

CAPTURING SPONTANEOUS CONVERSATION AND SOCIAL DYNAMICS: A PRIVACY-SENSITIVE DATA COLLECTION EFFORT

Danny Wyatt
Dept. of Computer Science
University of Washington
Seattle, WA 98195
danny@cs.washington.edu

Tanzeem Choudhury
Intel Research
1100 NE 45th St., 6th Floor
Seattle, WA 98105
tanzeem.choudhury@intel.com

Henry Kautz
Dept. of Computer Science
University of Rochester
Rochester, NY 14627
kautz@cs.rochester.edu

ABSTRACT

The UW Dynamic Social Network study is an effort to automatically observe and model the creation and evolution of a social network formed through spontaneous face-to-face conversations. We have collected more than 4,400 hours of data that capture the real world interactions between 24 subjects over a period of 9 months. The data was recorded in completely unconstrained and natural conditions, but was collected in a manner that protects the privacy of both study participants and non-participants. Despite the privacy constraints, the data allows for many different types of inference that are in turn useful for studying the prosodic and paralinguistic features of truly spontaneous speech across many subjects and over an extended period of time. This paper describes the new challenges and opportunities presented in such a study, our data collection effort, the problems we encountered, and the resulting corpus.

Index Terms—Data acquisition, privacy, speech analysis, oral communication

1. INTRODUCTION

The dynamic social network study is an effort to extend traditional social network analysis with automatically collected longitudinal data (data repeatedly collected from the same population over an extended period of time) about real-world interactions. While automatically collected longitudinal data has been used to study virtual interactions [1], few studies have attempted the same in the real-world—despite the fact that face-to-face communication is still people’s dominant mode of interaction [2]. To this end, we instrumented a subject population with wearable sensors that recorded them as they went about their lives over the course of 9 months. Since this data contains truly spontaneous speech recorded *in situ*, we believe this project presents exciting new avenues for speech research.

To date, much of the speech research in modeling conversations has been done on data in meeting room scenarios [3, 4] or with acted speech [5] (which is known to poorly reflect natural emotion [6]). The data sets that do capture real emotion (e.g. [7, 8, 9]) are generally limited to a handful of observations of each subject and cannot be used to compare one person’s speech across different situations or over time. Most are also recorded in relatively unnatural settings (television shows, interviews) that are not representative of ordinary human communication. We have found only one other attempt at collecting data in settings as spontaneous as ours [10], but it only recorded single participants in isolation (i.e. only one side of a conversation) and it does not seem to have continued beyond an initial piecemeal gathering of 250 hours of data.

But the largest problem by far with all of the work we have found is that it all assumes access to raw audio recordings. Collecting truly

spontaneous speech requires recording people in unconstrained and unpredictable situations, both public and private. There is little control over who or what may be recorded. Uninvolved parties could be recorded without their consent—a scenario that, if raw audio is involved, is always unethical and often illegal. Thus, assuming access to raw audio is impractical for most real-world situations and stands in the way of collecting truly natural data. Recording spontaneous data in real-world situations will require protecting the privacy of those involved by not always storing complete audio. While that limits the analyses that can be done on the data, it does not render it useless. A broad range of inferences can be made from privacy-sensitive features. It is also easier to gather far larger amounts of privacy-sensitive data than it is to gather raw audio (which is perhaps a fundamental trade-off between quantity and quality when collecting spontaneous speech data).

In the remainder of the paper we discuss our data collection method, our privacy-sensitive features and how those features can be used, the data we collected, and some problems we encountered along the way.

2. DATA COLLECTION METHOD

The 24 subjects in our study were recruited from the incoming class of the graduate program of a single department at a large research university. To collect data each subject wore an HP iPAQ hx4700 Personal Digital Assistant (PDA) with an attached multi-sensor board (MSB) containing 8 different sensors. The PDA was carried in a small over-the-shoulder bag and the MSB was connected to the PDA via a USB cable that ran discreetly down the bag’s strap (Figures 1(a) and 1(b)). The MSB was worn on the front of the wearer’s shoulder, similar in placement to a lapel microphone (Figure 1(c)). For long-term usability, a close-talking microphone was not appropriate. Recording could be started and stopped with the press of a single hardware button on the side of the PDA and the screen provided simple feedback to show whether the device was recording (Figure 1(d)). The PDA has an Intel XScale PXA270 624 MHz processor, with no FPU, and 64 MB of RAM. All data was saved to a 2 GB Secure Digital flash memory card in half hour chunks. Of the 8 sensors on the MSB, for conversation detection, the most important sensor is clearly the microphone. The MSB’s microphone is an inexpensive electret condenser microphone that records 16 bit audio at a rate of 15,360 Hz. The raw audio was not recorded, but a set of privacy-sensitive features that preserve information about conversation style and dynamics (described in Section 2.1) were computed and saved on the PDA.

The MSB also contains 7 other sensors that sample at varying rates: triaxial accelerometer (550 Hz), visible light (550 Hz), digital compass (30 Hz), temperature and barometric pressure (15 Hz), infrared light (5 Hz), and humidity (2 Hz). These sensors can be used



(a) Front: MSB is on right shoulder (b) Back: PDA is in bag. (c) MSB. Microphone is at top. (d) PDA and data collection program.

Fig. 1: The data collection kit worn by each subject.

to infer the wearer’s physical activity (e.g. walking, sitting, standing, etc.) and whether she is indoors or outside [11]. In addition to the data gathered via the MSB, the PDA records (at 0.5 Hz) the MAC addresses and signal strengths of the 32 strongest WiFi access points nearby. This WiFi data can be used to determine the wearer’s location [12]. Unlike audio, the raw data from the additional sensors and the WiFi readings are saved in their entirety with no initial feature processing.

Data was collected during working hours for one week each month over the 9 month course of an academic year. The first week had only 3 working days and the last only 4, for a total of 42 collection days. Aside from the days and hours, no other restrictions were placed on data collection. The subjects recorded data everywhere they went, inside and out: class, lunch, study groups, meetings, spontaneous social gatherings, etc. Subjects were asked to upload their collected data at the end of each collection day, but because their memory cards could hold an entire week of data most waited until the end of the week. The subjects were paid for each day of data that they submitted. They were also allowed to use the PDA during non-collection weeks and were given the PDA at the end of the study.

At the end of every collection week each subject filled out a survey. The survey asked 5 questions: (1) which other participants the subject interacted with over the previous month and how (e.g. homework collaboration, research collaboration, socially, etc.), (2) which 5 non-participant students within the same department the subject interacted with and how, (3) which sub-areas of the discipline the subject was interested in pursuing, (4) which faculty members the subject was interested in collaborating with, and (5) which faculty members the subject had collaborated with. Once each term, the survey also asked which classes the subject was taking, how she was funded, and whom she considered her advisor. An initial survey administered on the first day of data collection asked questions about the subject’s previous interactions with anyone in the department.

2.1. Privacy-Preserving Audio Features

For our data we devised a feature set that would preserve enough information to allow us to infer when conversations occur between study participants and also include enough prosodic and paralinguistic features to allow us to infer conversation types and speaker states. To protect the privacy of anyone who comes within the range of the microphone, however, we must ensure that the acoustic information that is saved cannot be used to reconstruct intelligible speech. A further challenge that constrains our feature set is that all features must be computed in real-time (no raw audio can ever be stored, even

temporarily) within the limited computational resources of the PDA.

To ensure that we cannot reconstruct content, the features we record do not include formant information. At the same time, they preserve the ability to infer when and how a person is speaking, and—if the speaker is one of our subjects—which person is speaking. We believe that this is possible using an approach based on detecting regions of audio that contain speech—and voiced speech in particular. Since our data can be recorded in widely varying noise situations, it is important that our features are robust to noise. Mean and variance of spectral entropy have been shown to be a good features for distinguishing speech from non-speech in noisy conditions [13]. Additional features that have been shown to be useful in robustly detecting voiced speech under varying noise conditions are: (1) non-initial maximum autocorrelation peak, (2) the total number of autocorrelation peaks, and (3) relative spectral entropy [14].

To compute these features, we use 33.33 ms frames (512 samples) with overlaps of 16.67 ms (256 samples). For each frame, we first compute its power spectrum and total energy. We normalize the power spectrum to a distribution by dividing by the total energy and use that distribution to compute spectral entropy. Relative spectral entropy is computed between the normalized spectrum of the current frame and a normalized running average of the power spectra of the last 500 frames (8,335 ms).

When using autocorrelation peaks to detect voiced speech, narrow spectrum noise (e.g. air conditioning, hard drives, etc.) can cause confusion because it also creates strong autocorrelation peaks. Fortunately, such noise is usually low energy (compared to voiced speech) and its autocorrelation can be disrupted by adding low energy white noise to the signal. To that end, we uniformly whiten the power spectrum with additional energy equal to 1% of the maximum energy possible per frame, and then compute the autocorrelation in the frequency domain. Logarithms (required for entropy) are impractical given the PDA’s limited processing power. However, the PDA’s comparatively large amount of RAM allows us to instead use a lookup table pre-populated with logarithms for all 16 bit values.

Summarizing, the complete list of acoustic features we compute and save are: (1) spectral entropy, (2) relative spectral entropy, (3) non-initial maximum autocorrelation peak (both its value and lag), (4) the total number of auto-correlation peaks, and (5) energy.

2.2. Types of Inference

While the set of features may sound limited, it still allows a broad range of inference. Mean and variance of spectral entropy can be used to find spoken regions [13]. Relative spectral entropy and autocorrelation peaks can be used to find voiced regions [14] and spoken

regions [15]. And we have had promising early results using only energy for speaker diarization (inferring who is speaking when) [16]. From the autocorrelation peaks we can infer pitch [17], and pitch and energy are the two features most commonly used to infer emotion from speech [18].

Knowing who is speaking when, with what energy, and at what pitch is enough to enable many classes of dialog understanding. Interruptions and speaking time reveal information about status and dominance [19]. Speaking rate reveals information about a speaker’s level of mental activity [20]. Energy can reveal a person or group’s interest in the conversation [21]. Pitch alone has a long history as a fundamental feature for inferring emotion [22], and energy and duration of voiced and unvoiced regions are also informative emotional features [23].

2.3. Ground Truth

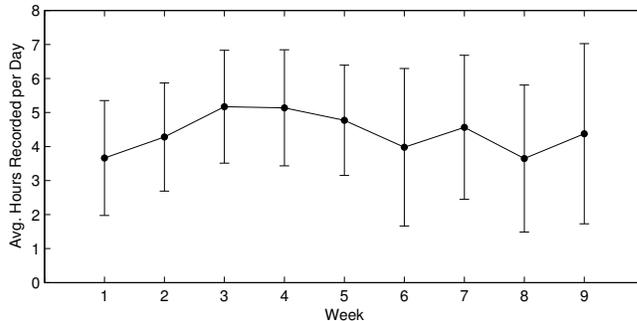
The above inferences can be separated into two classes: low level speaker diarization and conversation detection, and higher level inference of emotion, interest, social role and status, etc. To support work on the first class of low level inference, we have gathered a smaller development set of data (on the order of hours) with the raw audio preserved. This data has been labeled with the ground truth of who was speaking when, and who was in conversation with whom. These labels can be used to test the accuracy of any speaker-independent diarization or conversation detection algorithms that are to be used on the larger, unlabeled data set.

Ground truth for the second class of higher level inferences is more difficult to collect. Since capturing spontaneous conversations under the given privacy-constraints requires recording many hours of data without access to the raw audio, it is infeasible to label even a subset of the data after its collection. Nevertheless, we do have coarse ground truth from the survey data that we collected. We know the subjects’ self-reported interactions and how they categorize those interactions. That information can be used to guide analysis of the acoustic data.

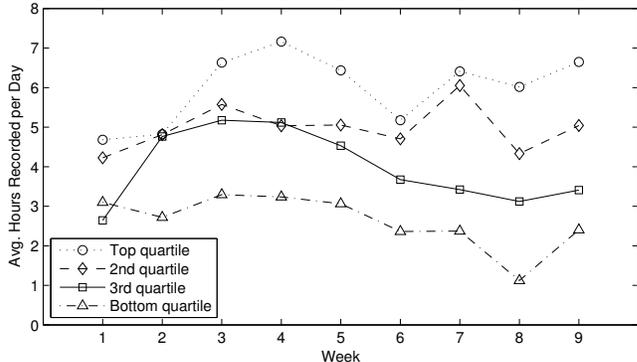
To be sure, this is not a corpus amenable to traditional research that classifies utterances and evaluates the classification according to correspondence with ground truth (e.g. speech recognition). This is a corpus for studying broad questions about prosody in a large amount of spontaneous speech, for example: Does prosody vary with social relationship? Does prosody change over time, and is that change correlated with a change in relationship? Can conversations be clustered according to prosody? Is there a correlation between those clusters and conversation location? In general, this data set will lend itself more to such unsupervised inference and exploratory analysis. As suggested by [5], such analysis in turn can enable the bootstrapping of new corpora that mix the prosodic features of truly spontaneous speech with less spontaneous acted data.

3. RESULTING DATA

Our subjects gathered a total of 4,469.48 hours—186.23 days—of data. The amount of data collected per participant varied greatly, from a maximum of 321.64 hours to a minimum of 88.78 hours. On average, each participant recorded 4.43 hours of data per collection day (max: 7.66, min: 2.11). The amount of data recorded also varied over the course of the study. Figure 2(a) shows the average hours collected per day for each collection week, averaged over all subjects. For Figure 2(b), subjects were ranked into quartiles according to how much data each subject submitted for the entire year. The average number of hours recorded per day was then computed for each quartile.



(a) Average over all subjects. Error bars show standard deviation.



(b) Average over subjects ranked in quartiles by total amount of data submitted.

Fig. 2: Average hours of data collected per day per subject.

The first three weeks show an increase in the amount of data collected as the subjects became more comfortable with the device and its use, and battery life was improved (see Section 3.1). We believe that collected amounts decrease in weeks 4 through 6 as the participants become fatigued and the study becomes less novel. This decline is more pronounced in the subjects who submitted less data overall. (The sharp drop in data collected by the top quartile at week 6 is due to one subject not being able to collect data that week. With that anomaly removed the top quartile’s mean for week 6 rises from 5.18 to 6.21, making the overall trend more stable.) Before weeks 7 and 9 we sent additional messages of encouragement to the cohort, and those may be responsible for the subsequent upturns in collection amounts.

Since subjects were only asked to record data during what each considered his working day, and since different graduate students keep very different hours, we worried about overlap in recording times. A few subjects started their days much earlier than others and spot checks of their data suggested that their PDAs’ batteries were dying at the time that most of the cohort was together. To counter this, we asked that no subject begin recording any earlier than 10 a.m. for the last 3 collection weeks. Figure 3 shows the average number of simultaneous recordings at each collected second for each collection week. From this it appears that changing the collection method did not effect the amount of overlap in the data. Indeed, overlap is strongly correlated with the amount of data collected ($r^2 = 0.69$, $p = 0.006$) more than it seems to be influenced by strict collection hours. While there is no moment when all subjects are recording (the maximum number of simultaneous recordings is 21), there is enough overlap in the data for it to contain many possible conversations. The average number of simultaneous recordings per second is 8.133 for the entire corpus, and 88.32% of all recorded

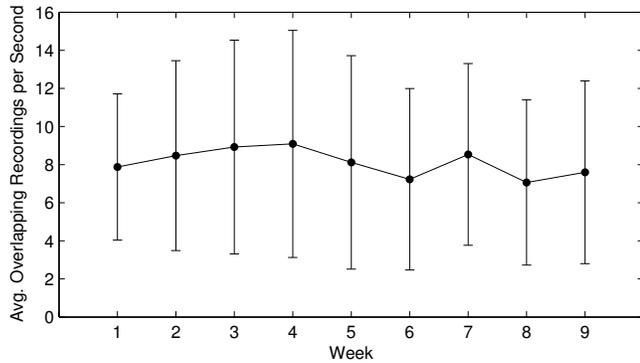


Fig. 3: Average number of subjects recording simultaneously at each collection second. Error bars show standard deviation.

seconds are covered by at least two recordings. Additionally, there is overlapping data for all pairs of subjects.

3.1. Problems Encountered

We encountered four significant problems during data collection. (1) Batteries died faster than anticipated. We discovered that the operating system was attempting to connect to known WiFi networks in weak signal conditions that we had not previously tested. We alleviated this problem by reconfiguring the OS to never attempt to connect to any network while the data collection application was running. (2) All of the PDA's software and settings are stored in volatile RAM and are completely lost if the battery fully discharges. Subjects found it easy to recharge their PDAs at the end of each collection day, but would often forget to charge them between collection weeks. This led to many Monday mornings of lost recording time while PDAs were reconfigured. (3) The PDAs' clocks are shockingly unreliable. We found them to drift up to 5 minutes between collection weeks, thus needing frequent resynchronization with a time server—which required periodically re-enabling the WiFi connection. (4) The cable that connected the MSB to the PDA's USB card was not durable enough for many weeks of continuous use. Over time, the cable would become loose, and the PDA would intermittently lose its connection to the MSB.

Each of these problems ultimately arises from our stretching the PDA well beyond its intended use. It is meant to be turned on only sporadically for short tasks, not to run continuously as its user goes about her day. The PDA is also intended to be attached to a computer regularly, providing it with the opportunity to charge its battery and synchronize its clock. While PDAs are handy portable platforms for short data collection efforts, we conclude that they are not suitable to long term collection efforts such as ours.

4. CONCLUSION

Our longitudinal data collection effort resulted in a corpus of over 4,400 hours of spontaneous interaction data among 24 subjects, sampled monthly over the course of 9 months. The recordings contain a set of acoustic features which preserve information about voicing and prosody that is useful in analyzing conversational dynamics, style, emotional content, and (at a higher level) social dynamics, but cannot be used to reconstruct intelligible speech. Saving only these privacy-preserving features was necessary for our wide scale deployment, and we believe that any extensive collection of *in situ* spontaneous speech must consider privacy and the attendant trade-off between quantity and quality. In addition to the acoustic features, the dataset contains the raw sensor data from the seven other sensors

on the MSB, and MAC addresses and signal strength of the WiFi access points near the user. After additional privacy reviews, we plan to freely share this data with the broader research community.

5. REFERENCES

- [1] G. Kossinet and D. J. Watts, "Empirical analysis of an evolving social network?" *Science*, vol. 311, pp. 88–90, 2006.
- [2] N. Baym, Y. B. Zhang, and M. C. Lin, "Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face," *New Media and Society*, vol. 6, pp. 299–318, June 2004.
- [3] NIST, "NIST rich transcription evaluations - <http://www.nist.gov/speech/tests/rt/rt2006/spring/>," 2006.
- [4] A. Dielmann and S. Renals, "Multi-stream segmentation of meetings," in *Proc. IEEE Workshop on Multimedia Signal Processing*, 2004.
- [5] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, 2003.
- [6] A. Batliner, K. Fisher, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *Proc. of the ISCA ITRW on Speech and Emotion*, 2000.
- [7] P. Greasley, J. Setter, M. Waterman, C. Sherrard, P. Roach, S. Arnfield, and D. Horton, "Representation of prosodic and emotional features in a spoken language database," in *Proc. of the XIIIth ICPHS*, 1995.
- [8] E. Douglas-Cowie, R. Cowie, and M. Schroeder, "A new emotion database: considerations, sources and scope," in *Proc. of the ISCA ITRW on Speech and Emotion*, 2000.
- [9] J. Ang, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. of ICSLP*, 2002.
- [10] N. Campbell, "The recording of emotional speech: JST/CREST database research," in *Proc. LREC*, 2002.
- [11] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A hybrid discriminative-generative approach for modeling human activities," in *Proc. of IJCAI*, 2005.
- [12] B. Ferris, D. Haehnel, and D. Fox, "Gaussian processes for signal strength-based location estimation," in *Proc. of Robotics: Science and Systems*, 2006.
- [13] A. Subramanya, J. Bilmes, and C.-P. Chen, "Focused word segmentation for ASR," in *Proc. of Eurospeech*, 2005.
- [14] S. Basu, *Conversational Scene Analysis*, PhD Thesis, MIT, 2002.
- [15] S. Basu, "A linked-HMM model for voicing and speech detection," in *Proc. of ICASSP*, 2003.
- [16] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, "A privacy sensitive approach to modeling multi-person conversations," in *Proc. of IJCAI*, 2007.
- [17] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 24–33, 1977.
- [18] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *Proc. of Eurospeech*, 2003.
- [19] K. Hawkins, "Some consequences of deep interruption in task-oriented communication," *J. Language and Social Psychology*, vol. 10, pp. 185–203, 1991.
- [20] R. Hurlburt, M. Koch, and C. Heavey, "Descriptive experience sampling demonstrates the connection of thinking to externally observable behavior," *Cognitive Therapy and Research*, vol. 26, no. 1, pp. 117–134, 2002.
- [21] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. of ICASSP*, 2005.
- [22] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. of ICSLP*, 1996.
- [23] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," in *Proc. of ICASSP*, 2004.