Conversation Detection and Speaker Segmentation in Privacy-Sensitive Situated Speech Data

Danny Wyatt[1]*, Tanzeem Choudhury*[2]*, Jeff Bilmes*[3]

[1]Department of Computer Science and Engineering, University of Washington, U.S.A.
[2]Intel Research, Seattle, U.S.A.
[3]Department of Electrical Engineering, University of Washington, U.S.A.
danny@cs.washington.edu, tanzeem.choudhury@intel.com, bilmes@ee.washington.edu

Abstract

We present privacy-sensitive methods for (1) automatically finding multi-person conversations in spontaneous, situated speech data and (2) segmenting those conversations into speaker turns. The methods protect privacy through a feature set that is rich enough to capture conversational styles and dynamics, but not sufficient for reconstructing intelligible speech. Experimental results show that the conversation finding method outperforms earlier approaches and that the speaker segmentation method is a significant improvement to the only other known privacy-sensitive method for speaker segmentation.

Index Terms: conversation modeling, speaker diarization, privacy, context-aware computing

1. Introduction

Existing spontaneous speech processing efforts have considered settings—meetings, phone conversations, interviews [1, 2, 3]—where the content of the speech is unpredictable, but the decision to have a conversation is made in advance. In these scenarios the dialogue is spontaneous, but the existence of the conversation is not. Many important interactions are unplanned: a chance meeting in an elevator or hallway, a random visit to a colleague's office, or an impromptu trip to a coffee shop. There are currently very few projects that consider such *situated* spontaneous speech: speech recorded "in the wild" as people go about their lives and their conversations occur spontaneously.

Portable devices capable of such recordings have grown in storage capacity while becoming smaller, cheaper, and more powerful. But obstacles to gathering situated spontaneous speech still remain, and perhaps no other obstacle is as prominent as privacy.

Collecting situated speech requires recording people in unconstrained and unpredictable situations, both public and private. There is little control over who or what may be recorded: we cannot just carry a microphone into any café, office, or restroom. 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 requires protecting the privacy of those involved by not always storing complete audio.

In this paper we present a method for discovering and segmenting multi-person conversations in privacy-sensitive audio data. We hope to employ this method to model the social network of a group of 24 people from over 4400 hours of privacy-sensitive, situated audio data that we have collected from them [4]. Beyond social network analysis, our techniques can be employed in any application that needs to know the social context of its users while respecting their privacy by not recording everything they say. We believe research in meeting understanding, medical assessment (e.g. depression, autism), and context-aware computing would greatly benefit from increased access to spontaneous speech data even if the raw audio is not available.

The key contributions of this paper are: (1) A method to automatically discover multi-person conversations. (2) A privacy-sensitive speaker segmentation model that outperforms previous privacy-sensitive techniques. And (3) the evaluation of these methods on truly situated speech in different settings and with different numbers of speakers. Additionally, our method is lightweight, fast, and easily decomposable into distributed components that could be deployed on a network of wearable devices that infer their wearers' social behavior in real-time.

This work was supported by NSF grant IIS-0433637.

2. Privacy-Sensitive Situated Speech Data

Data in our corpus is gathered using a wearable multi-sensor platform containing a microphone which is connected to a PDA that extracts a set of privacy-sensitive features in real-time [4]. Streams of these features are available from individuals whose conversations are to be modeled. The assumption that all involved participants are wearing microphones allows for more restrictions on the data collected and greater privacy assurances for both participants and uninvolved 3rd parties. Additionally, wearable devices enable the collection of situated speech in any location, not just specially instrumented rooms. The data considered in this paper is identical to the larger corpus mentioned above, but with the raw audio also saved in order to establish ground truth.

2.1. Privacy-Preserving Features

To protect the privacy of anyone within the range of the microphone we must ensure that the acoustic information that is saved cannot be used to reconstruct intelligible speech. At the same time we must preserve enough information to infer (1) when conversations occur and (2) who was speaking when and how (e.g. pitch, energy, rate).

To ensure that we cannot reconstruct verbal content, the features we record do not preserve formant information. Our approach for inferring when and how a person is speaking, and which person is speaking (if the speaker is wearing a device), is based on detecting regions of audio that contain voiced speech. Since situated speech data can be recorded in widely varying noise situations, it is important that our features are robust to noise. 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 [5].

We use 33.33ms frames with overlaps of 16.67ms, which is the same as in [5]. In this paper we use two different frame rates, so let us call these 60 Hz frames *voicing frames*. For each voicing frame, the relative entropy (Kullback-Leibler divergence) is computed between the normalized power spectrum of the current voicing frame and a normalized running average of the power spectra of the last 500 voicing frames ($\approx$8.33s). In

addition to the above three features, we also save each frame's energy.

3. Finding Conversations

Let us assume that we have K separate streams of audio from K different people. The goal of the conversation discovery step is to determine, for a window of time, a partitioning of the K streams into groups such that all of the people within a group are engaged in conversation with each other, and not with people in other groups.

To the best of our knowledge, there are only three existing proposals for finding conversations in streams of audio. [6] uses an HMM whose states correspond to every possible partitioning of the speakers and whose observations are binary speaker activity indicators (e.g. whether person A is speaking). The method in [7] computes normalized cross-correlation between raw audio signals and concludes that two people are in a conversation if their correlation coefficients are above a threshold estimated from labeled data. Similarly, the method in [8] computes the mutual information (MI) between binary signals that represent voiced/unvoiced speech and places two people in a conversation if their MI is above a given threshold. Our method extends that of [8] to (1) handle multi-person (beyond pairwise) conversations (2) operate at a finer time granularity, and (3) learn the MI threshold in an unsupervised manner.

We treat conversation finding as a clustering problem where the goal is to cluster windows of audio streams together if the individuals who recorded those streams were in a conversation during the window. Our method is as follows.

First, for each audio stream, we infer the posterior distribution of voiced speech in each voicing frame. This inference is done with an HMM whose observations are the three voicing features described above and whose hidden state is a binary variable indicating whether or not the frame contains voiced speech. Each observation probability is modeled with a single three-dimensional, full covariance Gaussian. The parameters of the voicing HMM are learned from data that does not contain any of the speakers in our evaluation data or in the larger corpus. This voicing HMM has been shown to be speaker-independent and robust across different environmental conditions [5]. For each stream, we use the forward-backward algorithm to infer the voicing posteriors given the entire recorded stream.

Once the voicing posteriors are computed, we aggregate W voicing frames into longer non-overlapping conversation windows. To determine whether two people are in conversation together, we examine the MI between simultaneous conversation windows from each of their streams. The MI between streams A and B for conversation window w is:

$$I(A_w; B_w) = \sum_{v,v'} P(A_w = v, B_w = v') \log \frac{P(A_w=v, B_w=v')}{P(A_w=v)P(B_w=v')}$$

Letting $P_A(V_t = v)$ be the voicing posterior from stream A at voicing frame t, the voicing distributions for conversation window w (beginning at voicing frame τ) are estimated as:

$$P(A_w = v, B_w = v') \triangleq \frac{1}{W} \sum_{t=\tau}^{\tau+W} P_A(V_t = v) P_B(V_t = v')$$

$$P(A_w = v) \triangleq \frac{1}{W} \sum_{t=\tau}^{\tau+W} P_A(V_t = v)$$

That is, we estimate the conversation window voicing distributions using "soft" counts from the voicing posteriors instead of hard counts from actual observations so that the uncertainty in the voicing inference can be carried through to the conversation inference.

While there are many methods for computing the similarity between two signals, MI between voicing streams is well suited for finding conversations between people wearing microphones. At physical distances that are normal for face-to-face conversations, all the microphones are likely to pick up the speech of any speaker in the conversation. It is extremely unlikely that two microphones that are not close enough for their wearers to be in a conversation will observe the same speech signal. Note that this is precisely the opposite of the justification presented in [8] which assumed voiced signals to be complementary. Other metrics (e.g. correlation between energy, which we consider below) do not have this property.

To enforce some temporal smoothing we do not use the MI from a single window alone, but rather a similarity metric that uses MI from multiple, neighboring conversation windows. The similarity metric is defined as $D(A_w, B_w) = \sum_{\tau=w-n}^{w+n} \alpha_\tau I(A_w; B_w)$ where α is a triangular window of length $2n+1$ and $\sum_\tau \alpha_\tau = 1$. We experimented with various conversation window sizes, both overlapping and non-overlapping, and achieved the best results for a window size of 20s (W=1200 voicing frames) with n=1.

Given the similarity metrics between all pairs of streams, we use agglomerative clustering to group the streams into conversations. Agglomerative clustering is fast and does not require advance knowledge of the number of clusters in the data—a useful property since we do not know how many conversations are occurring at a given time. Our application of agglomerative clustering requires a similarity threshold below which it should stop merging clusters. Unlike the earlier methods which learn thresholds from labeled data, we instead take advantage of the nature of our similarity metric to learn a threshold in an unsupervised way. The histogram of D across all conversation frames and pairs is distinctly bimodal. One mode corresponds to the frames from pairs that are not in conversation and for which $D \approx 0$. The other mode corresponds to the frames from pairs that are in conversation for which $D > 0$. Thus, we can cluster the values of D into two groups using k-means (with k=2) and use the midpoint between the two cluster means as the threshold for the agglomerative conversation clustering.

4. Speaker Segmentation

Once conversations have been segmented, we want to infer who was speaking when in each conversation. This is a task known as speaker diarization and there are a number of existing methods for it [9, 10, 11]. However, all of these methods use features (primarily MFCCs) from which the linguistic content of the signal can be easily reconstructed, i.e., they do not meet our privacy requirements. In [12] we presented a technique for speaker segmentation that uses only privacy-sensitive features. In this section we present a refinement of our previous technique that achieves better empirical results on the same privacy-sensitive feature set while using a simpler, faster model.

4.1. Pairwise Speaker Segmentation

Similar to our approach to conversation detection, our speaker segmentation method begins with pairwise comparisons. For a pair of speakers A and B, we aggregate the voicing frames into longer *speaking frames*. The longer speaking frames reduce the sensitivity of the speaker segmentation algorithm to small errors in the voicing inference. We use a speaking frame size of 0.26s (T=16 voicing frames) with an overlap of 0.13s. The NIST standard for evaluating speaker segmentation [13] allows for 0.25s of forgiveness around speaker turn transitions, so we are operating at the maximum conventional granularity.

Let w_s be a binary random variable that indicates whether or not a person speaks during speaking frame s. We define the probability that a person is speaking during frame s in stream A as:

$$P_A(w_s) \triangleq 1 - e^{-\lambda v_s}$$

where $v_s = \frac{1}{T} \sum_{t=\tau}^{\tau+T} P_A(v_t)$ is the proportion of voiced

frames in speaking frame s. Intuitively, this model assumes that the probability that no one is speaking decreases exponentially with the number of voiced frames observed. For each speaking frame s in stream A, we also compute the mean energy (g_s^A) of its constituent voicing frames.

For these speaking frames, we instantiate a new HMM whose hidden state S has four values: (1) no one is speaking, (2) A is speaking, (3) B is speaking, or (4) someone other than A or B is speaking. Call these states n, a, b, and u. The observations for this speaker HMM are the log ratios of the speaker frame energies: $r_s = \log g_s^A - \log g_s^B$. The speaker HMM observation probabilities, $P_o(r_s|S_s)$, are modeled as a one-dimensional Gaussian distribution. The mean of the Gaussian observation probability for states n and u is set to 0. The mean for states a and b is learned from 3 minutes of data collected in a location and from a set of speakers that are different from those in our evaluation data. A single mean $\hat{g}$ is estimated for all pairs of speakers, and states a and b have their means set to $\hat{g}$ and $-\hat{g}$. The variances of the Gaussians for all the four states (identical for a and b) are also estimated from this training data, as is the speech probability parameter λ.

Generally, $r_s > 0$ when $S = a$, $r_s < 0$ when $S = b$, and $r_s \approx 0$ when $S = n$ or $S = u$. To disambiguate between n and u, the speaker HMM also incorporates the posteriors from the voicing HMMs as virtual evidence [14] which introduces a pseudo-observation vector X whose value is always observed to 1, i.e. $\forall s\ x_s = 1$. This entails adding additional observation probabilities based on the speech probabilities:
$P(x_s = 1|S_s = a) \triangleq P_A(w_s)$,
$P(x_s = 1|S_s = b) \triangleq P_B(w_s)$,
$P(x_s = 1|S_s = n) \triangleq 1 - \frac{1}{2}(P_B(w_s) + P_A(w_s))$, and
$P(x_s = 1|S_s = u) \triangleq \frac{1}{2}(P_B(w_s) + P_A(w_s))$.

The transition probabilities are set to intuitive initial values which are refined for each conversation using expectation-maximization (EM). Once the EM step converges, we infer the posterior distribution for each speaker frame using the forward-backward algorithm.

Learning the transition probabilities from the separate training set reduced overall accuracy, as did learning the observation probabilities using EM. This suggests that speaker transitions vary for different pairs of people in different conversations, but that energy ratios are somewhat the same.

4.2. Combining Pairwise Segmentations

Once a posterior distribution over speaker states has been inferred for all pairs, these pairwise distributions are combined into a single, global distribution for the entire conversation. This is done by first expanding each pairwise distribution into a larger distribution with more than four states. This expanded distribution has a state for each speaker who has been placed in the conversation (by the clustering step), as well as a state for no speaker and a state for any other unmiked speakers. The probability assigned to state u for the pair is divided evenly among the other speakers' states (i.e. all but A and B) and the unmiked speaker state.

The expanded distributions from each pair are then combined to form the global distribution. We evaluated two simple methods of combining the distributions: summing $P(S_s = y) = \frac{1}{Z}\sum_k P_k(S_s = y)$ and multiplying $P(S_s = y) = \frac{1}{Z}\prod_k P_k(S_s = y)$, where $P_k(S_s = y)$ is posterior probability computed by pair k and Z is a normalizing term. The summing approach achieved better empirical results.

5. Evaluation

To evaluate our methods we collected 50 minutes of data from 5 people who wore our recording devices while moving around a building and entering and leaving different conversations with one another. The participants were told where to go and whom to speak to, but not what to talk about. They are all friends and had no trouble filling the time with casual conversation. The two primary locations were a quiet meeting room and a loud and noisy atrium (where most of the background noise is other speech), but conversations also occurred while the participants walked together and rode elevators between locations.

mics	accuracy	precision	recall	partial prec.
5	99.2	95.1	92.9	99.0
4	98.5	96.5	91.5	98.2
3	97.5	97.1	91.5	98.2
2	96.1	98.1	93.3	98.1

(a) Multiperson

mics	accuracy	precision	recall
5	97.2	97.5	96.4
4	96.8	97.7	95.2
3	96.2	97.8	94.0
2	96.1	98.1	93.3

(b) Pairwise

Table 1: Conversation detection results using voicing MI

To test the performance of our methods in the presence of unmiked speakers, we selectively removed streams from the data set and performed inference using only the remaining streams. Results reported for fewer than five microphones are computed over all permutations of that number of microphones.

5.1. Conversation Detection

We measured the accuracy, precision, and recall of conversation detection in two ways: per conversation and per pair. The per conversation measurements consider all possible multi-person conversations that could occur between the given number of speakers. In the per conversation evaluation, a true positive means that all participants in a conversation are grouped correctly. False positives, and true and false negatives are defined similarly. The per pair measurements consider each pair of speakers separately. A pairwise true positive indicates that two people were grouped correctly in a conversation without considering the placement of the other participants. False positives, and true and false negatives are defined similarly. For the multi-person evaluation we also compute the proportion of false positive conversations that are subsets of true conversations. We compute a "partial precision" by reducing the error in the true precision by that proportion.

Table 1 shows our conversation detection results. Overall, they are very promising and are a significant improvement on the earlier technique proposed in [8]. That approach modeled pairwise conversations only and achieved the following performance rates on our data: accuracies ranging from 78.6 to 81.2, precisions from 95.9 to 98.4, and recalls from 53.9 to 61.4. That earlier technique used 60 second conversation windows to avoid false positives, but at a significant reduction in recall. By using shorter frames while still considering the MI of neighboring frames we can achieve higher precision with greatly increased recall. Additionally, we get a small benefit ($0.5\% - 2.0\%$) from using soft counts instead of hard counts from voicing inference when computing the MI scores.

Because we do not perform speaker segmentation before conversation detection, we cannot evaluate the technique from [6] on our data. Nevertheless, the best accuracy reported in [6] is 87.5, which our technique exceeds in all cases.

We also cannot compare our conversation detection technique directly to that from [7] since that technique made use of the entire audio signal and does not protect the speakers' privacy. We can, however, approximate it in a privacy-sensitive way by considering the correlation between energies aggregated into voicing frames. When cross correlation between energies is

mics	accuracy	precision	recall	DER
5	81.2	82.9	94.4	11.1
4	77.9	79.2	93.9	15.9
3	74.8	76.0	93.2	20.4
2	73.1	74.8	92.4	23.5

(a) All data

mics	accuracy	precision	recall	DER
4	82.0	81.7	97.8	16.7
3	79.3	78.5	98.1	18.0
2	74.4	73.5	97.0	23.2

(b) Subset of the data used in [12]

Table 2: Speaker segmentation results

used (instead of MI between voicing inferences) to detect conversations, multi-person accuracies ranged from 92.6 to 98.3, precisions from 86.9 to 89.3, and recalls from 98.9 to 82.5. Accuracy decreases notably, but precision decreases even more severely. Since one goal of this technique is to build a model of a social network, low precision is more detrimental than low recall. For accurate sociological analyses, ties should not be inferred where none exist.

5.2. Speaker Segmentation

To evaluate speaker segmentation, for each speaker frame we choose the most likely state from the combined speaker distributions and compare it to the ground truth. From this comparison, we compute four evaluation metrics: (1) accuracy–the fraction of frames in which the inferred state matches the ground truth state, (2) precision–the fraction of the frames for which the correct speaker is inferred, (3) recall–the fraction of truly spoken frames for which any speaker is inferred (i.e., the accuracy of basic speech-detection), and (4) the diarization error rate (DER)–a standard metric used by NIST [13] to measure the performance of speaker segmentation systems. DER is a relaxed version of error rate that merges pauses shorter than 0.3s long and ignores 0.25s of data around a change in speaker.

Table 2(a) shows speaker segmentation results for all 50 minutes of data. These results are comparable with current speaker diarization results: 18.6 is currently the best reported DER (achieved with non-privacy-sensitive features) for meeting data [13]. Unfortunately the dataset from that evaluation is not readily available, so we could not compare results directly. The results presented in this paper are also an improvement on our earlier technique [12], which was evaluated on a subset of this data containing 6 conversations (separated by hand, not automatically) between four speakers. It had accuracies ranging from 53.4 to 71.8, precisions from 53.8 to 72.5, recalls from 90.2 to 98.9, and DERs from 41.9 to 13.6. Table 2(b) shows the performance of the current speaker segmentation on that same subset.

In [12], we used a dynamic Bayesian network to simultaneously infer voicing and speaker segmentation. That joint model introduced correlations between energy observations and voicing inference which reduced the accuracy of the otherwise energy-independent voicing detector. Separating the voicing inferences from the speaker segmentation step and combining them through virtual evidence removed this correlation and improved the our system's performance.

6. Social Context-Aware Applications

Beyond improving performance, the separation of voicing inference from speaker segmentation also allows our technique to be modified to work in real-time and the computation to be distributed among wearable devices. Even though the results reported here use posteriors computed from 50 minutes of data, we have empirically determined that fixed lag smoothing with a lag of 916ms (55 voicing frames) is enough to yield identical posterior distributions.

If devices could compute their own voicing posteriors, they could share them with one another and infer whether their wearers were in a conversation together. This information could then be used by applications that need to know their users' immediate social context. If the devices share their observed mean energies along with their voicing posteriors, then the pairwise speaker segmentation inferences could be shared between devices (with each device in an n person conversation responsible for $\binom{n}{2}/n$ speaker segmentations, on average). Once the pairwise segmentations have been combined into a global inference (which is simple addition), each device would have a model of the conversation's turn-taking dynamics.

Since all of the information shared between devices in those scenarios is privacy preserving and cannot be used to reconstruct speech it is safe for devices to share it without fear that they are broadcasting their entire acoustic environment beyond where it could normally be heard. Additionally, people—especially uninvolved 3rd parties—may be more receptive to intelligent environments that use audio data if they know that the data being used is not raw audio and that their speech is not being recorded or recognized and transcribed.

7. Conclusion

We have presented a method for automatically finding conversations in privacy preserving data and segmenting those conversations into speaker turns. The method makes use of several small, fast inference procedures and combines them using virtual evidence. The conversation finding technique outperforms the other three known methods for conversation discovery. The speaker segmentation model improves the performance of our earlier system—the only other privacy-sensitive speaker segmentation work that we are aware of. The decomposition of the problem also allows for our method to be extended to work in a distributed network of wearable devices that model their users' social contexts.

8. References

[1] 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.

[2] 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.

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

[4] D. Wyatt, T. Choudhury, and H. Kautz, "Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort," in *Proc. of ICASSP*, 2007.

[5] S. Basu, "A linked-HMM model for voicing and speech detection," in *Proc. of ICASSP*, 2003.

[6] O. Brdiczka, J. Maisonnasse, and P. Reignier, "Automatic detection of interaction groups," in *Proc. of ICMI*, 2005.

[7] S. R. Corman and C. R. Scott, "A synchronous digital signal processing method for detecting face-to-face organizational communication behavior," *Social Networks*, vol. 16, pp. 163–179, 1994.

[8] S. Basu, "Conversational scene analysis," PhD Thesis, MIT, 2002.

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

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

[11] X. Anguera, "Robust speaker diarization for meetings," PhD Thesis, UPC, 2006.

[12] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, "A privacy sensitive approach to modeling multi-person conversations," in *Proc. of IJCAI*, 2007.

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

[14] J. Bilmes, "On soft evidence in bayesian networks," Dept. of EE, U. of Washington, Tech. Rep. UWEETR-2004-0016, 2004.