A Privacy-Sensitive Approach to Modeling Multi-Person Conversations

Danny Wyatt
Dept. of Computer Science
University of Washington
danny@cs.washington.edu

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

Jeff Bilmes
Dept. of Electrical Engineering
University of Washington
bilmes@ee.washington.edu

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

Abstract

In this paper we introduce a new dynamic Bayesian network that separates the speakers and their speaking turns in a multi-person conversation. We protect the speakers' privacy by using only features from which intelligible speech cannot be reconstructed. The model we present combines data from multiple audio streams, segments the streams into speech and silence, separates the different speakers, and detects when other nearby individuals who are not wearing microphones are speaking. No pre-trained speaker specific models are used, so the system can be easily applied in new and different environments. We show promising results in two very different datasets that vary in background noise, microphone placement and quality, and conversational dynamics.

1 Introduction

Automatically modeling people's spontaneous, face-to-face conversations is a problem of increasing interest to many different research areas. Yet there is very little data available that captures truly spontaneous speech—speech recorded *in situ* as people go about their lives. Portable devices capable of such recording have grown in storage capacity while becoming smaller, cheaper, and more powerful. But obstacles to gathering spontaneous speech still remain, and perhaps no other obstacle is as prominent as privacy.

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. Recording spontaneous data in real-world situations will require protecting the privacy of those involved by not always storing complete audio. More specifically, any data that is saved must not allow the linguistic content of a person's speech to be reconstructed.

While that limits the analyses that can be done on the data, it does not render the data useless. A broad range of inferences can be made from privacy-sensitive features. There are many applications that would benefit from increased access to spontaneous speech data while not needing to know the content of the speech.

For example, research in speech and emotion often uses only information about pitch, volume, or duration [Schuller *et al.*, 2004]. But the data used in such research has been either acted speech [Campbell, 2000] or datasets gathered in constrained situations [Greasley *et al.*, 1995; Douglas-Cowie *et al.*, 2000; Ang, 2002]. Acted speech is known to poorly reflect natural emotion [Batliner *et al.*, 2000]; and the constrained datasets are recorded in relatively unnatural settings (television shows, interviews) that are not representative of ordinary human communication. There is a demand for more natural data sets for the study of speech and emotion [Douglas-Cowie *et al.*, 2003a].

A second example is the study of social networks. Traditional social network analysis has relied on data gathered either through surveys, which are vulnerable to known biases [Bernard *et al.*, 1979; Marsden, 1990], or third party observers, which is costly, labor intensive, and does not scale. Recent studies have used automatically gathered data about on-line interactions [McCallum *et al.*, 2005; Kossinet and Watts, 2006], but there are few studies involving automatically recorded face-to-face conversations—despite the fact that face-to-face communication remains people's dominant mode of interaction [Baym *et al.*, 2004]. To study social networks, it is sufficient to know only who spoke with whom, not what was said.

Finally, non-linguistic aspects of spoken communication are also useful features in medical and meeting understanding applications. Speaking rate is an indicator of mental activity [Hurlburt *et al.*, 2002] and a behavioral symptom of mania [Young *et al.*, 1978]. Abnormal conversation dynamics are symptoms of Asperger syndrome [Wing and Gould, 1979] and autistic individuals often speak in a high-pitched voice or lack intonation [Tager-Flusberg, 1994]. In meetings, interruptions and speaking time can reveal information about

status and dominance [Hawkins, 1991] and gender specific differences in interruptions and the consequences of those differences are active topics of research [Tannen, 1993]. None of these features require access to linguistic content, and all of these applications would benefit from increased access to natural speech data.

1.1 Problem Description

Our specific long-term goal is to model the evolution of spontaneous face-to-face interactions within groups of individuals over extended periods of time. In order to protect the privacy of both research subjects and the non-subjects with whom they come into contact, we must ensure that the acoustic information that is saved cannot be used to reconstruct intelligible speech. The stored features must contain enough information to serve as input to models of conversational and social dynamics, but at the same time have insufficient information to reconstruct what is being said, or to positively identify individuals who are not wearing microphones.

The work presented in this paper is the first step toward our goal. We present an unsupervised approach to separating speakers and their turns in a multi-person conversation, relying only on acoustic features that do not compromise privacy. The features employed are useful for modeling conversational dynamics—who is speaking and how—but are not sufficient for speech recognition.

Our work is novel in several ways. The key contribution is a joint probabilistic model that combines streams of acoustic features from a set of individuals wearing microphones, infers when there is speech present, separates the different speakers from each other, and also detects when other individuals around them—who are not equipped with microphones—are speaking. It does not require pre-trained speaker specific models, and thus scales to any number of users and can be used with new speakers and in new environments. Our model can be extended to a dynamically varying number of speakers, where new audio streams come and go whenever a new person enters or exits a group. We also introduce a novel feature set that is useful for segmenting speakers and for modeling conversation attributes, but that cannot be used to transcribe the actual words spoken during a conversation.

1.2 Related Work

Most of the work in modeling spoken conversations has been done in the domain of meeting understanding [NIST, 2006; Dielmann and Renals, 2004]. One of the goals in meeting understanding is speaker diarization: determining who spoke when [Reynolds and Torres-Carrasquillo, 2005]. All of the work that we have found in this domain assumes access to the full audio and that it is not necessary to remove information that can be used to transcribe speech.

Much previous work has been done in linguistic conversation analysis [Ochs *et al.*, 1996; Sacks, 1992]. Clearly, that work relies almost solely on information about the words that are spoken, and not the basic acoustics of speech. As such, the model presented in this paper is complementary to traditional conversation analysis. However, inasmuch as conversations are considered turn taking between speakers, our model's ability to infer speaker turns can be seen as a necessary low-level enabler for higher level conversation understanding.

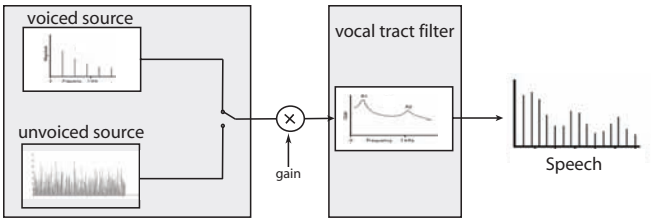

Figure 1: The source-filter model for speech production.

Finally, as previously mentioned, there has been much research into recognizing emotions associated with speech [Douglas-Cowie *et al.*, 2003b]. Many of these emotion-recognition applications may not need to know the words that are spoken. [Shriberg, 2005] mentions the importance of modeling natural speaking behavior and identifies it as a fundamental challenge for spoken language applications.

2 Speech Features and Privacy

We begin by giving a very simple description of speech production based on the source-filter model [Quatieri, 2001] (see Figure 1). Most speech sounds can be modeled with two independent components: (i) the source sound generated in the glottis and (ii) the filter (the vocal tract) that shapes the spectrum of the source sound. The source can be either voiced with fundamental frequency F0 (the pitch) or unvoiced with no fundamental frequency. Prosodic information about speech (intonation, stress, and duration) is described by how the fundamental frequency and energy (volume) change during speech. The frequency response of the vocal tract—particularly the resonant peaks (the formants)—contains information about the actual phonemes that are the basis for words. To reproduce speech intelligibly, information on at least three formants is required [Donovan, 1996]. Any processing of the audio that removes information about the formants will ensure that intelligible speech can not be synthesized from the information that remains, and privacy will be preserved.

To detect speech (and specifically, voiced speech) and model how something is being said, we extract features that contain information about the source and prosody but not about the formants. Three features that have been shown to be useful for robustly detecting voiced speech under varying noise conditions are: (i) non-initial maximum autocorrelation peak, (ii) the total number of autocorrelation peaks and (iii) relative spectral entropy [Basu, 2002] computed on 16 ms chunks of audio. Because of the periodic components of voiced speech (see Figure 1), the autocorrelation will have a small number of sharp peaks. Similarly, the relative spectral entropy between the spectrum at time t and the local mean spectrum (800 ms window) will be high for voiced speech even in the presence of indoor and outdoor noise (e.g. wind, fan).

Beyond detecting the spoken regions, the system needs additional information to separate the different participants in

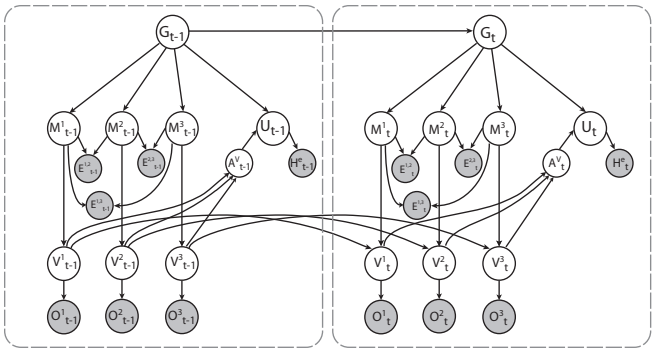

Figure 2: DBN model for multi-person conversation

the conversation. We found two features to be useful for this purpose: (i) the absolute energy, and (ii) the entropy of the energy distribution across the different microphones (described in more detail in Section 3).

Summarizing, the complete list of acoustic features that must be saved for our model are: (i) non-initial maximum autocorrelation peak, (ii) the total number of autocorrelation peaks, (iii) relative spectral entropy, and (iv) energy.

3 Multi-Person Conversation Model

Let us assume there are N individuals wearing microphones. Given acoustic features from these N microphones, we want to detect when one of the wearers is speaking as well as when the microphones are picking up speech from others in the area not wearing microphones. A dynamic Bayesian network (DBN) [Dean and Kanazawa, 1988] is a flexible way to combine all of the features into a unified model used to infer who is speaking when. The state factorization of DBNs makes it relatively simple to express complex dependencies between different variables in the system. Figure 2 depicts our DBN model for inferring both spoken segments as well as identifying the speaker of those segments ($N = 3$ in this example). The shaded nodes are the observed variables whose values are inputs to the system and the hidden nodes are the variables whose values are to be inferred.

Each time step in the DBN corresponds to a small chunk, or frame, of audio data. In all of our experiments we used frames with a length of 33.33 ms 16.67 ms of overlap with the previous frame.

Below we describe how the different variables are specified and the dependencies between them.

Group state $\mathbf{G}_t$

The group node G_t determines who is holding the floor or taking a turn to speak. It is a discrete random variable with cardinality $N + 2$: one state for no speaker (silent regions), one state for each of the N people wearing microphones, and one state for any other speakers not wearing microphones. The group state G_t depends on G_{t-1}, and the conditional probability $P(G_t|G_{t-1})$ encodes the probability of turn transitions between speakers. At time $t = 0$ all states are equally likely: $P(G_0) = 1/(N + 2)$. The group node allows us to constrain the individual states described below, and reduces the complexity of the conditional probability table (CPT) we would otherwise have to model to capture dependencies between speakers.

State of the individuals wearing microphones $\mathbf{M}_t^i$

The binary random variable M_t^i indicates whether the i-th individual wearing a microphone is speaking. The conditional probability $P(M_t^i|G_t)$ is set to be semi-deterministic: $P(M_t^i = 1|G_t = g) \approx 1$ when $g = i$, and ≈ 0 otherwise. This imposes the constraint that—most of the time—people do not talk simultaneously during a conversation. Note that it is still possible, though highly unlikely, for multiple M_t^i variables to be true while G_t is held to a single speaker.

State of unmiked others $\mathbf{U}_t$

Similar to M_t^i, the unmiked other node U_t is a binary random variable that indicates whether anyone not wearing a microphone is speaking. If there are multiple unmiked persons present, they are all modeled by this node. U_t is conditioned on the group node G_t and the aggregate voicing node A_t^V which indicates whether any microphone detected voiced speech (and is described in more detail below). The conditional probability $P(U_t|G_t, A_t^v = 1)$ is defined identically to $P(M_t^i|G_t)$, and $P(U_t = 1|G_t, A_t^v = 0) \approx 0$.

Voicing states $\mathbf{V}_t^i$ and aggregate voicing $\mathbf{A}_t^V$

The voicing states V_t^i are binary variables that indicate whether microphone i has recorded sound consistent with voiced human speech. The parents of V_t^i are M_t^i and the previous V_{t-1}^i. Since each microphone can pick up speech from its wearer as well as other speakers nearby, the conditional probabilities of the V_t^i nodes are defined as $P(V_t^i = 1|M_t^i = 1) \approx 1$ and $P(V_t^i = 1|M_t^i = 0) = 0.5$. In other words, if person i is speaking it is highly likely her microphone will record voiced speech, and if she is not speaking there is a uniform probability that her microphone will record voiced speech.

The node A_t^V is an aggregate voicing node that is the deterministic logical OR of all the V_t^i nodes. We describe below how A_t^V helps distinguish other individuals speaking from silent regions.

Observations $\mathbf{O}_t^i$, $\mathbf{E}_t^{i,j}$, and $\mathbf{H}_t^e$

The observed variables obtained from the acoustic features of the N microphones are included at various points in the DBN as children of M_t^i, U_t, and V_t^i.

O_t^i is a three-dimensional variable that includes the three features previously mentioned as having been useful for detecting voiced speech (non-initial maximum autocorrelation peak, number of autocorrelation peaks, and relative spectral entropy). $P(O_t^i = o|V_t^i = v)$ is modeled by a 3D Gaussian with full covariance matrix. The $P(O_t^i|V_t^i)$ parameters are learned from a set of labeled data (where V_t^i is given) containing speakers who are not present in any of the data we evaluated here. Learning these features in this manner has been shown to be speaker-independent and robust across different environmental conditions [Choudhury and Basu, 2004].

$E_t^{i,j}$ is a two-dimensional variable containing the log energies of microphones i and j averaged over a 333 ms window centered at time t. The conditional distribution

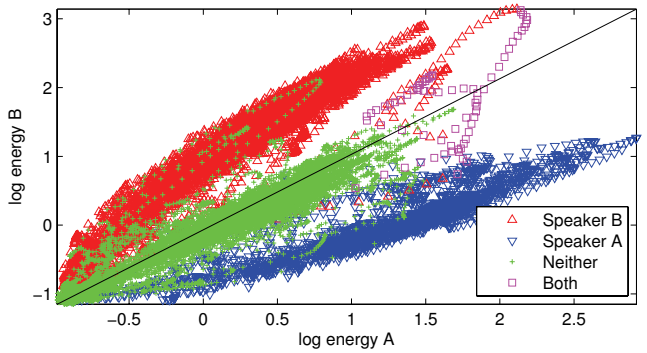

Figure 3: Pairwise log energies.

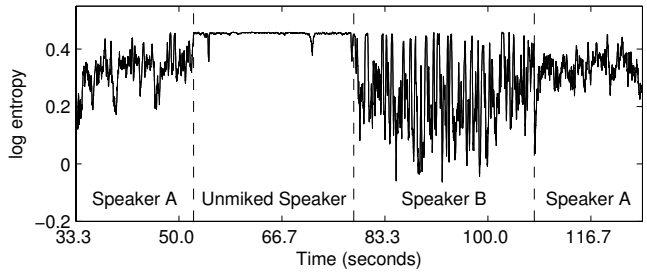

Figure 4: Log entropy of energy distribution across microphones.

$P(E_t^{i,j}|M_t^i, M_t^j)$ is modeled with a full covariance 2D Gaussian. This pairwise energy feature associates voiced regions with their speaker. If person i speaks at time t, then her energy should be higher than j's (and vice versa). When both i and j speak, both of their microphones have high energy and when neither speaks both of their microphones have low energy. Figure 3 illustrates an example of this.

H_t^e is the entropy of the log-energy distribution across all N microphones. This feature is useful for determining whether voiced regions come from a speaker not wearing a microphone. When a person wearing a microphone speaks, his microphone will be significantly louder than the others' and the entropy will be low. When a person not wearing a microphone speaks, her energy will be spread more uniformly across all microphones and the entropy will be high. Figure 4 illustrates an example of this.

H_t^e is computed as follows. First, all microphones' energies are normalized to a distribution: $P_e(\tau) = \frac{e_i(\tau)}{\sum_j e_j(\tau)}$. Then H_τ is computed as the entropy of $P_e(\tau)$. Finally, H_t^e is the average of H_τ over a 333 ms window centered at time t. We use the log of H_t^e and model that with a 1D Gaussian random variable conditioned on U_t.

Although H_t^e is a useful feature for distinguishing whether speech comes from a person wearing a microphone or from one who is not, it does not help distinguish between when an unmiked person is speaking and when no one is speaking. Entropy is high in both of those cases. However, information about the voicing states V_t^i taken together with the entropy can distinguish those situations. If there is voicing and the entropy is high then it is likely that someone else (i.e. someone not wearing a microphone) is speaking. It is here that the aggregate voicing node A_t^V is useful. We define $P(U_t|G_t = u, A_t^V = 1) \approx 1$ (where u is the state of node G_t that indicates an unmiked person is speaking), and $P(U_t|G_t, A_t^V) \approx 0$ if $G_t \neq u$ or if $A_t^V = 0$. Loosely, this means that the model will only infer an unmiked speaker if at least one microphone picked up voiced speech and that speech cannot be assigned to any of the miked speakers.

3.1 Parameter Learning and Inference

Learning is done in an entirely unsupervised manner using expectation maximization (EM). Unsupervised learning is important for this application, given the privacy constraints associated with recording spontaneous speech. Raw audio will not be available for labeling speaker-specific data, so the model must be able to fit itself to unlabeled data.

However, with a large number of parameters, EM can often converge to values that do not result in accurate inferences. To prevent this, we clamp most of the above parameters to their pre-defined or pre-trained values. Indeed, only the Gaussians associated with the energy-based observations ($P(E_t^{i,j}|M^i, M^j)$ and $P(H_t^e|U_t)$) are learned during EM. (As mentioned, the Gaussians associated with the voicing observations are pre-trained in a speaker-independent way.) All of the transition probabilities and semi-deterministic conditional probabilities are fixed at predefined values. We did experiment with leaving more parameters free and the resulting inferences were much less accurate.

Once the free parameters are learned with EM, exact inference is done using the junction tree algorithm. During decoding, we infer the most likely state sequence for the group node G, speaker nodes M_t^i and U_t, and voicing nodes V_t^i. We use the Graphical Models Toolkit (GMTK) for all of our learning and inference [Bilmes and Zweig, 2002].

4 Experiments and Results

Experimental evaluations were performed on two datasets: (i) the publicly available scripted meeting corpus (M4) from IDIAP [McCowan *et al.*, 2003] and (ii) a labeled dataset of natural interactions that we collected. The M4 corpus contains 27 four person meetings, each of them about five minutes long. The dataset has audio recordings from 12 microphones—one 8 microphone array and 4 lapel microphones. The speakers in each recording followed a script for certain meeting-wide activities (e.g., discussion, argument, monologue) but were not told what to say. In our experiments, we used data only from the lapel microphones as it most closely resembles our data collection setup. Overall, this dataset is quite clean and does not have much background noise. We evaluated the performance of our model on 13 randomly selected meetings from this corpus.

The dataset we collected is much more challenging. It has a significant amount of background noise and distant speech. There are 6 conversations collected in 4 different locations: a meeting room, an elevator, a hallway, and a loud and noisy atrium. The speakers were told where to go but not what to talk about. They are all friends and had no trouble filling

Figure 5: Two people wearing our portable recording equipment. The sensing unit is at their right shoulders.

the time with spontaneous conversation. For recording, we used an inexpensive condenser microphone, which is part of a multi-modal sensing unit (dimensions: 60 mm x 30 mm x 25 mm) [Welbourne *et al.*, 2005]. The sensing unit is clipped to the strap of a small over the shoulder bag. The unit sits near the upper right shoulder, but can move so the microphone is not always at a fixed location from the mouth. A PDA in the bag records the audio data. Figure 5 shows two people wearing the equipment. Unlike the M4 data, where all lapel microphones are tethered to the same recording computer, each person in our data carries with her all the equipment needed for recording. Thus, our participants can move about independently and interact in a more natural manner.

To evaluate the speaker segmentation performance, we learned the unclamped parameters of the model in an unsupervised manner for each meeting independently. Once the learning was done, we inferred the most likely state sequence for the group state node G_t and speaker nodes M_t^i and U_t.

We compute four evaluation metrics that compare the inferred value of G_t to ground truth. (i) The per frame error rate is the fraction of frames in which the value of G_t does not match the ground truth speaker. We do not consider frames that have more than one ground truth speaker. (ii) The diarization error rate (DER) is a standard metric used by NIST [NIST, 2006] to measure the performance of speaker segmentation systems. It is a relaxed version of frame error rate that merges pauses shorter than 0.3 s long and ignores 0.25 s of data around a change in speaker. These relaxations account for perceptual difficulties in labeling speech at such a fine time granularity. (iii) Precision is the fraction of the total number of inferred speaker frames that are correct. (iv) Recall is the fraction of truly spoken frames for which any speaker is inferred (in other words, the accuracy of basic speech-detection).

The results for the M4 corpus are in Table 1 and the results for our dataset are in Tables 2(a) and 2(b). In both datasets, all participants wear microphones. To test the performance of our model with unmiked speakers, we selectively ignored the data from some participants' microphones. Results shown for

mics	frame err.	DER	prec.	recall
4	19.48	15.94	83.27	95.57
3	21.62	18.15	81.22	95.52
2	22.98	19.79	79.82	95.33

Table 1: Results for the M4 corpus. All meetings had 4 participants.

speakers	mics	frame err.	DER	prec.	recall
4	4	31.73	24.10	68.18	95.69
	3	31.27	23.72	70.74	95.23
	2	30.52	23.18	70.45	96.25
3	3	33.07	27.60	65.49	90.17
	2	35.47	29.99	64.09	90.66
2	2	28.23	17.47	72.49	93.86

(a) Quiet environments.

speakers	mics	frame err.	DER	prec.	recall
4	4	42.80	39.70	56.27	95.35
	3	44.77	40.79	54.55	94.60
	2	46.28	41.87	53.82	96.02
3	3	23.96	13.56	76.77	98.52
	2	25.23	14.91	75.52	98.90
2	2	40.87	29.59	60.92	95.98

(b) Noisy environments

Table 2: Results on our data.

cases with fewer microphones than speakers are the averages of results for all permutations of that number of microphones across that number of speakers.

The first thing to note is that the DER scores on the M4 data are comparable with current speaker diarization results. 18.6 is currently the best DER (achieved with features that do not preserve privacy) for meeting data [NIST, 2006]. (Unfortunately the dataset used in that evaluation is not generally available, so we cannot compare our results directly.) To the best of our knowledge, there is only one other published diarization result for the M4 corpus [Ajmera *et al.*, 2004]. That technique has better frame error rates (7.4%) but lower precision and recall (their reported average of the two is 80.8, our average of the two is 88.5)—thus making for an inconclusive comparison. That technique also used features (low order cepstral coefficients) that contain information about the words spoken, so it does not protect privacy.

Our error rates are significantly better on the M4 dataset than on the dataset we collected. Our dataset, however, has substantially more difficult characteristics than M4. Our data includes significantly more background noise, and the conversations are more fluent and fast paced with much more speaker overlap. For example, the M4 data's mean turn duration is 6.5 s (median: 2.5 s). For our dataset the mean turn duration is 1.52 s (median: 1.1 s). Clearly, more work remains to be done in handling noisy environments and conversations with faster (or even variable) pacing, but the results are still quite promising.

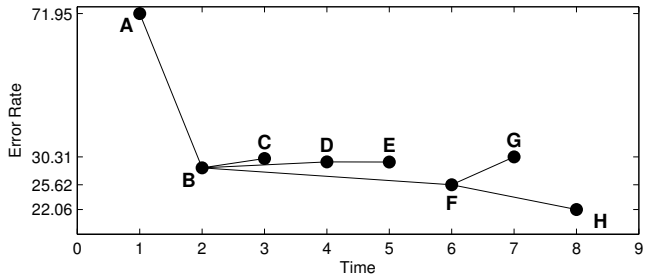

Figure 6: Model history

5 Discussion

Our goal of separating the different speakers in conversations and identifying when they speak while not preserving information from which intelligible speech can be reconstructed led us to explore many different variations of the model proposed in this paper. Having the correct graphical structure and acoustic features is critical in solving this problem.

A number of variations were attempted along the way. This is depicted in the development history plot shown in Figure 6. We started by using only pairwise relative energy ratios along with the voicing features for our observations (A). This performed quite poorly, failing in cases where no one was speaking and small absolute changes in energy caused large changes in their ratios. In (B), we modeled the full distribution of pairwise energies which significantly improved our error rates. Next, we experimented with the effect of making the individual speaker variables M_t^i deterministically dependent on G_t which worsened things (C) and lead to the semi-deterministic CPT we ultimately used. In (D) and (E) we tried adding various energy statistics as children of U_t to determine when unmiked persons were speaking. For (D) we added the maximum microphone energy, and in (E) we added the mean and variance of all the energies, but these models were not appreciably better. Removing these energy observations and using an aggregate voicing indicator A_t^V alone in model (F), did help—(F) is 3% lower than (B). Model (G) was an attempt to add back the mean and variance of energy, but those features (even after normalization) continued to hurt performance. This suggests that those features are not sufficiently discriminative which causes EM to converge to unhelpful parameter values. Lastly, model (H) uses the smoothed log entropy of the energy distribution, as described earlier in Section 3.

6 Future Work

There are many ways that our model could be extended. Currently, it requires at least two microphones, and we are exploring ways to allow it to work with only one microphone. And while the model is theoretically capable of inferring overlapping speakers, the semi-deterministic conditional probabilities from the G_t node prevent it from reliably doing so. We are exploring ways of adding a switching node [Bilmes, 2000] to explicitly represent changes in speaker turns as well as interruptions and interjections. Additionally, we have not experimented with very many different values of the pre-defined parameters. It may be possible to learn some of these from other data, as we do with the voicing parameters, or to allow EM to learn them within given bounds. Learning better turn transition probabilities will probably help our results the most, since turn lengths seem to vary with conversation type.

We also plan to apply this model to the analysis of a much larger data set. We have recently collected over 4,400 hours of data from 24 subjects over the course of 9 months. The subjects wore our data collection equipment and all feature extraction was done in real-time on the PDA. (The data also contains features beyond those listed in this paper, but none from which intelligible speech can be reconstructed.) We hope to use our model (along with other techniques) to study the evolution of the subjects' conversational styles and social network. To the best of our knowledge, this is the most detailed study of face-to-face social interactions ever done using automatically gathered data. We would not have been able to collect this data without the privacy guarantees provided by our feature set.

7 Conclusion

We have presented a DBN and privacy-sensitive feature set that are capable of inferring who was speaking when in a conversation. The feature set does not include any information that could be used to reconstruct intelligible speech. In clean data, its performance is comparable to that of systems that use much richer features from which the original speech can be reconstructed. Even though other features useful for speech synthesis and recognition could have been used, we believe there is a huge advantage to protecting privacy. In the long run, this will allow us to collect and model more interesting and spontaneous conversations and extend this work to capture richer conversation dynamics and handle varying numbers of speakers and overlaps.

Acknowledgements

This work was supported by National Science Foundation grant IIS-0433637.

References

[Ajmera *et al.*, 2004] J. Ajmera, G. Lathoud, and L. McCowan. Clustering and segmenting speakers and their locations in meetings. In *Proceedings of ICASSP*, 2004.

[Ang, 2002] J. Ang. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of ICSLP*, 2002.

[Basu, 2002] S. Basu. *Conversational Scene Analysis*. PhD Thesis, MIT, 2002.

[Batliner *et al.*, 2000] A. Batliner, K. Fisher, R. Huber, J. Spilker, and E. Noth. Desperately seeking emotions or: Actors, wizards and human beings. In *Proceedings of the ISCA ITRW on Speech and Emotion*, 2000.

[Baym *et al.*, 2004] 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*, 6:299–318, 2004.

[Bernard *et al.*, 1979] H. Bernard, P. Killworth, and L. Sailer. Informant accuracy in social network data iv. *Social Networks*, 2:191–218, 1979.

[Bilmes and Zweig, 2002] J. Bilmes and G. Zweig. The graphical models toolkit: An open source software system for speech and time-series processing. In *Proceedings of ICASSP*, 2002.

[Bilmes, 2000] J. A. Bilmes. Dynamic bayesian multinets. In *Proceedings of UAI*, 2000.

[Campbell, 2000] N. Campbell. Databases of emotional speech. In *Proceedings of the ISCA ITRW on Speech and Emotion*, 2000.

[Choudhury and Basu, 2004] T. Choudhury and S. Basu. Modeling conversation dynamics as a mixed memory markov process. In *Proceedings of NIPS*, 2004.

[Dean and Kanazawa, 1988] T. Dean and K. Kanazawa. Probabilistic temporal reasoning. In *Proceedings of AAAI*, 1988.

[Dielmann and Renals, 2004] A. Dielmann and S. Renals. Multi-stream segmentation of meetings. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2004.

[Donovan, 1996] R. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University, 1996.

[Douglas-Cowie *et al.*, 2000] E. Douglas-Cowie, R. Cowie, and M. Schroeder. A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA ITRW on Speech and Emotion*, 2000.

[Douglas-Cowie *et al.*, 2003a] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1–2):33–60, 2003.

[Douglas-Cowie *et al.*, 2003b] E. Douglas-Cowie, R. Cowie, and N. Campbell. Speech and emotion. *Speech Communication*, 40(1–2):1–3, 2003.

[Greasley *et al.*, 1995] 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 *Proceedings of the XIIIth ICPhS*, 1995.

[Hawkins, 1991] K. Hawkins. Some consequences of deep interruption in task-oriented communication. *Journal of Language and Social Psychology*, 10:185–203, 1991.

[Hurlburt *et al.*, 2002] R. T. Hurlburt, M. Koch, and C. L. Heavey. Descriptive experience sampling demonstrates the connection of thinking to externally observable behavior. *Cognitive Therapy and Research*, 26(1):117–134, 2002.

[Kossinet and Watts, 2006] G. Kossinet and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

[Marsden, 1990] P. V. Marsden. Network data and measurement. *Annual Review of Sociology*, 16:435–463, 1990.

[McCallum *et al.*, 2005] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In *Proceedings of IJCAI*, 2005.

[McCowan *et al.*, 2003] I. McCowan, S. Bengio, D. Gatica-Perez, F. Lathoud, D. Moore, P. Wellner, and H. Bourlard. Modeling human interactions in meetings. In *Proceedings of ICASSP*, 2003.

[NIST, 2006] NIST. NIST rich transcription evaluations - http://www.nist.gov/speech/tests/rt/rt2006/spring/, 2006.

[Ochs *et al.*, 1996] E. Ochs, E. A. Schegloff, and S. A. Thompson, editors. *Interaction and Grammar*. Cambridge, 1996.

[Quatieri, 2001] T. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 1st edition edition, 2001.

[Reynolds and Torres-Carrasquillo, 2005] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proceedings of ICASSP*, 2005.

[Sacks, 1992] H. Sacks. *Lectures on Conversation*. Blackwell, 1992.

[Schuller *et al.*, 2004] 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 *Proceedings of ICASSP*, 2004.

[Shriberg, 2005] E. Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *Proceedings of Eurospeech*, 2005.

[Tager-Flusberg, 1994] H. Tager-Flusberg. Dissociations in form and function in the acquisition of language by autistic children. In H. Tager-Flusberg, editor, *Constraints on language acquisition: Studies of atypical children*, pages 175–194. Lawrence Earlbaum, 1994.

[Tannen, 1993] D. Tannen. Interpreting interruption in conversation. In *Gender and Discourse*, pages 53–83. Oxford, 1993.

[Welbourne *et al.*, 2005] E. Welbourne, J. Lester, A. LaMarca, and G. Borriello. Mobile context inference using low-cost sensors. In *Proceedings of LoCA*, 2005.

[Wing and Gould, 1979] L. Wing and J. Gould. Severe impairment of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, 9:11–29, 1979.

[Young *et al.*, 1978] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer. A rating scale for mania: Reliability, validity and sensitivity. *British Journal of Psychiatry*, 133:429–435, 1978.