from 99.1 to 99.9 percent (see Table 1).

Conciseness. A detailed analysis was made of the distribution of informative responses. It can be seen that about 97 percent of the responses contained the desired word, either by itself ("Male") or in a common phrase ("I'm male"). About 3 percent of the informative responses did not contain the exact word or phrase, but did provide the desired information. Subsequent analysis of this category revealed that, in fact, well over half of the 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".

Table 1: Percentage of responses which contain a target word/are informative.

Task	Contains Target Word	Informative
day	96.4	99.3
ever married	98.2	99.7
first name	93.0	99.7
gender	98.4	99.6
hispanic	99.3	99.7
last name	93.3	99.7
marital status	91.0	99.1
middle initial	98.7	99.8
month	94.0	99.9
race	95.8	99.4
spell first name	96.6	99.5
spell last name	97.8	99.6
year	96.6	99.4

Recognition Performance

Word Recognition. We have developed task-dependent recognizers for each question. They were trained on the hand-transcribed portion of the corpus, using automatically located phoneme boundaries. Table 2 shows the system performance for the test set. Only responses containing a target word were run. The data collected for this task is noisier than other corpora we have collected. It is also regionally very diverse.

Table 2: System performance on test calls which contain one of the target words.

Task	Test Calls	With Target	Percent Correct	After Rejection
month born	777	743	91.9	93.9
race (group 1)	404	385	96.9	99.4
marital status	635	618	99.5	98.8
gender	787	776	98.8	99.3
yes/no	1827	1813	98.7	99.5
day born	769	763	75.8	NA
year born	762	758	76.0	NA

Name Recognition. The OGI name retrieval algorithm is designed for spelling with pauses between the letters [6]. However, callers were not asked to pause during the collection of this corpus in order to create a data set on which to develop fluent letter recognition. Other than retraining the classifier, the system architecture has not yet changed. In particular, alternate segmentations are not considered. This has lowered the percentage correct significantly.

The name-retrieval algorithm was modified to use prior probabilities estimated from name counts in the Seattle white pages, augmented with another list of 50,000 last names. For last names, there were 236 responses in the test set. The remaining were not yet transcribed. Of these, 190 were in the list derived from

the Seattle phone book. The system correctly identified 135/190 for 71 percent.

For first names, there were 259 responses in the test set. Of these, 238 were in the list derived from the Seattle phone book. The system correctly identified 192/238 for 81 percent.

Rejection

Since the overwhelming majority of the callers gave cooperative responses, we first focused our efforts on measuring confidence of recognition of words within each task vocabulary (e.g., "February" being recognized as "January"). For this task, we have had some success by measuring the difference between the score of the top and second ranking alternatives. Table 2 shows the performance using a rejection threshold which eliminated the lowest-scoring 5 percent of the responses in the development set. (In the dialogue, rejection causes the question to be repeated once.) This rejection scheme did not work for numbers.

DISCUSSION

Our research suggests that a spoken language interface can be designed to produce concise and informative responses for a census task. Callers who completed the protocol produced the desired information about 99 percent of the time.

The amount of information captured is likely to improve dramatically when the protocol is incorporated into a spoken language system. The system will be able to detect uninformative responses and repeat the question. At the end of the protocol, the system reviews the information recognized, and give the caller yet another chance to respond appropriately. Unlike the simple data-collection protocols evaluated in this study, a spoken language system can create several opportunities to obtain the desired information.

For the low-perplexity recognition tasks (e.g., gender, marital status, those requiring yes/no recognition), recognition performance is acceptable when classifiers are trained on task-specific data. Most remaining errors can be identified automatically using measures of confidence applied to the recognition scores. For the more difficult recognition tasks, additional work is required to achieve acceptable levels of performance. We are currently investigating two promising approaches: using context dependent phonetic networks, following [7]; and performing word-specific reclassification following the initial recognition.

There is much work yet to be done to produce a robust and graceful spoken language system for this task. The prototype system described was developed as part of a feasibility study. The next step will be to develop and deploy a production prototype.

ACKNOWLEDGEMENTS

We thank the U.S. Bureau of the Census, U.S. Office of Naval Research, Apple Computer Inc., Digital Equipment Corporation, Lernout & Hauspie Speech-Products Inc., LINKON Corporation, and U S WEST for their research support.

We are indebted to the research work of Brian Hansen. Many thanks to Mike Noel and Terri Lander for their work on corpus development for this project. Also, thanks to the CSLU Corpus Development Staff for their labeling and transcribing efforts.

REFERENCES

J.-M. Boite, H. Bourlard, B. D'hoore, and M. Haesen (1993), "A new approach towards keyword spotting," Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin, Sep. 21-23, 1993, pp. 1273-1276.

M. Cohen, H. Franco, N. Morgan, D. Rumelhart and V. Abrash. "Context-dependent multiple distribution phonetic modeling with mlps," in J.D. Cowan S.J. Hanson and C.L. Giles, editors, Advances in Neural Information Processing Systems, volume 5, pages 649-657. Morgan Kaufmann, 1993.

H. Bourlard and C.J. Wellekens, (1989), "Links between Markov models and multilayer perceptrons," Advances in Neural Information Processing Systems 1, (Morgan Kaufmann, San Mateo, CA, 1989), pages 502-510.

M. Fanty, R. A. Cole and K. Roginski, (1992), " English alphabet recognition with telephone speech," Advances in Neural Information Processing Systems 4, Morgan Kaufmann, San Mateo, CA, 1992).

H. Hermansky, (1990), "Perceptual linear predictive PLP analysis for speech," Journal of the Acoustical Society of America, Vol. 87(4), pp. 1738-1752.

N. Morgan and H. Bourlard (1990), "Continuous speech recognition using multilayer perceptrons with hidden Markov models," Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing, pp. 413-416.

R. Schwartz and S. Austin, "A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses," Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing, (IEEE, 1991), pp. I1-I4.