Presented at the ACM Conference on Computer Supported Cooperative Work (CSCW). New Orleans, Louisiana. November 2002.

Collaboration from Conversation

Nathan Eagle
MIT Media Laboratory
20 Ames St.
Cambridge, MA 02139 USA
+1 617 253 0370
nathan@media.mit.edu

Alex (Sandy) Pentland
MIT Media Laboratory
20 Ames St.
Cambridge, MA 02139 USA
+1 617 253 0648
sandy@media.mit.edu

ABSTRACT
There are several different types of information inherent in conversations: speech features, participants, and content. By recording conversations in organizations and aggregating this information, we claim that high-potential collaborations and expertise can be identified. Preliminary results for learning context from computer transcribed conversations show encouraging results. Current research involves collecting a larger conversation dataset to be used for social network analysis, knowledge management and as training data for a simulation to model organizational disruptions.

Keywords
conversation, collaboration, speech recognition, privacy, knowledge management, social network analysis

INTRODUCTION
The majority of working professionals continually carry a microphone and speaker in the form of a cellular phone. Many are also carrying PDAs with computational horsepower similar to those found in desktop computers only a few years ago. This emerging foundation of mobile communications and processing power within the workplace will enable an exciting suite of business applications.

The importance of voice communications within the workplace is widely acknowledged. Critical pieces of information rarely disseminate though alternate mediums. The value of face-to-face interactions is exemplified by the money and time spent on business travel and conferences. This paper postulates that if an organization were able to collect all the relevant conversations of its employees, it would have an extraordinary resource for collaboration, team formation, knowledge management, and social network analysis. We will first discuss the information that can be obtained from many streams of audio conversations, describe on how this information can be applied to a variety of disciplines, and then discuss the privacy implications of such a dataset. The second half of the paper will discuss the current state of the research to date and the potential directions for it to evolve.

BACKGROUND
There are three types of information inherent within streams of audio recorded from individuals of a common social network: conversation features, participants, and context. Speech features can be extracted from conversations such as speaker energy (volume), transitions and duration. By analyzing the mutual information between audio streams the participants of a conversation can be identified and the internal social networks can be documented. And lastly, assuming a voice model for each individual exists, conversation content can be inferred from distinctive keywords using commercial speech recognition engines.

Conversation Features
In [2], Sumit Basu analyzed features such as speech energy, duration, and speaker transitions in conversations. It was shown that these transitions could be learned and predicted given enough training data. Extrapolating from these results, one can assume that these conversation features are also be indicative of the type of relationship between the two people involved. Conversations that are dominated by a single person allowing no time for interjections are quite different from those conversations that have lower speech energy and regular speaker transitions. Conversation analysis allows group dynamics to be quantified and provides individual behavior profiles that can assist in assembling a project team.

Conversation Participants
Social network analysis typically involves self-report surveys given at random to individuals in an organization. In these surveys participants are asked to simply list with whom they spoke during that particular day. Social network analysis has traditionally consisted of employing a variety of visualization tools on self-report survey data regarding these internal communication habits of an organization [1]. However, the inherent sparsity and uncertainty of the data have been limiting factors in the research. Relying on this subjective information creates a large amount of uncertainty in the already limited data set. To get around the problem, Tanzeem Choudhury's current research attempts to sense social proximity using IR beacons worn over the shoulder [3]. In this way the social networks are detected and can be visually represented automatically. The location of these interactions is also documented with ceiling mounted IR

beacons. She augments the system with accelerometer and audio data.

IR is not the only way to sense social interactions. By finding the mutual information between streams of audio conversations, one can determine who are participating in a conversation and who are in proximity of the conversation but not participating [1]. In this way an organization's verbal communication network structure can be mapped with only audio. Other work has shown that an analysis of ambient noise and light is also a method to derive location [3]. It is the premise of this paper that simply looking at the audio streams from people will provide this type of contextual and social information. When aggregated with data from email conversations and web postings, we posit that the resultant information can be used to identify high potential collaborators.

Conversation Content

A trained speech recognition engine typically has accuracy rates subject to wide variation. Despite this noisy output, typically keywords can be identified and the topic of a conversation can then be inferred. The experimental section of this paper empirically demonstrates that the vocabulary transcribed by the speech recognition engine is indicative of context and location. Vocabulary also lends insight into the nature of the conversation: a question, response, and the two people involved can all be documented. Such information is critical to distinguish the relationship types, expertise areas, and conversation relevancy to enable appropriate collaborations.

CONVERSATION-ENABLED COLLABORATION TOOLS

As headsets and lapel microphones continue to proliferate within the work environment, harnessing these sensors for alternate tasks will feasible. Leveraging the recent advances in speech recognition, keywords within conversations can be archived and individual profiles can be automatically generated. By querying profiles, a manager can form a team that has synergistic skills and social behavior. Clusters of people working on similar projects within a large organization can be identified to instigate collaboration and avoid redundant work. Experts can be identified similarly from conversation features, keywords and participants. In a real-time system, ad hoc information passing can be enabled so that relevant people who could be interested in an ongoing public conversation can be patched directly into the conversation. Many other applications across a range of fields will emerge from such data mining. This section will only discuss two: an expert/collaborator finder and ad hoc conversation patching. Initial results towards these applications will be discussed in the following section.

Expert and Collaborator Locator

The emerging area of knowledge management attempts to capture and visualize the collective knowledge of an organization from individuals' posted profiles and web documents [7]. Although existing corporate information repositories can be easily analyzed using standard data mining operations, the output reflects a severely limited and static view of an organization's human and social capital. Augmenting knowledge management and traditional social network analysis with information gathered by unobtrusive wearable sensors has enormous potential benefit for organizational collaboration.

With the current rate of disruptive technology emerging, it is becoming virtually impossible to keep track of the collective knowledge in a large organization. To even begin to infer this knowledge, the information disseminated verbally must be harnessed. This system will enable managers to visualize who is working with who, infer the type of relationship, the related interests, and generate a database of the categories of knowledge within the organization.

A database of employee profiles that dynamically updates to represent changes in social network, email and verbal conversation behavior should spark many high-potential relationships and lend insight into in-house expertise. Querying this database for interests, skills, or simply recent vocabulary would be an efficient way to instigate collaboration. When combined with the social network information, searching for keywords should immediately show who are working on similar problems and whether they are indeed collaborating. When speech features are added to the aggregate data, the intonation of conversations and speaker transition data provides sufficient data to make plausible expertise classifications.

Ad Hoc Conversation Patching

There is no technical obstacle from prohibiting the system described above from operating in real-time. As conversations occur throughout an office space, they can be automatically streamed to a server that spots keywords using an individual's voice model (or speaker independent model if accuracy constraints decrease) that are relevant to his or her profile. These keywords of running conversations can be parsed and clustered into topic categories [6]. If a public brainstorming session occurs and a specific technology is repeatedly mentioned, another employee who is interested and experienced with it could be automatically patched into the conversation. Perhaps a more compelling example would be that of emergency response. Words like "Help!" could be automatically sent to a rescuer's headset along with location information.

PRIVACY CONCERNS

Continually recording, transcribing, and archiving all conversations within an organization may seem unreasonable, and if misused, could be potentially dangerous. Despite many companies having both employees and visitors sign contracts agreeing that their actions can be recorded, enforcing a policy that mandates microphones may raise considerable backlash. Instead of a company-wide declaration that everyone needs to be archiving all conversations, an organization might create an opt-in program that rewards individuals who are using the

system efficiently. Large corporations are spending millions of dollars on knowledge management systems that fail to capture the channel for most knowledge flow. An alternative could be a bonus program for employees who choose to participate in sharing their relevant conversation data. Regardless of a successful opt-in policy however, genuine privacy issues remain. In an attempt to assuage some of these legitimate concerns, several methods of collecting this data will be discussed.

Conversation Postings

For the most privacy conscious, all the conversations can be stored locally on the individual's machine. At the end of the week a topic-spotting algorithm would be used to summarize each conversation, allowing the user to get a list of the week's conversations, the participants involved, and the duration. By each conversation there would be box to check if the conversation is private, public, or should be deleted. If the conversation was relevant to the company and the individual wished to add it to the corporate database, the keywords and features within the conversation would automatically used to update the individual's profile. The time management benefits of being able to archive the details of each conversation could convince people that keeping their microphone on provides enough value to offset the time and privacy issues. Types of environments where the system would flourish could be places where individuals need to keep careful track of how they spend every minute of their day. For billing purposes, law firms could use such a system to account for their lawyers' time.

Ten Minute Delete / Mute Button

Simply put, the device recording the audio could also be equipped with a button to delete the last ten minutes of the recording, or mute the audio for ten minutes into the future. In this way, employees could have a private conversation while at work with a push of the button.

Low Energy Filtering

To insure that other people's voices are not transcribed, a low energy filter can be applied to the audio files after the features are extracted from the conversation. This will essentially eliminate the audio that does not originate from the speaker wearing the microphone. The speaker energy and participant energy in a conversation transcription is shown in Figure (1).

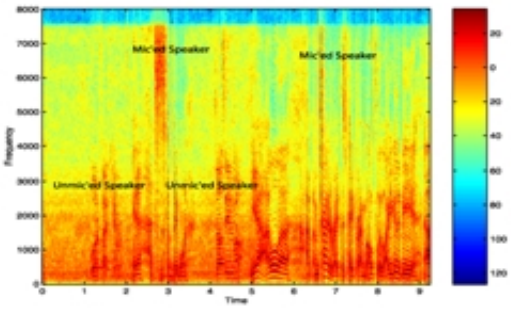

Fig. 1. – Conversation Energy

Demanding Environments

In some instances, the environmental demands may supersede privacy concerns. Intel has expressed interest in prototyping this system within a chip fabrication facility, where each minute of downtime costs the company many thousands of dollars. Environments such as these have minimal private conversations, and the needs for all available information is so great, that many of the privacy concerns may not be relevant. Other testing grounds for such a system could be emergency response teams or military applications.

EXAMPLE RESULTS: CONTEXT FROM AUDIO

Currently, over two hundred conversations have been transcribed and approximately thirty conversations have been labeled with location and the participants. A Markov model was trained on some of the labeled data to identify location based on vocabulary. Although the initial task was fairly straight forward, the results point to the fact that such a model can learn distinguishing words that help identify the context and the people involved using speech recognition engines.

Enabling Software

Leveraging the advances in a commercial speech recognition system, a TCL script was written to interface with the ViaVoice SDK. Recorded conversations in a 22kHz, 16 bit wav file format were placed in a select directory where they were input into the script. The script's output corresponds to the transcribed words from the ViaVoice recognition engine. Along with each word is a time stamp marking the beginning of the utterance and a recognition confidence factor ranging from -100 to 100. The three data types in each output data file (transcribed word, timestamp, and confidence factor) are then input into a program that parses the data into a 3xn cell array. The data is then used by Matlab to train the weighted Markov model that will be discussed in the next section.

The Classifiers

Using word frequency to explicate location is an obvious choice for this classification problem. Intuitively, it is clear there should be many words in our vocabularies that are extremely context dependent. These words are expected to have a frequency that correlates to the social setting. The vocabulary that people use in different contexts can be

clustered into sets that, although they may contain significant overlap, has features that will remain relatively distinct. The zeroth order Markov model simply creates a normalized 5000-by-1 histogram of word frequency for each of the training data (home vs. lab), and finds the log-likelihoods of a test stream of data. The first order Markov model in contrast, considers pairs of neighboring words and places them into a similar 5000-by-5000 histogram.

Both of the Markov models are classifiers whose outputs depend on a log probability to determine the appropriate classification. Pseudovalues are added to the count histograms in order to prevent log(0) problems. Finally, the methods are normalized such that each value corresponds to an actual probability of the word or sequence. Each word in a test set corresponds to one of the 5000 words in the model's working vocabulary. The test set of n words is then turned into a stream of n numbers using a 'key' which correlates every individual word with a number from 1-5000.

Eq 1.) 0th Order Weighted Markov:

$$\sum_{j=1}^{n} \log(count1(stream(j)) * conf(j)^{q})$$

Eq 2.) 1st Order Weighted Markov:

$$\sum_{j=1}^{n-1} \log(count2(stream(j), stream(j+1)) * conf(j)^{q} * conf(j+1)^{q}$$

What distinguishes the two classifiers from standard Markov models is the additional information incorporated from the TCL script related to the confidence of each word generated from the speech recognition engine. The confidence factors are distributed in a Gaussian manner with values ranging from zero to one. By multiplying the confidence factor with the initial count probabilities, the classifier becomes biased towards the words that have a greater probability of being correct. Thus, the classifier attempts to overcome the fact that every other word in the data set is incorrect by shifting the weights of the features. Changing the exponential q on the confidence factor (conf(I)) determines how heavily to weight the confidence data.

After training on approximately 20 hours of data, the classifiers were given 10 hours of test data. Two methods were used to assess the success of each classifier: certainty and speed. Certainty was defined as the differences between the two log-likelihoods. Speed was defined as the number of words required before the classifier converges on what will be a final answer.

Evaluation of Results

The weight of the confidence factor (signified by q in equation 1 and 2) was initially high due to prior thinking that the poor word accuracy from the recognition engine would dramatically degenerate the underlying classification performance. However, it was soon learned that setting the value of q above one had little effect on classification certainty. Although the optimal number varied between test sets, q was finally set to one for the following tests.

When the 0th order approach was compared with the 1st order Markov model it was surprising to see that the 0th order model output a higher certainty than the Markov model in most of the data sets.

Initial data supports the theory that similar location elicits the common vocabulary from multiple people. Over 90% of the data from an individual uninvolved in generating the training data was classified correctly.

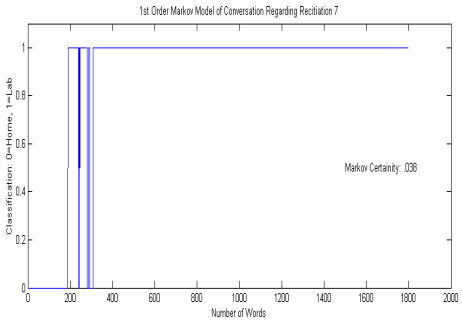

Fig. 2. - Typical data set classified with a 1st order Markov model

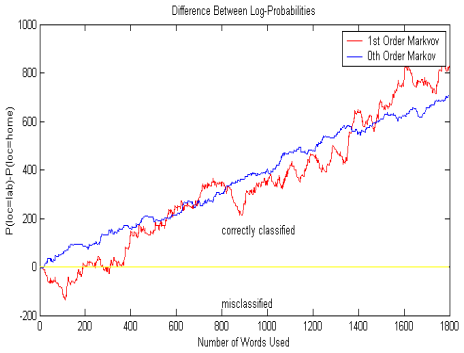

Fig. 3. - Certainty of the 0th and 1st order Markov Models

CURRENT RESEARCH

The results above suggest that people have a unique vocabulary that adapts to a given context. The next step will be to show how an individual's knowledge can be extracted from this type of unique vocabulary.

The goal of our current research is to map a group of people's knowledge of specific technical areas using unobtrusive wearable devices and autonomous keyword scanning software. It is our hope that conversation data, when combined with the email and survey data traditionally used for knowledge management, will yield a considerably more accurate model of an organization's knowledge resources.

Experimental Design

To discover the reality of an organization's knowledge assets, an individual's social behavior is monitored using an unobtrusive wearable device and is analyzed by a software

program that that scans conversation data for the specific keywords. The experiment is designed to unobtrusively learn about people's knowledge of specific subjects while also mapping their social networks. Users will have agreed to wear the device while they are at work. Information we will collect from the device consists initially of audio recorded at 16bit/16 KHz and streamed to a central server. The Zaurus PDA platform was chosen because it supports other sensor inputs which may become appropriate in the future.

To quantify a participant's knowledge, a software program will autonomously scan through spoken conversations, emails and personal web sites for keywords. For example, one such area of specialization will be 'wireless networking' and includes the following ten words:

1. 802.11 2. Bluetooth 3. Transceiver 4. Radio 5. PCMCIA 6. Antenna 7. SNR 8. GSM 9. Fourier 10. Convolution

The participants also have duplicates of their email automatically forwarded to an alias program that scans for the same keywords and then deletes all mail.

The data will be used to define the participant's social network, using conversations features to quantify the type of relationship and keyword spotting to look at the content. These high-level features will be used as training data for a dynamic Bayesian network simulation of organizational behavior currently under development. The idea of simulating the effects of an organizational disruption in existing networks is one that is immediately applicable to every organization. This tool could give managers the ability to see how their organization would react to changes such as merging two departments or relocating a group. Indeed, such a data-driven model offers the potential to transcend the traditional org-chart, perhaps by drawing parallels to ad-hoc network optimization. Forming groups based on inherent communication behavior rather than rigid hierarchy or formal education may yield significant insights to the organizational structure community.

CONCLUSIONS

Simply analyzing conversations of the individuals within an organization can yield significant insight into expertise, social network dynamics and immediate social context. By clustering people based on the profiles from these conversations, one can identify potential collaborations, real-time knowledge, and redundant work. In this way, project teams can be formed from people who have both synergistic skill sets and social behavior. In the not-so-distant future, algorithms will be able to quantify these relationships in even greater detail. Management tools will incorporate this new wealth of 'human' information and provide strategies to optimize organizational effectiveness while maintaining a harmonious work environment.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Allen, T. J. *Organizational Structure for Product Development*, 2002 - Working Papers.
2. Basu, S. "Conversation Scene Analysis," in *Dept. of EECS*. Cambridge: MIT, 2002.
3. Choudhury, T. *The Shortcuts Project.* http://www.media.mit.edu/~tanzeem/shortcuts
4. Choudhury, T. Sociameter: A Wearable Device for Understanding Social Networks. (2002) To be published in the *Proceedings of the Ad hoc Communications and Collaboration in Ubiquitous Computing Environments Workshop*, New Orleans, La.
5. Clarkson, B., Sawhney, N.,and Pentland, A. Auditory Context Awareness via Wearable Computing, (1998) *Proceedings of the Perceptual User Interfaces Workshop*, San Francisco, CA.
6. Jebara,T; Ivanov, Y; Rahimi, A; Pentland, A.. Tracking Conversational Context (1999) in *Media Lab Tech Reports*.
7. Manox, D. Expert Finder, Mitre, 1998. http://www.mitre.org/pubs/edge/june_98/third.htm