GINA: Group Gender Identification Using Privacy-Sensitive Audio Data

Jiaxing Shen[1], Oren Lederman[2], Jiannong Cao[1], Florian Berg[2], Shaojie Tang[3], and Alex 'Sandy' Pentland[2]

[1]The Hong Kong Polytechnic University. Email: {csjshen,csjcao}@comp.polyu.edu.hk
[2]Massachusetts Institute of Technology. Email: {orenled,fberg,pentland}@mit.edu
[3]The University of Texas at Dallas. Email: shaojie.tang@utdallas.edu

***Abstract*—Group gender is essential in understanding social interaction and group dynamics. With the increasing privacy concerns of studying face-to-face communication in natural settings, many participants are not open to raw audio recording. Existing voice-based gender identification methods rely on acoustic characteristics caused by physiological differences and phonetic differences. However, these methods might become ineffective with privacy-sensitive audio for two main reasons. First, compared to raw audio, privacy-sensitive audio contains significantly fewer acoustic features. Moreover, natural settings generate various uncertainties in the audio data. In this paper, we make the first attempt to identify group gender using privacy-sensitive audio. Instead of extracting acoustic features from privacy-sensitive audio, we focus on conversational features including turn-taking behaviors and interruption patterns. However, conversational behaviors are unstable in gender identification as human behaviors are affected by many factors like emotion and environment. We utilize ensemble feature selection and a two-stage classification to improve the effectiveness and robustness of our approach. Ensemble feature selection could reduce the risk of choosing an unstable subset of features by aggregating the outputs of multiple feature selectors. In the first stage, we infer the gender composition (mixed-gender or same-gender) of a group which is used as an additional input feature for identifying group gender in the second stage. The estimated gender composition significantly improves the performance as it could partially account for the dynamics in conversational behaviors. According to the experimental evaluation of 100 people in 273 meetings, the proposed method outperforms baseline approaches and achieves an F1-score of 0.77 using linear SVM.**

***Index Terms*—gender detection, group gender identification, nonlinguistic audio analysis**

I. Introduction

Group gender plays an essential role in understanding social interaction and group dynamics [1], [2]. It is also the foundation of promising research like gender inequality [3] and gender difference [4]. With the prevalence of studying spontaneous face-to-face communication in natural settings [5]–[7], it becomes unprecedentedly important to identify group gender through privacy-sensitive audio data. Because face-to-face conversation is a dominant and the richest communication modality available to humans [8], [9]. Such communication could capture real emotions and represent true information flow within an organization [10], [11].

Gender identification using privacy-sensitive data is based on ethical and practical needs. Collecting truly spontaneous conversation requires recording people in unconstrained and unpredictable situations, both public and private. There is little control over who or what might be recorded. Private content and uninvolved parties could be recorded without their consent - a scenario that, if raw audio is involved, is always unethical and sometimes illegal. Therefore, assuming access to raw audio is impractical for most real-world situations and impedes collecting truly natural data [10]. An alternative is to collect privacy-sensitive audio [11]. The microphone signal is sampled at 700 Hz and generates an average amplitude reading every 50 milliseconds to ensure raw audio is not recorded nor can it be reconstructed.

Existing voice-based gender identification methods rely on distinctive acoustic characteristics caused by physiological differences (like glottis, vocal tract thickness) and phonetic differences [12], [13]. Those features are extracted from raw audio. Various identification systems have been proposed with different acoustic features and classification models [13]–[17]. The most frequently used features are pitch [14] and first formant [15] which are related to voice sources and vocal tract, respectively.

Despite extensive efforts on voice-based methods, existing solutions might become ineffective with privacy-sensitive audio for two main reasons. First, compared to raw audio, privacy-sensitive audio is too coarse-grained and it is extremely hard to extract valuable acoustic features from it. Moreover, due to natural settings, privacy-sensitive audio contains various uncertainties like background noises. These uncertainties pose serious challenges for existing methods. For example, estimating fundamental frequency with different levels of noises is difficult [13].

In this paper, we aim to achieve group gender identification using privacy-sensitive audio (GINA). Instead of extracting acoustic features from privacy-sensitive audio, we focus on conversational behaviors. The rationale is that conversational behaviors could reflect gender difference. Many sociology studies have reported explicit relationships between gender and conversational behaviors including turn-taking behaviors and interruption patterns [18]–[21]. Take the length of speaking turns as an example, women have shorter speaking turns [22]. Also, men are more likely to interrupt women than the opposite [23]. Different from previous studies whose data are collected in laboratories, we conduct extensive experiments using data collected in natural settings and observe similar patterns. For

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00061

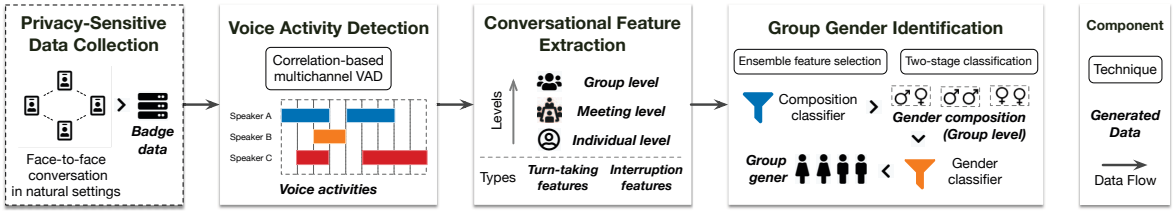


Figure 1: Overview of GINA.

example, we find that the average turn length of women (2.6 seconds) is shorter than that of men (3.2 seconds). Besides, contrary to most existing findings on interruption, we find that women interrupt men more often than vice versa.

The vision of GINA, however, entails two significant challenges when applied to real conditions. *1) Transforming privacy-sensitive audio into voice activities encounters problems including low-resolution audio and unexpected dynamics of spontaneous conversation.* On one hand, the low-resolution audio hinders extracting acoustic features. This makes existing approaches, like multi-class classification, ineffective. On the other, spontaneous conversation in natural settings contains various uncertainties. For example, unpredictable noise and people movement could affect the robustness of existing methods. *2) Although conversational behaviors reflect gender difference to some extent, their instability reduces the robustness and effectiveness of gender identification.* People's conversational behaviors are affected by many factors including internal factors (like emotions) and external factors (like gender composition of the meeting [24], [25]). For example, people behave differently when in mixed-gender and same-gender groups [24]. This results in unstableness and even inconsistency of conversational behaviors and thus affects the performance of gender identification.

To address the first challenge, we propose a correlation-based multichannel voice activity detection (VAD) algorithm. Traditional approaches try to separate voice signals from other people (crosstalk) because crosstalk imposes negative effects on voice applications. However, we observe that crosstalk is beneficial as it generates correlation in privacy-sensitive audio. Based on the observation, we could identify moments when only one person speaks. Then we extract their speaking features to detect voice activities adaptively. For the second challenge, we have made two efforts. To reduce the variance of the performance, we adopt ensemble feature selection which reduces the variance of F-score by over 10%. It is often reported that several different feature subsets may yield equally optimal results, and ensemble feature selection may reduce the risk of choosing an unstable subset [26], [27]. To improve the general identification performance, we propose a two-stage classification method. In the first stage, we predict one of the external factors (gender composition) as an additional input feature for gender identification in the second stage. This approach could improve F-score by over 10% because gender composition could partially explain the dynamics of conversational behaviors.

According to our experimental evaluation of 100 people in 273 meetings, with a total length of 438 hours, GINA improves the performance of baseline approaches by 8.5% on average. GINA could achieve an F1-score of 0.77 using linear SVM. The contribution of this paper is summarized as follows.

- We propose a privacy-sensitive modality (conversational behaviors) for gender identification. The performance is improved by ensemble feature selection and a two-stage classification method.
- An adaptive correlation-based multichannel VAD algorithm for privacy-sensitive audio is proposed.
- We analyze group conversation in natural settings and bring new insights of gender difference in interruption.

The remainder of this paper is organized as follows. An overview is introduced in Section II. We elaborate on design details of the proposed system in Section III. Section IV illustrates the experimental evaluation of the data collected in real-life scenarios. Related works are introduced in Section V, and we conclude this work in the last section.

II. System Overview

In this section, we give an overview of GINA. As illustrated in Figure 1, the proposed system consists of four main components, including Privacy-Sensitive Data Collection, Voice Activity Detection, Conversational Feature Extraction, and Group Gender Identification.

GINA is motivated by the ethical and legal issues arising from studying spontaneous face-to-face conversation. To this end, we exploit electronic badges [11] to collect privacy-sensitive audio data in *Privacy-Sensitive Data Collection*. We briefly introduce this components as it is not our main contribution. More details could be found in [11]. After collecting the badge data, it is processed with the devised multichannel VAD algorithm in *Voice Activity Detection*. This step mainly transforms privacy-sensitive audio data into voice activities or conversational behaviors. In *Conversational Feature Extraction*, we extract two kinds of features, namely turn-taking features and interruption features, for group gender identification. These features could be further divided into individual level, meeting level and group level. We also demonstrate the effectiveness analysis of those features and new insights of gender difference in interruption patterns. Lastly, we introduce the proposed the two-stage classification method in *Group Gender Identification*. It is related to two classifiers: composition classifier and gender classifier. In the composition classifier, we predict the latent information

of gender composition as an additional group level feature. Because people's conversational behaviors vary in groups with different gender composition (mixed gender and same gender). Then we apply ensemble feature selection to three different levels of features to select stable feature subsets. Finally, we exploit the gender classifier to identify group gender based on the selected features.

III. System Design

A. Privacy-Sensitive Data Collection

As indicated in [10], it is a large problem to assume access to raw audio recordings in collecting spontaneous face-to-face conversational data. Therefore, we adopt a platform that uses a privacy-sensitive data collection style [11]. The platform exploits electronic badges [28] which embed multiple sensors like RFID, Bluetooth, and microphone to monitor face-to-face interaction of badge wearers.

The badge samples the microphone signal at 700 Hz and creates an average amplitude reading every 50 milliseconds. The averaged amplitude generally reflects the fluctuation of badger wears' volume. In one second, every badge generates 20 volume data points. We call the timespan of one second as a *frame*. The collected badge data is privacy-sensitive as no raw audio is recorded and the audio cannot be re-generated from the stored samples.

B. Voice Activity Detection

Multichannel voice activity detection (VAD) is to detect whether a user in a channel speaks or not. Privacy-sensitive though the badge data is, it brings new challenges in VAD due to the low resolution of the badge data and unpredictable dynamics of spontaneous conversation in natural settings.

One type of traditional VAD is based on multi-class classification. Related features are extracted from raw audio first and then classification models like Hidden Markov Model [29] or Gaussian Mixture Model [30] are utilized to detect voice activities. However, most of the features could not be extracted from the privacy-sensitive audio data. Besides, it might be difficult to adapt to scenarios without training data.

Another type of methods regards VAD as blind source separation and solves it using independent component analysis (ICA) [31]. However, ICA assumes stationary mixing of the signal, i.e., requires participants to remain fixed at locations. This constraint is hard to satisfy in natural settings as participants would walk around and show some demos during the meeting. Apart from this, it is also difficult to find thresholds to separate speech and noise on the de-mixed signals, which are not resilient to different environments.

Traditional approaches try to separate voice signals from other people (crosstalk) because crosstalk imposes negative effects on voice applications. However, we find that crosstalk is beneficial as it generates correlation in privacy-sensitive audio. When only one badge wearer speaks, other people's badge signals are highly correlated with the speaker's badge signal due to crosstalk. Voice signal from different people could be regarded as independent random variables. Without the effect of crosstalk, the correlation of voice signals from two speakers should obey a zero mean normal distribution. Given a set of participants $\mathbf{P}$ within a meeting, the badge data $\mathbf{S}_i$ of wearer i in a frame could be represented as:

$$\mathbf{S}_i = \underbrace{\mathbf{V}_i}_{\text{Local speech}} + \underbrace{\sum_{j \in \mathbf{P}} \phi_{ij} \cdot \mathbf{V}_j}_{\text{Crosstalk}} + \underbrace{\rho_d + \rho_e}_{\text{Noise}}, j \neq i$$

where $\mathbf{V}_i$ is the voice signal from the wearer in the same frame, ϕ_{ij} is a attenuation factor of voice over the distance between wear i and j, ρ_d and ρ_e are device and environmental noise respectively. The badge signal is a mixture of *local speech* (voice from the badge wearer), *crosstalk* (voice from other participants), and noise (device and environmental noise). When only participant i speaks during frame k, the badge signal of $\mathbf{S}_i(k)$ and $\mathbf{S}_j(k)$ could be reduced to Equation 1. It is clear that approximated $\mathbf{S}_i(k)$ and approximated $\mathbf{S}_j(k)$ are linearly correlated.

$$\begin{cases} \mathbf{S}_i(k) = \mathbf{V}_i(k) + \rho \approx \mathbf{V}_i(k) \\ \mathbf{S}_j(k) = \phi_{ij} \cdot \mathbf{V}_i(k) + \rho \approx \phi_{ij} \cdot \mathbf{V}_i(k) \end{cases} \quad (1)$$

Based on the observation, we propose a correlation-based multichannel VAD algorithm as shown in Algorithm 1[1]. The algorithm takes the badge data $\mathcal{F}_b$ from a whole meeting as input and derives voice activities $\mathcal{F}_r$ for all participants. It

Algorithm 1: Correlation-based multichannel VAD.

Input : $\mathbf{P}$: a set of participants in a meeting;
$\mathcal{F}_b$: a directory of badge data for all participants;
Output: $\mathcal{F}_r$: a directory of voice activities for all participants
Initialize empty directories: $\mathcal{F}_g, \mathcal{F}_a, \mathcal{F}_r$;
$\mathbf{F} \leftarrow \bigcup_l \mathcal{F}_b(l)(frame)$; // $\mathbf{F}$ is a set of all frames in the meeting
/* Step 1: Detect **genuine speak** information */
foreach *frame* $k \in \mathbf{F}$ **do**
  $p \leftarrow argmax\Big(mean\big(\mathbf{S}_i(k)\big)\Big), i, p \in \mathbf{P}$;
  if $\forall j \in \mathbf{P}$, $corr(p, j) > \theta$ **then**
    Add frame k to $\mathcal{F}_g(j)$;
/* Step 2: Detect **all speak** information */
$\mathcal{C} \leftarrow$ *get-clf-rules*$(\mathcal{F}_g)$; // Find classification rule for each person
foreach *frame* $k \in \mathbf{F}$ **do**
  foreach $p \in \mathbf{P}$ **do**
    if $mean(\mathbf{S}_p) \geq \mathcal{C}(p, \text{'mean'})$ ***or*** $std(\mathbf{S}_p) \geq \mathcal{C}(p, \text{'std'})$ **then**
      Add frame k to $\mathcal{F}_a(j)$;
/* Step 3: Detect **real speak** information */
$\mathbf{F}_r(p) = \mathcal{F}_g(j) \cup \mathcal{F}_a(j)$;
foreach *frame* $k \in \bigcup_l \mathcal{F}_a(l)$ **do**
  if $\forall i, j(j \neq i) \in \mathbf{P}, corr\big(\mathbf{S}_i(k), \mathbf{S}_j(k)\big) > \theta$ **then**
    $p \leftarrow argmin\Big(mean\big(\mathbf{S}_i(k)\big), mean\big(\mathbf{S}_j(k)\big)\Big), p \in \mathbf{P}$;
    Remove frame k from $\mathcal{F}_r(p)$;
Function *get-clf-rules* $(\mathcal{F}_g)$;
Input : $\mathcal{F}_g$: A directory of frames when only one person speaks
Output: $\mathcal{C}$ A directory of classification rules for each person
foreach $p \in \mathbf{P}$ **do**
  $\mathbf{D}_t(p, \text{'mean'}) \leftarrow$ distribution of mean volume in a frame when p talks;
  $\mathbf{D}_s(p, \text{'mean'}) \leftarrow$ distribution of mean volume when p remains silent;
  $\mathbf{D}_t(p, \text{'std'}), \mathbf{D}_s(p, \text{'std'}) \leftarrow$ distributions of standard deviation of volume;
  $\mathcal{C}(p, \text{'mean'}) \leftarrow$ intersection of $\mathbf{D}_t(p, \text{'mean'})$ and $\mathbf{D}_s(p, \text{'mean'})$
  $\mathcal{C}(p, \text{'std'}) \leftarrow$ intersection of $\mathbf{D}_t(p, \text{'std'})$ and $\mathbf{D}_s(p, \text{'std'})$
return $\mathcal{C}$

[1]The code: https://github.com/HumanDynamics/openbadge-analysis

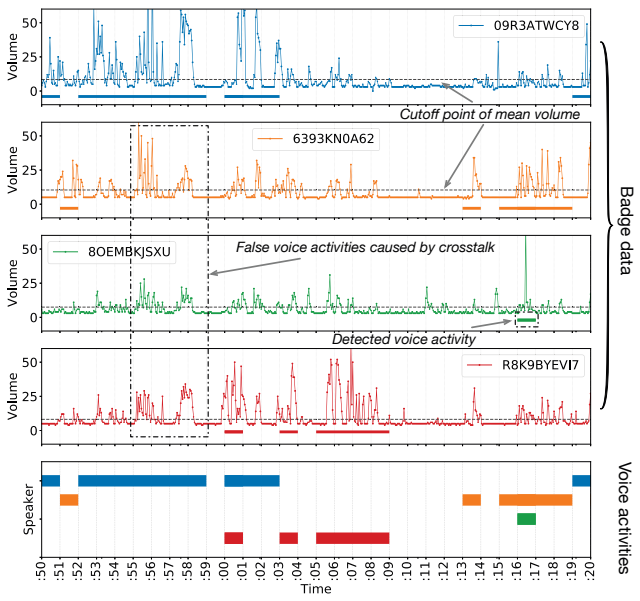


Figure 2: An example result of multichannel VAD on a meeting with four participants between 18:12:50 and 18:13:20.

consists of three main steps. We extract a set of all frames by the union of each participant's frames (line 2). The first step (line 3 $\sim$ 6) is to find a subset of frames $\mathcal{F}_g$ that only one wearer speaks or only one local speech exists (denote as *genuine speak* information). The selection criteria are two-fold. First, the person p must has the highest mean volume to make sure his badge signal is not caused by crosstalk. Second, other people's badge signal are all highly correlated with the person p which ensures p is the only speaker. Parameter θ is a threshold of correlation to detect crosstalk. We further discuss this parameter in Experimental Evaluation.

The second step (line 7 $\sim$ 11) detects all frames that a person is likely to speak by applying classification rules learned from $\mathcal{F}_g$ (*all speak* information). Given $\mathcal{F}_g$, we could identify frames of two situations for a person: talking and silence. Through comparison of both situations, we could identify cutoff points of the statistical features (mean and variance) of the volume.

Since the detected voice activities could be caused by crosstalk, the last step (line 12 $\sim$ 16) is to remove such false activities (*real speak* information). The voice signal of two speakers are expected to be random independent variables, so do their badge data. For pairwise wearers, if their badge signals are strongly correlated (correlation $\geq \theta$), we remove the frame for the wearer who has the weaker volume as it might be caused by crosstalk.

An example result of multichannel VAD is illustrated in Figure 2. The first four sub-figures reveal the badge data collected from four participants. It is clear that participants' badge signals in the box exceed their cutoff point of mean volume. However, these false activities are just caused by crosstalk from the blue participant. The last sub-figure illustrates the

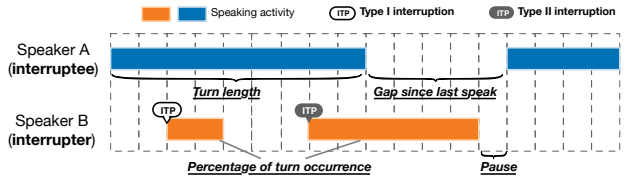


Figure 3: Illustration of conversational features. Underlined bold text represent turn-taking features, the other bold text represent interruption features.

detected voice activities for all participants.

C. Conversational Feature Extraction

After Voice Activity Detection, privacy-sensitive audio data is transformed into voice activities. From the detected voice activities, we define and extract two kinds of conversational features, turn-taking features and interruption features, which are shown in Figure 3.

1) ***Turn-taking features****:* Turn-taking features include turn length (how long a person's turn lasts), the percentage of turn occurrence (how frequently a person speaks), pause between consecutive turns, and gap since last speak as indicated in the literature [18].

Through analysis of the data collected from MIT Sloan Fellows program (See Section IV), we find that some of these features might not be effective. Figure 4(a) $\sim$ (d) depict the probability density functions (PDFs) of four different features. As shown in Figure 4(a), females have shorter turn length than males. According to Figure 4(b), females have larger turn-taking variations. Besides, there seem no significant gender difference in gap since last speak and turn pauses as indicated by Figure 4(c) and (d).

To compare the effectiveness of those turn-taking features in gender identification, we exploit Receiver Operating Characteristic (ROC) curve, which is usually used to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold varies. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. As shown in Figure 4(e), it is clear that the effectiveness of turn length is much better than the others.

2) ***Interruption features****:* According to literature, interruption consists of cooperative and disruptive interruption which could reflect gender difference [23], [32]. Cooperative interruption is usually words of agreement and support or anticipation of how other people' sentences and thoughts would end. Disruptive interruption, on the other hand, is described as having a tendency to switch the topic or take the floor. The detailed description of interruption and gender difference is stated in Related Work (Section V-B).

However, cooperative and disruptive interruption might be too complex and difficult to detect without context information. In Figure 3, we define two roles in interruption. An interrupter is a person who starts his turn before others' turns finish while an interruptee is a person that is interrupted. Besides,

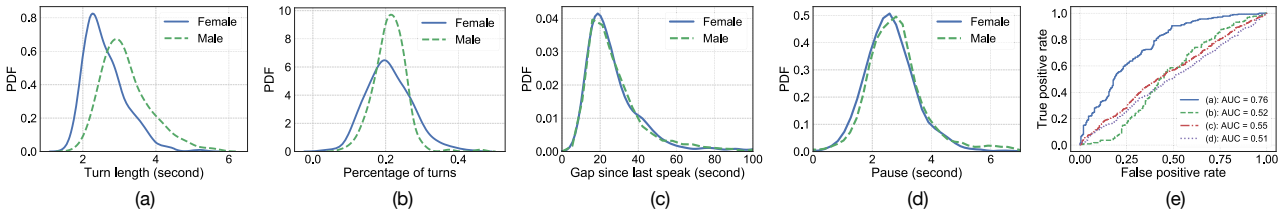


Figure 4: Effectiveness analysis of turn-taking features. (a) $\sim$ (d) PDFs of different features; (e) ROC curves of features.

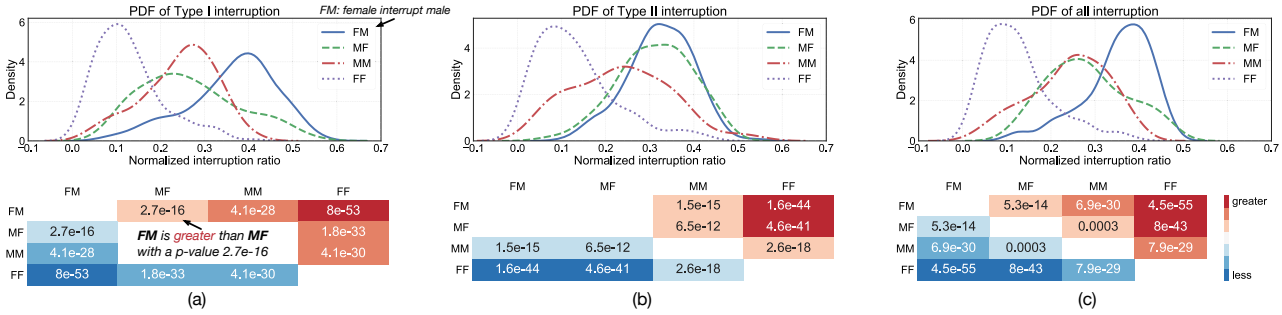


Figure 5: Analysis of who interrupts who with PDFs of four-class interruption and results of Mann-Whitney U test for different types of interruption. (a) Type I interruption; (b) Type II interruption; (c) Type I and Type II interruption.

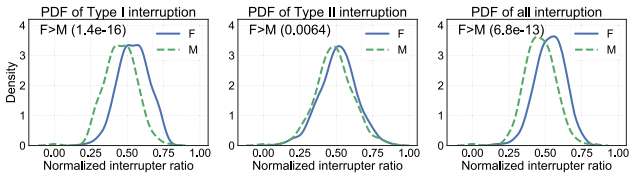


Figure 6: Analysis of inters under three types of interruption.

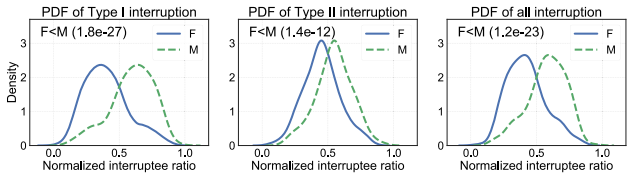


Figure 7: Analysis of intees under three types of interruption.

we also define two types of interruption. Type I interruption is more likely to be a mixture of unsuccessful interruption and cooperative interruption, while Type II interruption is mostly successful interruption.

After analyzing the collected data, we find that generally women interrupt men more frequently which is contrary to the most existing findings in sociology studies [19], [23]. The analysis of interruption consists of three parts, who interrupts who, interrupter, and interruptee.

Who interrupts who: There are four classes of interruption, namely FM (female interrupt male), MF, MM, and FF, in a mixed-gender group meeting. Given the fact that the numbers of both genders are different, we calculate interruption ratios as shown in the matrix.

$$\textbf{Interruption ratios}\quad \begin{array}{cc} \textbf{FF} & \textbf{FM} \\ \textbf{MF} & \textbf{MM} \end{array} = \begin{array}{cc} \frac{\mathbf{I}_{FF}}{\mathbf{I}_F \cdot \mathbf{N}_F} & \frac{\mathbf{I}_{FM}}{\mathbf{I}_F \cdot \mathbf{N}_M} \\ \frac{\mathbf{I}_{MF}}{\mathbf{I}_M \cdot \mathbf{N}_F} & \frac{\mathbf{I}_{MM}}{\mathbf{I}_M \cdot \mathbf{N}_M} \end{array}$$

$\mathbf{I}_{FF}$: Number of FF interruption
$\mathbf{I}_F$: Number interruption started by females
$\mathbf{N}_F$: Number of females in group

The normalized interruption ratio is a normalization of each ratio over their total sum. As shown in Figure 5, we plot PDFs of four classes of interruption in three different situations. To show the relation of pairwise classes of interruption, we resort to Mann-Whitney U test which is a nonparametric test. The null hypothesis of the test is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. We derive interesting results that in different situations, the relations of four-class interruption are also different. For all interruption, the relationship of four-class interruption is FM $>$ MF $>$ MM $>$ FF. For Type I interruption, the relationship mostly holds except there is no significant difference between MF and MM. The PDFs of Type II interruption indicate that there is no significant difference in Type II interruption between female interrupt male and male interrupt female.

Interrupter: The role of gender as interrupters is analyzed in Figure 6. We show PDFs of male and female interrupters under three different types of interruption. The normalized interrupter ratio is simply calculated using the percentage of male or female interrupter over all interrupters. We could find that females are more likely to initiate interruptions especially Type I interruption. This is reasonable since a significant part of Type I interruption is cooperative interruption which is favored by women.

Interruptee: Similar to the analysis of interrupters, we also analyze interruptees. The results in Figure 7 indicate males are far more likely to be interrupted in different types of

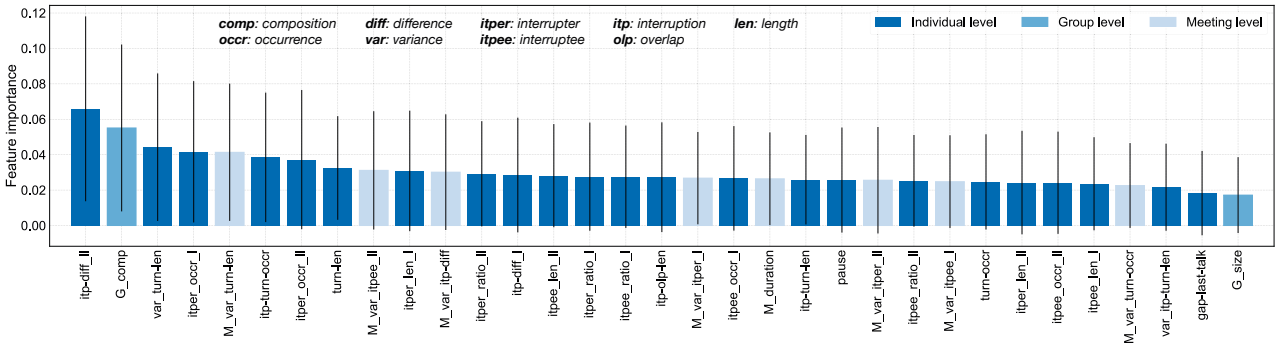


Figure 8: Feature importance of all the features in a Random Forest consisting of 100 trees.

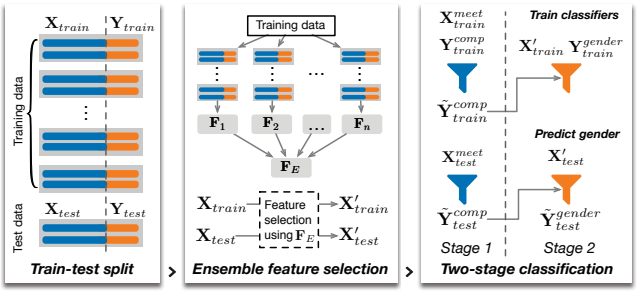


Figure 9: An illustration of ensemble feature selection and the two-stage classification in an iteration of cross-validation.

interruption.

Turn-taking behaviors and interruption patterns could both reflect gender difference. Therefore we devise three levels of features based on turn-taking and interruption. Figure 8 includes different levels of features we use. Features start with an 'M' is a meeting level feature, 'G' indicates group level features, while the rest are individual level features. For example, feature *itper_len_I* means the average length of Type I interruption when a participant acts as an interrupter. Feature *itpee_occr* means the occurrence of interruption when a participant acts as an interruptee. Feature *itp-diff* is the difference between *itper_occr* and *itpee_occr*.

We also show the importance of those features in Figure 8. A Random Forest of 100 trees is used to evaluate their importance on an artificial classification task. Each bar represents the importance of a certain feature, along with its inter-tree variability. We could notice two things. First, it is nontrivial to select a subset of features that are very informative. Second, almost all the features have a large deviation in different trees. This also reflects the instability of conversational behaviors.

D. Group Gender Identification

The last step is to predict group gender based on the extracted features. Specifically, it consists of the following 2 steps: ensemble feature selection and two-stage classification which are illustrated in Figure 9. First of all, in an iteration of k-fold cross-validation, we choose $(k-1)$ folds as training data and the rest fold as test data. The input data ($\mathbf{X}$) consists of three counterparts, individual level feature, meeting level feature, and group level features: $\mathbf{X} = \{\mathbf{X}^{idl}, \mathbf{X}^{meet}, \mathbf{X}^{group}\}$. The label ($\mathbf{Y}$) consists of two parts, composition and gender: $\mathbf{Y} = \{\mathbf{Y}^{comp}, \mathbf{Y}^{gender}\}$. Each fold contains the data from one or more groups. Second, we further separate the training data into n folds for training ensemble feature selector $\mathbf{F}_E$. The selector $\mathbf{F}_E$ is applied to select a subset of features for both training ($\mathbf{X}'_{train}$) and test ($\mathbf{X}'_{test}$) data respectively. Lastly, the training data is used to train two classifiers (composition classifier and gender classifier). During the testing, the estimated composition ($\tilde{\mathbf{Y}}^{comp}_{test}$) and selected input data ($\mathbf{X}'_{test}$) are fed into the gender classifier to infer genders ($\tilde{\mathbf{Y}}^{gender}_{test}$).

*1) **Ensemble feature selection**:* As introduced in Introduction, although conversational behaviors could reflect gender difference, such behaviors are unstable sometimes inconsistent. The potential reasons for those changes rely on the complex nature of human dynamics. Many factors could affect people's conversational behaviors including internal factors like emotions and external factors like gender composition of a meeting [24].

To improve the performance of using conversational behaviors, feature selection is essential. The objectives of feature selection are usually three-fold: improving the prediction performance, providing faster and more cost-effective predictors, and facilitating a better understanding of the underlying process. Furthermore, to handle the instability of conversational behaviors, we adopt ensemble feature selection (EFS). The idea of ensemble feature selection resembles ensemble learning. It is often reported that several different feature subsets may yield equally optimal results in large feature or small sample size domains. EFS could reduce the risk of choosing an unstable subset [33]. Besides, the representational power of a particular feature selector might constrain its search space such that optimal subsets cannot be reached. With EFS this problem could be alleviated by aggregating the outputs of several feature selectors [33].

Among several ways of ensemble, we adopt homogeneous ensemble [27]. It is not only easy to implement, but also more fair to evaluate its effectiveness with the standalone feature

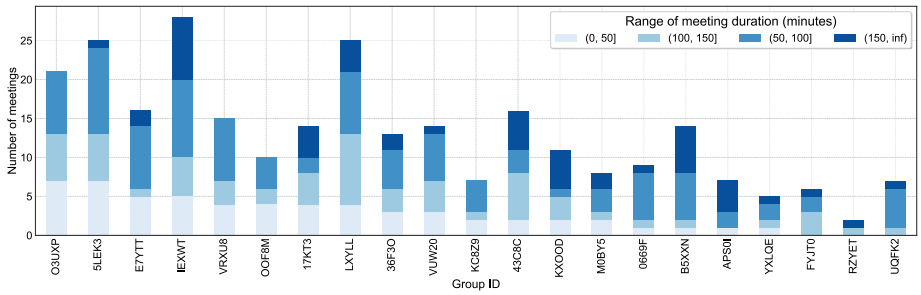


Figure 10: Stacked histogram of number of meetings and meeting duration for all study groups.

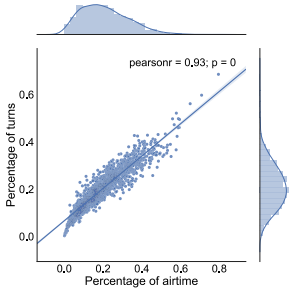


Figure 11: Joint plot of percentage of airtime and percentage of turns.

selector. Homogeneous ensemble applies the same feature selection method to different training data. As illustrated in Figure 9, we separate the training data into n folds and apply n feature selectors of the ensemble $\mathbf{E} = \{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$ to each $(n-1)$ folds of training data. Each selector $\mathbf{F}_i$ outputs a weight vector ($\mathbf{f}_i$) of all features with $\mathbf{f}_i^j$ representing the weight of the j-th feature. To derive a general weight vector $\mathbf{f}_E$ from all weight vectors, we use an average as shown in Equation 2.

$$\mathbf{f}_E^j = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{f}_i^j \tag{2}$$

Lastly, a subset of features is selected with the mean feature weight of $\mathbf{f}_E$ as a threshold.

*2) **Two-stage classification**:* We find gender composition, one of the external factors, could be inferred accurately using meeting level features. Therefore, we propose a two-stage classification method as shown in Figure 9. In the first stage, we infer the latent information of gender composition and treat it as an additional input feature for group gender identification in the second stage. In both stages, we choose popular classification models like linear SVM and Random Forest.

In the first stage, we leverage meeting level features of each group to predict its gender composition. Each participant in the meeting has two roles, interrupting others (as interrupter) and being interrupted by others (as interruptee). The variance of the difference between interrupter and interruptee in a meeting (*M_var_itp-diff*) is a good indicator of gender composition. Same-gendered groups tend to have smaller variance. Because interruption is reported more evenly distributed in same-gendered groups [24]. In the second stage, we combine the selected features and the inferred gender composition as input to predict gender for the whole group.

IV. Experimental Evaluation

A. Settings

*1) **Setup**:* The privacy-sensitive audio data is collected from spontaneous face-to-face meetings of MIT Sloan Fellows class of 2016/17 for about 4 weeks. 100 out of the 110 students participated in the study, including 31 females and 69 males.

Approach	Feature space (in levels)	Feature selection
T-E	Three levels (no composition)	Ensemble feature selection
TC-S	Three levels + composition	Single feature selection
GINA	Three levels + composition	Ensemble feature selection

A *Evaluate group level feature* B *Evaluate ensemble feature selection*

Figure 12: Illustration of baseline approaches.

They came from 35 different countries and had an average age of 37.41 ± 4.45 years (mean standard deviation) as well as an average work experience of 13.78 ± 4.24 years. All participants gave written informed consent about their participation in the study.

Great importance is attached to group collaboration in the MIT Sloan Fellows program. Therefore, Sloan Fellows are assigned to study groups of four or five students before the program starts. The guideline of the group assignment ensures if it is a mixed gender group there are at least 2 students of the same gender. There are 21 study groups including 5 same-gender groups and 15 mixed-gender groups. These groups are consistent over the whole program, and the students within these groups regularly meet to study and work on the courses together. They are free in how often and how long they meet.

*2) **Dataset**:* During the experiment, we collect 273 effective meetings with a total length of 438.25 hours from 21 groups. We show the number of meetings and their duration for each group in Figure 10. On average, each group had 13 meetings, but still, some groups had no more than 5 meetings. Besides, over half of those meetings last for more than 100 minutes.

Through the analysis of detect voice activities, we find that individual's percentage of airtime and percentage of turns are highly correlated with a Pearson correlation over 0.9 as shown in Figure 11. This finding indicates that airtime and turn might be redundant features, use both of them could have a negative impact as specific models are affected in different ways and to varying extents.

B. Evaluation

*1) **Baseline approaches**:* To evaluate the effectiveness of ensemble feature selection and gender composition, we pro-

pose two other approaches as baselines. The detailed configuration of the approaches are illustrated in Figure 12.

Feature selection techniques can be divided into three categories based on how they interact with the classifier. Filter methods directly operate on the dataset by providing a feature weighting, ranking or subset as output. The advantage of being fast and independent of the classification model but at the cost of inferior results. Wrapper methods perform a search in the space of feature subsets, guided by the outcome of the model (like classification performance on cross-validation of the training set). Their results are reported better than filter methods, but at the cost of an increased computational cost. Lastly, embedded methods use internal information of the classification model to perform feature selection (e.g., use of the weight vector in support vector machines). They often provide a good trade-off between performance and computational cost [34]. Therefore, a decision tree based embedded feature selection method is used.

*2) **Evaluation metrics**:* Gender identification is essentially a binary classification problem. We use metrics based on precision, recall, and F1-score to evaluate the performance of the proposed system. When the target label is male (i.e., X is set to male), precision, recall and F1-score for male is calculated as follows.

$$\begin{cases} \text{precision}(p) = \frac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fp}} \\ \text{recall}(r) = \frac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fn}} \\ \text{F1-score} = 2 \cdot \frac{p \cdot r}{p+r} \end{cases}$$

Truth / Prediction	X	$\tilde{X}$
X	**tp**	**fp**
$\tilde{X}$	**fn**	**tn**

X Target label {female, male}; $\tilde{X}$ Non-target label

Consider the imbalance in numbers of females and males, we use a weighted version of those metrics. The weighted F1-score is calculated with Equation 3 where S_F is the support of female or the number of true female instances and F1_F is the F1-score for females. The weighted precision and weighted recall are derived in a similar way.

$$\text{F1} = \frac{S_F}{S_F + S_M} \cdot \text{F1}_F + \frac{S_M}{S_F + S_M} \cdot \text{F1}_M \qquad (3)$$

*3) **Parameter selection**:* Parameter θ in Voice Activity Detection (Section III-B) is a threshold for detecting crosstalk. Different values of θ lead to different genuine speak information ($\mathcal{F}_g$, in Section III-B).

Generally, large θ could derive better accuracy because the frames selected as genuine speak ($\mathcal{F}_g$) becomes more strict. However, it will also lead to large deviation as the number of frames in $\mathcal{F}_g$ decreases. On the contrary, small θ will result in more false detections of genuine speak and thus reduce the accuracy but the number of frames are more than adequate. Given two distributions $\mathbf{D}_t(p, \textit{'mean'})$ (distribution of mean volume when p talks) and $\mathbf{D}_s(p, \textit{'mean'})$ (p remains silent), the larger their distance measured in KL divergence the better. As shown in Figure 13, with the increase of θ, the mean distance also increases with while the number of genuine speak frames decreases. This is a trade-off between accuracy and deviation, we experimentally set $\theta = 0.5$ in our scenario.

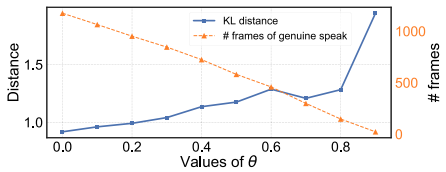


Figure 13: The impact of different θ.

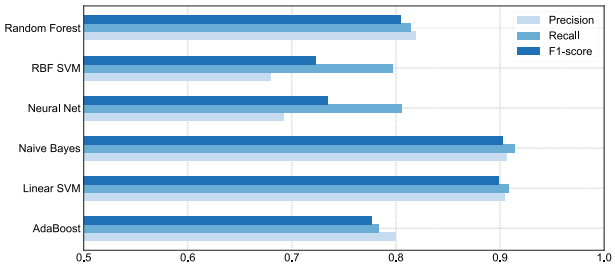
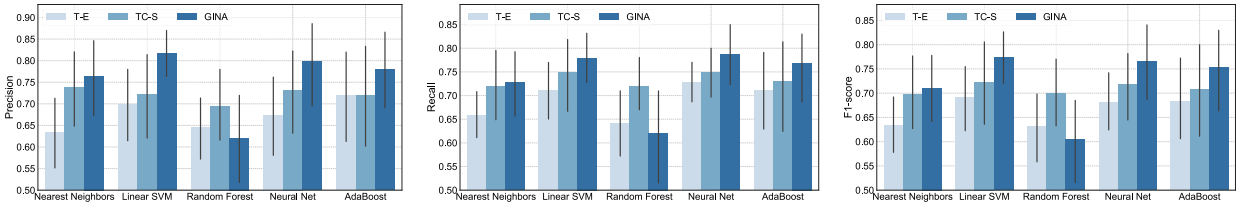


Figure 14: Performance of gender composition detection with different models.

*4) **Performance of gender composition detection**:* We evaluate the performance of gender composition detection with 10-fold cross-validation. Because the number of groups is small, we repeat the cross-validation process for 5 times and show the average performance in Figure 14. Naive bayes and linear SVM outperform other models and achieve a weighted F1-score around 0.9. This indicates the meeting level features we extract have the potential to capture gender composition effectively. Because same-gendered groups and mixed-gendered groups have distinct meeting behaviors. Same-gendered groups have evenly distributed interruption patterns. The gap between a person being an interrupter and an interruptee are close to each other in the same-gendered groups. While in mixed-gendered groups, women tend to have large gap while men are likely to have small gaps. This is reflected in the analysis on *who interrupts who*. Therefore the variance of gaps is larger in the mixed-gendered groups.

*5) **Performance of group gender identification**:* We evaluate the performance of baseline approaches on selected classification models including Nearest Neighbor, Linear SVM, Random Forest, Neural Network and AdaBoost. The parameter settings for all models are consistent with different baselines. As shown in Figure 15, for most of the models, the order of performance is GINA $>$ TC-S $>$ T-E. On average, GINA outperforms T-E and TC-S by 11.62% and 5.37% in F1-score respectively except on Random Forest. This indicates that the inferred gender composition and ensemble feature selection is effective in improving the performance of gender identification.

Not only the performance, but ensemble feature selection could also reduce the variance of performance. Without Random Forest, on average GINA reduces the variance of Precision and Recall by 17.28% and 7.15%. As explained, ensemble feature selection could reduce the risk of choosing



(a) (b) (c)

Figure 15: Comparison of performance using different classification models. (a) Precision; (b) Recall; (c) F1-score.

an unstable subset of features by aggregating the outputs of several feature selectors.

As shown in Figure 8, the feature of gender composition is the second most important feature. On average, this additional feature could improve the Precision and Recall by 15.99% and 9.15%. Gender composition could partially account for the instability of conversational behaviors and thus increase the interpretability of conversational features.

V. Related Work

A. Gender detection

Gender identification has been studied for decades in different areas. Various modalities like vision, online behaviors and voice have been utilized for this purpose. Different application scenarios have varying preferences of modalities. For example, vision-based methods are the first choice in systems where user cooperation is not required, like surveillance systems. In speech recognition, voice-based approaches are preferred.

Vision-based approaches exploit information from the face and whole body (either from a still image or gait sequence) to recognize human gender. It is usually based on appearance differences like face and body, and behavior differences like gait. More details on the utilized techniques and challenging issues could be found in the survey [35].

Vision, voice as well as handwriting are traditional modalities for gender identification. With the development of digital advertising, users' online behaviors like video viewing behaviors [36] and web browsing behaviors [37] are used for gender identification recently. This type of approaches is based on preference differences and behaviors differences.

Among all different modalities, voice is the most related to conversational behaviors. Voice-based methods rely on discriminative features extracted from human voices. The intuition is that different genders have different acoustic characteristics due to physiological differences (like glottis, vocal tract thickness) [13] and phonetic differences [12]. Various identification systems with different classification models and different types of features have been reported in the literature [13]–[17]. The most frequently used features are pitch [14] and first formant [15], which are closely related to voice sources and vocal tract, respectively. Generally, the pitch and the formant frequencies of females are higher than that of males. Moreover, as pointed out in [13], other traditional acoustic features like linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLP), and relative spectral PLP coefficients (RASTA-PLP) are used in the literature for gender identification.

The majority of aforementioned acoustic features depend on accurate estimation of the fundamental frequency which itself is a challenging task. Therefore, Alhussein et. al. propose a new single-value feature in the form of area under the modified voice contour (MVC) in [13]. The proposed feature is independent of fundamental frequency and is proved promising compared to existing features.

Besides, there is a trend of combining multiple features for gender identification in recent work [16], [38]. For example, Abouelenien et. al. extract features from five different modalities, including acoustic, linguistic, visual, thermal, and physiological, to identify gender [16].

B. Gender differences and interruption

The occurrence of overlap and interruption have been found closely related to gender in many sociology studies [19], [23]. The classic study by Zimmerman and West found that in same-sex conversations, interruptions were rare and appeared to be evenly distributed between speakers, whereas in cross-sex conversations, almost all the interruptions were initiated by male speakers [19]. A well-adopted explanation is males tend to show dominance by interrupting females. Many other works have found similarly that men interrupt more than women.

However, a few studies have different findings. For example, Hannah et. al. found no significant difference between interruption and gender [20]. Murray and Covelli even had a contrary discovery that women interrupt more than men [21]. One potential reason for the diverse findings is multiple conceptual and operational definitions of interruptions used in the literature [23]. Interruption is a complex interactional phenomenon with rich meanings, diverse functions, and various structural features [23]. There exist two different types of interruption, cooperative and disruptive, in literature [23], [32]. Cooperative interruption is usually words of agreement and support or anticipation of how other people' sentences and thoughts would end. This type of interruption is reported characteristic of women's style of speech [18] that might have a potentially positive influence on the interpersonal relationship between speakers. Disruptive interruption, on the other hand, is described as tending to switch the topic or take the floor. This type of interruption is attributed to men's

style that might have the potential to bear negatively on the interpersonal relationship between speakers.

VI. Conclusion

In this paper, we propose a data mining system (GINA) to identify group gender through privacy-sensitive audio data. Our contribution are three-fold. First, we propose a privacy-sensitive modality for gender identification. The effectiveness and robustness are improved by ensemble feature selection and a two-stage classification. Second, an adaptive correlation-based multichannel VAD algorithm for privacy-sensitive audio is proposed. Last, we bring new insights of gender difference in interruption through analysis of group conversation in natural settings. According to experimental evaluation, GINA could effectively identify group gender with an F1-score 0.77 using Linear SVM.

Acknowledgment

The work is completed during the visit of the first author to MIT Media Lab. It was partially supported by the funding for Project of Strategic Importance provided by The Hong Kong Polytechnic University (Project Code: 1-ZE26). It was also supported by demonstration project on large data provided by The Hong Kong Polytechnic University (project account code: 9A5V) and NSFC Key Grant with Project No. 61332004.

References

[1] P. S. Tolbert, M. E. Graham, and A. O. Andrews, "Group gender composition and work group relations: Theories, evidence, and issues," 1999.

[2] P. Raghubir and A. Valenzuela, "Malefemale dynamics in groups: A field study of the weakest link," *Small Group Research*, vol. 41, no. 1, pp. 41–70, 2010.

[3] H. L. Ford, C. Brick, K. Blaufuss, and P. S. Dekens, "Gender inequity in speaking opportunities at the american geophysical union fall meeting," *Nature communications*, vol. 9, 2018.

[4] L. Zheng, R. Ning, L. Li, C. Wei, X. Cheng, C. Zhou, and X. Guo, "Gender differences in behavioral and neural responses to unfairness under social pressure," *Scientific reports*, vol. 7, no. 1, p. 13498, 2017.

[5] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] R. Bernstein, "Communication: spontaneous scientists," *Nature*, vol. 505, no. 7481, pp. 121–123, 2014.

[7] L. Ten Bosch, N. Oostdijk, and J. P. De Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *International Conference on Text, Speech and Dialogue*. Springer, 2004, pp. 563–570.

[8] N. K. Baym, Y. B. Zhang, and M.-C. Lin, "Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face," *New Media & Society*, vol. 6, no. 3, pp. 299–318, 2004.

[9] L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, "Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task," 2008.

[10] D. Wyatt, T. Choudhury, and H. Kautz, "Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–213.

[11] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, "Rhythm: A unified measurement platform for human organizations," *IEEE MultiMedia*, vol. 25, no. 1, pp. 26–38, 2018.

[12] A. P. Simpson, "Phonetic differences between male and female speech," *Language and Linguistics Compass*, vol. 3, no. 2, pp. 621–640, 2009.

[13] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, "Automatic gender detection based on characteristics of vocal folds for mobile healthcare system," *Mobile Information Systems*, vol. 2016, 2016.

[14] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.

[15] K. Rakesh, S. Dutta, and K. Shama, "Gender recognition using speech processing techniques in labview," *International Journal of Advances in Engineering & Technology*, vol. 1, no. 2, pp. 51–63, 2011.

[16] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Multimodal gender detection," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 302–311.

[17] M. Kumari and I. Ali, "An efficient algorithm for gender detection using voice samples," in *Communication, Control and Intelligent Systems (CCIS), 2015*. IEEE, 2015, pp. 221–226.

[18] D. Reznik, "Gender in interruptive turns at talk-in-interaction," *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, vol. 4, no. 3, 2004.

[19] D. H. Zimmermann and C. West, "Sex roles, interruptions and silences in conversation," *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pp. 211–236, 1996.

[20] A. Hannah and T. Murachver, "Gender and conversational style as predictors of conversational behavior," *Journal of Language and Social Psychology*, vol. 18, no. 2, pp. 153–174, 1999.

[21] S. O. Murray and L. H. Covelli, "Women and men speaking at the same time," *Journal of Pragmatics*, vol. 12, no. 1, pp. 103–111, 1988.

[22] C. L. Ridgeway, *Gender, interaction, and inequality*. Springer, 1992.

[23] X. Zhao and W. Gantz, "Disruptive and cooperative interruptions in prime-time television fiction: The role of gender, status, and topic," *Journal of Communication*, vol. 53, no. 2, pp. 347–362, 2003.

[24] A. Mulac, "Men's and women's talk in same-gender and mixed-gender dyads: Power or polemic?'," *Journal of Language and Social Psychology*, vol. 8, no. 3-4, pp. 249–270, 1989.

[25] Y. Xu, "Gender differences in mixed-sex conversations: A study of interruptions," 2009.

[26] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 313–325.

[27] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp. 124–139, 2017.

[28] O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. E. Murray, and A. Pentland, "Open badges: A low-cost toolkit for measuring team communication and dynamics," *arXiv preprint arXiv:1710.01842*, 2017.

[29] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 84–91, 2005.

[30] T. Pfau, D. P. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the icsi meeting recorder," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 107–110.

[31] S. Maraboina, D. Kolossa, P. Bora, and R. Orglmeister, "Multi-speaker voice activity detection using ica and beampattern analysis," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5.

[32] D. Tannen and D. Tannen, *You just don't understand*. Simon & Schuster Audio, 1991.

[33] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[34] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[35] C. B. Ng, Y. H. Tay, and B. M. Goi, "Vision-based human gender recognition: A survey," *arXiv preprint arXiv:1204.1611*, 2012.

[36] J. Zhang, K. Du, R. Cheng, Z. Wei, C. Qin, H. You, and S. Hu, "Reliable gender prediction based on users video viewing behavior," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 649–658.

[37] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 151–160.

[38] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceeding of Speech Prosody*, 2016, pp. 84–88.