USING A SPOKEN DIARY AND HEART RATE MONITOR IN MODELING HUMAN EXPOSURE TO AIRBORNE POLLUTANTS FOR EPA'S CONSOLIDATED HUMAN ACTIVITY DATABASE

Curry I. Guinn (Univ. of North Carolina Wilmington, USA)
Daniel J. Rayburn Reeves (Univ. of North Carolina Wilmington, USA)

ABSTRACT: In this paper, we describe technology used to collect fine-grained data of individual daily human activity including information related to time, exertion, specific activity, and location. This data uses the Consolidated Human Activity Database developed by the EPA's National Exposure Research Laboratory and is used to build models of human exposure to airborne pollutants. This paper will look specifically at subject compliance and quality in a small pilot study and the automatic classification of activity and location based on a spoken language diary collected from a digital recorder.

INTRODUCTION

The purpose of this study was to develop and test a method that would assist the Environmental Protection Agency, and the scientific community at large, in the generation of an activity/location/time/energy expenditure database of sufficient detail to accurately predict human exposures and dose. Collecting daily location/activity data along with measures of pulmonary dynamics to assess human exposure to airborne pollutants has proven to be expensive, to be burdensome to the participant, and to produce unreliable data. This study tested the feasibility of modifying existing technology to reduce respondent burden and cost and to improve the reliability of data collection.

One goal of this study was develop a protocol with highly detailed reliable information with low subject burden, two features which are generally inversely correlated. Our protocol tested the use of a digital voice recorder to collect activity/location diary data assuming it to be a less burdensome and a more reliable method than using paper and pencil diaries or hand-held computers. We expected the data to be more complete and reliable than retrospective reports (diaries filled out at the end of day) because the recorders are easy to use, the diary entries are made as the events occur, and we expected that participants would be more likely to complete the study because of the reduced burden. The data collection plan was also expected to show that the cost of the transcription of the diary can be reduced substantially by using speech and language processing to translate the digital diaries into the EPA's Comprehensive Human Activity Database (CHAD).

BACKGROUND AND PREVIOUS RESEARCH

Capturing Environmental Exposure Data. The definition of environmental exposure provided by Ott (Ott, 1982) and later adapted by others (NAS 1991;

USEPA, 1992) is "an event that occurs when there is direct contact at a boundary between a human and the environment with a contaminant of a specific concentration for an interval of time." This definition implies that four variables must be measured to accurately characterize exposure: location (L), time (T), activity (A), and concentration (C). The EPA developed a system to store and systematically analyze the available data, the Consolidated Human Activity Database (CHAD) (McCurdy et al., 2000).

Paper-based diaries, electronic diaries, voice-recorded diaries, and observational techniques have all been used to collect data about temporal activities, spatial locations, product use, and dietary consumption of research participants (Johnson et al., 2001). Diary methods relying on recall are not highly reliable and have a relatively high respondent burden, which negatively impacts participant compliance. Observational techniques are extremely costly and burdensome. Post-study processing of diary entries is labor intensive unless simplified reporting protocols are employed and automatic processing systems are developed.

In our study, participants are trained to record their current activity and location whenever the participant changes either of those conditions. They are trained to use simple English sentences such as "I am on the bus on my way to South Square Mall." The coder's task is to take spoken language utterances like "I am on the bus on my way to South Square Mall" and select the appropriate CHAD location code **(31140: Travel by bus)** and activity code **(18400: Travel for goods and services)**.

In order to reduce coder burden, one of our goals was to use the computer to assist in taking the participant statements and returning the appropriate CHAD codes. A common approach to text abstraction is to use statistical NLP techniques that analyze a training corpus to build probabilities that can be used to choose the most likely semantic categories for an utterance. In this study, a selection of the total corpus was set aside as training data. From this training set, human coders selected CHAD activity and location codes for each utterance. Using Bayesian statistics, unigram, bigram, and trigram probabilities were generated for each set of words for each activity and location code.

EXPERIMENTAL PLATFORM

Two physical devices were employed – one to capture the heart rate data; the other to capture the voice diary.

Heart Rate Monitoring. Each subject wore a custom-built heart rate monitor with three EKG electrodes. This device recorded heart rate at 2 minute intervals. Further, if the subject's heart rate changed by more than 15 beats per minute from the previous measurement, the device would beep – a signal that the subject should make a voice diary entry.

Voice Diaries. Subjects carried a holstered digital recorder (Sony ICD-MS1) to make their diary entries. Because of cumbersome headsets and unreliable wireless microphones, subjects were required to use the built-in microphone on the device.

During training, participants were instructed to make a voice diary entry whenever they changed location, changed activity, or if the heart rate monitor beeped.

EXPERIMENTAL RESULTS

Nine participants participated in the study for a duration of one week each in addition to the training period. A summary of some of the characteristics of these participants is given in Figure 1. Across the entire data collection period for the nine participants, the average daily experiment period was 8.56 hours.

Figure 1: Subject Characteristics

ID	Sex	Occupation	Age	Education
1	F	Manages Internet Company	52	Some College
2	F	Grocery Deli Worker	18	Some College
3	M	Construction Worker	35	High School
4	F	Database Coordinator	29	Graduate Degree
5	F	Coordinator for Non-profit	56	Some College
6	M	Unemployed	50	High School
7	M	Retired	76	High School
8	M	Disabled	62	High School
9	M	Environment Technician	56	Graduate Degree

Recordings per day. There were 1350 diary entries across the nine participants. The average number of recordings per participant was 29 per day. Given the average monitoring time of 8.56 hours, participants were making 3.39 recordings per hour. An interesting trend over the course of the field trial was a precipitous decline in the number of recordings as the week progressed. During the first three days of the trial, participants averaged 34.81 recordings a day. During the last 2 days of the trial, participants averaged 20.67 recordings a day. This suggests that participant fatigue of the recording process is significant. Figure 2 shows the diary reporting trend for participants over 7 days of the trial.

Prompting based on heart rate change. In an attempt to encourage participants to report their change in activities, the heart rate monitor would prompt the participant with a tone if the participant's heart rate changed by 15 beats per minute between two recording intervals. This tone was fairly unobtrusive. During training, participants were told to record their current location and activity whenever they heard the tone. Our trial study indicates that this tone encouraged diary entries with only limited success. By analyzing the heart rate data, we can determine when a tone was issued by the monitor. By cross-referencing those times with the times of diary entries, we can determine whether an entry was made around the time of the tone. For this analysis, we assumed the participant was following protocol if they made a diary entry at any point in a 3-minute interval surrounding the time of the tone. Figure 3 provides the average number of prompting tones issued per day and the percentage of times a subject made an entry corresponding to a tone.

Figure 2. Average Recordings Per Day Across Time

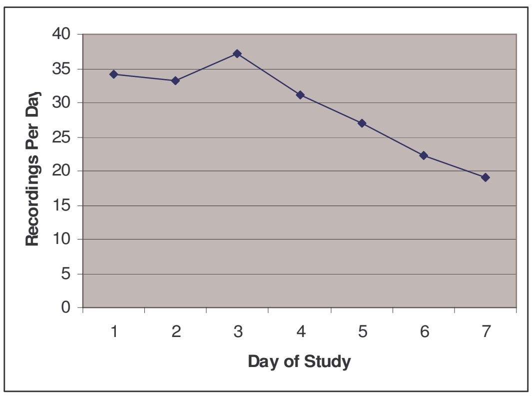


Figure 3: Heart Rate Change Indicator Tones and Subject Compliance

Subject	Number of Tones	% of Times Subject Made a Diary Entry
1	22.1	45%
2	41.8	29%
3	32.5	36%
4	33.0	55%
5	33.3	36%
6	15.6	40%
7	32.5	37%
8	26.0	22%
9	22.7	31%

Diary entry quality. Of the 1350 diary entries, 133 (9.85%) contained no encodable data. The subject's inadvertent start of recording or failure to stop a previous recording was the overwhelming cause of these false entries. One disadvantage of the voice diary is that participants are given no prompts (other than from the heart rate monitor) or guidelines interactively while making recordings. This problem was manifest in the data collected. In 90 entries (6.6%) the subject's activity could not be coded by a human listener. In 20 entries (1.5%) a human encoder could not determine the location. Examples of this sort of deficient diary entry might only include the location ("I'm in the kitchen") or only the activity ("I'm putting on my clothes"). It may be possible to infer the location or the activity in some cases, particularly if the previous diary entries were made in close time proximity.

NATURAL LANGUAGE PROCESSING METHODOLOGY

The natural language processing task is to take the words returned from the speech recognizer and determine the appropriate CHAD encoding for location and activity. As an example, "my bedroom making the bed" would be encoding as **30125 (Bedroom)** and **11200 (Indoor Chores)**. The strategy employed was to build statistical language models of the domain.

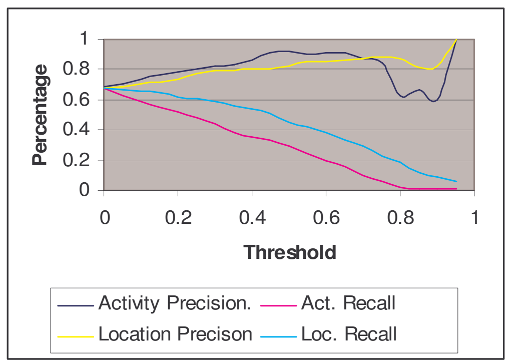
The valid diary entries were divided into a training set (925 entries) and a test set (279 diary entries). A human coder went through each set and determined the appropriate CHAD activity and location codes. Then uniqram, bigram, and trigram probabilities were determined for each entry in the training set. As each element of the test set is processed, these unigram, bigram and trigram scores are used to select maximally likely category (Guinn, 2006).

Statistical Processing Results. The training set was first processed using a human listener to perform the speech recognition to obtain raw transcripts of the voice diaries. The statistical techniques described above were applied to the training set. Then, using these statistics, the most likely activity and location was determined for the test set. We experimented with different thresholds by having the system not make a guess if the score was too low or if the relative scores of the top two choices were close. Figure 4 shows the precision and recall given a threshold of 0.3 while Figure 5 shows the relative tradeoff between precision and recall given different threshold values.

Figure 4: Statistical Processing Accuracy of Hand-Transcribed Data with Threshold of 0.3

Subject	Act. Prec.	Act. Recall	Loc. Prec.	Loc. Recall
1	93.3	66.6	76.5	61.9
2	54.5	27.9	62.5	46.5
3	55.6	35.7	62.5	47.6
4	82.5	51.5	74.1	60.6
5	87.5	66.0	78.0	73.6
6	87.9	39.9	67.3	48.6
7	68.1	64.0	55.9	80.6
8	60.4	40.4	56.0	68.4
9	89.5	72.3	93.0	85.1
Total	75.5%	51.6%	69.5%	63.7%

Figure 5 Threshold values affect the precision and recall: The higher the threshold, the greater the precision but the lower the recall



DISCUSSION AND FUTURE WORK

One drawback of the system described above is that data is captured passively by the system with no immediate feedback to the user. In practice, the only time the system communicated to the subject was when the system issued a beep if the subject's heart rate changed significantly. This beep was a signal to the subject to make a diary entry.

A better solution is to create a dialog between the subject and the system with more elaborate prompting and immediate error-correcting dialogs when the system is unable to process the user's diary entry. To create this system, we are currently building an input device with the PocketPC technology using a wireless mic/headset, built-in GPS, a pen-based diary, and peripherals for exertion and motion data gathering (Figure 5).

This system will also fuse data from multiple input sources and use discrepancies to trigger dialogs with the user. For instance, the system will have constant access to Global Positioning System data which includes not just coordinates but also speed. If the voice diary entry indicates that the person is at home but the GPS reports the position is away from home and traveling at 55 mph, then the system can prompt the user to re-enter the voice diary entry. In some situations, the system could prompt with very specific questions with expected Yes/No answers – a response which has a high likelihood of correct speech recognition.

ACKNOWLEDGMENTS

This work has been funded by the Environmental Protection Agency under contract 68-D-99-012 and as a subcontract to RTI International.

REFERENCES

Guinn, C., Crist, D, and Werth, H. A. 2006. "Comparison of Hand-Crafted Semantic Grammars Versus Statistical Natural Language Parsing in Domain-Specific Voice Transcription," *Proceedings of Computational Intelligence*, Ed. B. Kovalerchuk, San Francisco, CA, pp. 490-495.

Johnson, T., Long, T., and Ollison, W. 2001. *A survey of prospective technologies for collecting human activity data.* Contractor report for the American Chemistry Council.

McCurdy, T., Glen, G., Smith,L., and Lakkadi,Y. 2000. "The National Exposure Research Laboratory's Consolidated Human Activity Database". *J. Exp. Anal. Environ. Epidem*. 10:566-578.

National Academy of Sciences (NAS). 1991. *Human exposure assessment for airborne pollutants; advances and opportunities*. National Academy of Sciences, Washington, DC.

Ott, W.R. "Concepts of human exposure to air pollution." 1982. Environ. Intl. 7:179-196.

RTI, *Longitudinal activity data for selected susceptible population subgroups*. 2001. Final Report for EPA Contract No. 68-D-99-012 (Task Order 0012).

U.S. Environmental Protection Agency (USEPA). 1992. "Guidelines for exposure assessment". *Federal Register*, 57(104):22888-22938.