Submitted to IEEE Computer Magazine, Special Issue on Handheld Computing. September 2003.

Handhelds that Listen and Learn*

Nathan Eagle, Push Singh, Alex (Sandy) Pentland
{*nathan, push, sandy*}*@media.mit.edu*

It is no longer uncommon to see a businessman ranting to himself alone in a parking lot, or spy a woman driving to work in an empty car, engaged in a heated debate. Many of the same people who use their mobile phone's hands-free microphone also carry a 400 MHz PDA to store their "to-do" list. Could the handheld computer take advantage of the microphone's audio stream to infer useful information about the user's present situation? To find out, we have enabled over seventy linux-based, 802.11b-enabled handheld computers (the Sharp Zaurus) to record audio and other contextual information such as conversation participants and wireless access point IDs. With this hardware, we are exploring ways to infer aspects of a user's situation from spoken conversations using both context and commonsense knowledge.

Why is this problem hard? Even with the latest speech recognition engines trained to a user's voice, accuracy rates for spontaneous speech recognition fall below 35%. Conversation transcripts like the one shown below are difficult even for a human to comprehend.

> store going to stop and listen to type of its cellular and fries he backed a bill in the one everyone get a guess but that some of the past like a salad bar and some offense militias cambers the site fast food them and the styrofoam large chicken nuggets son is a pretty pleased even guess I as long as can't you don't have to wait too long its complicity sunrise against NAFTA pact if for lunch

To make sense of such transcriptions, we have developed a method for semantically filtering and regularizing the text using a commonsense knowledgebase. We use the Open Mind Common Sense (OMCSNet) semantic network, which contains over 250,000 pieces of commonsense knowledge such as '*when you go to a restaurant you look at a menu*' or '*cooking food requires an oven*'. While the words the speech recognition engine gets correct tend to be grouped around neighboring semantically-related nodes, errors in the transcriptions turn out to be distributed randomly over this network. The nodes surrounding the largest clusters of keywords are assumed to be potential aspects of the speakers' situation.

Additional context, such as information that the conversation participants are in a cafeteria, or more precisely, waiting in line, would help many people understand that the participants are deciding what to get for lunch. Thus robustness of the classifier can be further improved by biasing the prior probability of each node based on contextual information from the handheld computers, such as the user's location, conversation participants, time of day, or the people in local proximity.

The amount of help location information provides seems to vary significantly depending on the topic of conversation. For some people, conversations in restaurants regarding sports may occur just as much, if not more, than conversations about food. Once a user's recent history is incorporated, we expect the model's performance to increase considerably. Instead of using OMCSNet as a set of generic prior probabilities, we are beginning to implement online learning algorithms that incorporate subsequent observations into the classifier. This new model leverages the causal information within OMCSNet to yield a specialized model that better reflects an individual's behavior. Basically, the longer the handheld listens, the more it will learn.

Situational Inference. Two participants were standing in line, talking about what to order in the food court cafeteria. The situation classification with only the noisy transcript is shown in Table 1. Table 2 incorporates additional contextual information: the fact that the audio was streamed to the food court access point.

Table 1

Confidence	Classification with no context
5	Eat in restaurant
5	eat in fast food restaurant
5	buy hamburger
5	talk with someone far away
5	buy beer
4	go to hairdresser
4	wait in line

Table 2

Confidence	Classification with location context
27	eat in fast food restaurant
21	eat in restaurant
18	wait on table
16	you would go to restaurant because you
16	wait table
16	go to restaurant
15	know how much you owe restaurant

* This work was partially supported by the NSF Center for Bits and Atoms (NSF CCR-0122419).