# User Profiling Based on Nonlinguistic Audio Data

JIAXING SHEN and JIANNONG CAO, The Hong Kong Polytechnic University, China
OREN LEDERMAN, Massachusetts Institute of Technology, USA
SHAOJIE TANG, The University of Texas at Dallas, USA
ALEX "SANDY" PENTLAND, Massachusetts Institute of Technology, USA

User profiling refers to inferring people's **attributes of interest** (**AoIs**) like gender and occupation, which enables various applications ranging from personalized services to collective analyses. Massive nonlinguistic audio data brings a novel opportunity for user profiling due to the prevalence of studying spontaneous face-to-face communication. Nonlinguistic audio is coarse-grained audio data without linguistic content. It is collected due to privacy concerns in private situations like doctor-patient dialogues. The opportunity facilitates optimized organizational management and personalized healthcare, especially for chronic diseases. In this article, we are the first to build a user profiling system to infer gender and personality based on nonlinguistic audio. Instead of linguistic or acoustic features that are unable to extract, we focus on conversational features that could reflect AoIs. We firstly develop an adaptive voice activity detection algorithm that could address individual differences in voice and false-positive voice activities caused by people nearby. Secondly, we propose a gender-assisted multi-task learning method to combat dynamics in human behavior by integrating gender differences and the correlation of personality traits. According to the experimental evaluation of 100 people in 273 meetings, we achieved 0.759 and 0.652 in F1-score for gender identification and personality recognition, respectively.

CCS Concepts: • **Information systems** → **Data mining**; • **Social and professional topics** → **User characteristics**; • **Human-centered computing** → *Ubiquitous and mobile computing;*

Additional Key Words and Phrases: User profiling, nonlinguistic audio, personality recognition, gender identification, multi-task learning

**ACM Reference format:**
Jiaxing Shen, Jiannong Cao, Oren Lederman, Shaojie Tang, and Alex "Sandy" Pentland. 2021. User Profiling Based on Nonlinguistic Audio Data. *ACM Trans. Inf. Syst.* 40, 1, Article 17 (August 2021), 23 pages.
https://doi.org/10.1145/3474826

This work was supported by RGC CRF (C5026-18G) and CRF (C6030-18G). It was also supported by PolyU Internal Start-up Fund (P0035274).

Authors' addresses: J. Shen and J. Cao, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Hong Kong, China; emails: {jiaxshen,jiannong.cao}@polyu.edu.hk; O. Lederman and A. S. Pentland, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, US; emails: {orenled,pentland}@mit.edu; S. Tang, The University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, US; email: shaojie.tang@utdallas.edu.



1046-8188/2021/08-ART17 $15.00
https://doi.org/10.1145/3474826

## 1 INTRODUCTION

User profiling refers to the process of inferring users' **attributes of interest** (**AoIs**) like gender, occupation, and personality. Since AoIs are indispensable in various applications ranging from personalized services [8, 21, 30] to collective analyses [31, 44, 50], user profiling is thus increasingly valued in both academia and industry.

The accumulation of nonlinguistic audio data results from the prevalence of studying spontaneous face-to-face communication in naturalistic environments [6, 19], which brings a novel opportunity for user profiling. Nonlinguistic audio is a low-sampling audio signal processed with a mean filter so that the linguistic content cannot be recognized [44]. Collecting nonlinguistic audio is mainly due to privacy concerns since truly spontaneous conversation happens in unconstrained and unpredictable situations where private content and uninvolved parties could be recorded without consent [20]. If raw audio is involved, it is unethical and sometimes illegal like in doctor-patient dialogues and business meetings.

Effective user profiling with nonlinguistic audio is beneficial to stakeholders in different scenarios. In healthcare, patients could get personalized treatments based on the inferred personalities [15], especially for chronic diseases [47]. While for organization administrators, the estimated AoIs provide additional contextual information for organizational design and management [31]. For example, understanding what kind of person is more likely to influence group productivity could facilitate better administration [54].

Although different data modalities have been studied for user profiling, there is little research on nonlinguistic data, to the best of our knowledge. Existing audio processing methods mainly focus on linguistic and acoustic features extracted from raw audio [2, 5]. The way people choose words (linguistic features) and how they speak (acoustic features) could reflect their AoIs like gender and personality [46]. However, these methods are inapplicable to nonlinguistic audio for two reasons. First, compared to raw audio, nonlinguistic audio is too coarse-grained to extract valuable acoustic or linguistic features. Second, as collected in naturalistic settings, nonlinguistic audio contains various uncertainties, including background noises and unexpected voices. These uncertainties pose serious challenges for existing methods, e.g., estimating the fundamental frequency under different levels of noises [25, 33, 53].

In this article, we propose a user profiling system to infer gender and personality based on nonlinguistic audio data. Instead of acoustic or linguistic features that are unable to extract from nonlinguistic audio data, we focus on conversational features including turn-taking and interruption behaviors. Although existing sociology and psychology studies have qualitative findings on the relationship between conversational behaviors, gender, and personality, there are rarely any quantitative studies. For example, men have longer speaking turns [38] and are more likely to interrupt women than been interrupted by women [57]. Besides, particular turn-taking styles are related to personality. Extroverts, for example, tend to talk more, louder, faster, and have fewer hesitations [5]. Also, men's and women's personalities appear to differ in several aspects. For example, women scored notably higher than men in Neuroticism [18]. Different from previous studies whose data are mainly collected in laboratories, we quantify their correlations using extensive experiments from real study groups in natural settings.

Our vision, however, entails two grand challenges when applied to real conditions. (1) How to accurately detect individual voice activities from nonlinguistic audio. First, variations in people's vocal features and ways of collecting the audio data pose serious challenges to accurate **voice activity detection** (**VAD**). Second, due to physical proximity, the nonlinguistic audio may come from other participants, which leads to false-positive detections. (2) How to fill in the gap between dynamic conversational behaviors and stable AoIs. Both gender and personalities are consistent

over time [49], but conversational behaviors are dynamic and could be affected by many factors like emotions and environments. For example, people behave differently in conversations with different gender compositions [56]. Certian personalities may also have different interpretations under different social contexts [34].

To address the first challenge, we devise an adaptive Bayesian VAD algorithm based on the observation when only one person speaks, his audio signals are highly correlated with others' signals. We first exploit the correlation patterns to identify a fraction of audio data when only one person speaks while the others remain silent. According to the speaking and silent data of an individual, we could learn his vocal features and detect all the voice activities from that individual adaptively. Then we use the correlation again and rectify false-positive detections caused by crosstalk. For the second challenge, we have made the following three efforts. First, we manage to capture the dynamics of conversational behaviors by inferring multi-level features including individual-level, meeting-level, and group-level. Meeting-level features could illustrate intra-group interactions while group-level features could represent contextual factors like social context. Second, we find that whether the group is of the same gender is effective in predicting the gender information of each member. Third, due to gender differences in personalities, we propose a gender-assisted multi-task learning method to predict gender as extra input for personality recognition. Jointly learning the correlated personality traits could reduce the risk of overfitting and lead to better generalization performance.

According to the experimental evaluation of 100 people in 273 meetings, with a total length of 438 hours, the proposed method achieves average F1-scores of 0.759 and 0.652 for gender inference and personality recognition, respectively. Contrary to most existing findings on interruption [57, 58], we find that women interrupt men more often than vice versa. We also observe gender differences in both personality traits and conversational behaviors. For instance, male extroverts mostly meet the literature description that they tend to have more turn occurrence, longer turn duration, and larger variances of turn length [5]. However, female extroverts are observed positively correlated with turn occurrence only.

To summarize, the contributions of this article are threefold.

—We are the first to build a user profiling system from nonlinguistic audio which could effectively infer gender and personality by incorporating multi-level features into the proposed gender-assisted multi-task learning model.
—The proposed Bayesian algorithm is effective in detecting voice activities from nonlinguistic audio.
—We analyzed real group conversations in natural settings and provided evidence of gender difference in conversational behaviors and personality traits.

The rest of this article is organized as follows. In Section 2, the concepts of personality and its assessment are explained. Then we elaborate on the design details of the proposed system in Section 3. Section 4 illustrates the experimental evaluation of the data collected in real-life scenarios. Related works are introduced in Section 5, and we conclude this work in the last section.

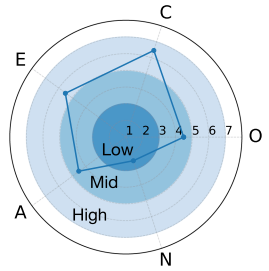
## 2 PERSONALITY AND THE GROUND TRUTH

"Personality is the latent construct that accounts for individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms-hidden or not-behind those patterns" [28].

Personality recognition refers to recognizing the true personality levels of given individuals rather than the personality impressions others attribute to them [49]. Personality recognition consists of two components: personality representation and personality measurement. To represent

| | |
|---|---|
| ***Openness*** | Artistic, Curious, Imaginative, Insightful, Original, Wide interests, etc. |
| ***Conscientiousness*** | Efficient, Organized, Planful, Reliable, Responsible, Thorough, etc. |
| ***Extraversion*** | Active, Assertive, Energetic, Outgoing, Talkative, etc. |
| ***Agreeableness*** | Appreciative, Kind, Generous, Forgiving, Sympathetic, Trusting, etc. |
| ***Neuroticism*** | Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying, etc. |

(a)




(b)

Fig. 1. Five personality traits (OCEAN). (a) Detailed explanations. (b) An example OCEAN data of an individual.

personality, the mostly adopted psychological model is the Five-Factor Model (aka Big Five) which describes five personality traits with five dimensions including **Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism** (**OCEAN**) [49]. To measure personality, there are various ways like surveys, questionnaires, and social media [49]. Audio is especially popular due to its convenient accessibility in different scenarios [45, 55]. To represent personality, numerous models have been proposed in the literature [26]. Despite the wide variety of terms at disposition, personality descriptors tend to group into five major clusters [28]. Therefore, the most commonly adopted personality model in both psychology and computer science is the Big Five [46, 49]. Big Five has five broad dimensions that "appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences" [41]. Figure 1(a) gives detailed explanations of the five traits (OCEAN). Each trait ranges from 1 to 7 and is further separated into three different levels as illustrated in the example of Figure 1(b). For instance, values of 1 and 7 in Extraversion represent extremely introverted and extroverted, respectively.

The main instruments for scoring Big Five are questionnaires where a person is assessed in terms of observable behaviors and characteristics. Several inventories have been developed for measuring the five dimensions, like **Revised NEO Personality Inventory** (**NEO-PI-R**) [10], the **Eysenck Personality Questionnaire** (**EPQ**) [11], and the **Big Five Inventory** (**BFI**) [17]. Accurate as these questionnaires are, they mostly contain dozens of questions. It is extremely difficult to ask a large number of people to fill in long tedious surveys which might discourage them from participating in the event. Under this situation, brief measurements of the Big-Five personality are proposed like **10-item BFI** (**BFI-10**) [36] and the **Ten-Item Personality Inventory** (**TIPI**) [13].

To collect the ground truth, we adopt TIPI as published analysis suggested that the TIPI "achieves slightly better validity than the other measures" after comparing several brief measures [4, 12]. A limitation of TIPI is that it is not adaptable to capturing the finer, narrow-bandwidth personality traits [4]. Therefore, we use three different levels (low, mid, high) to represent each trait as shown in Figure 1(b). One of our focuses in this work is to investigate the possibility of recognizing personality levels with the coarse-grained nonlinguistic audio.

## 3 SYSTEM DESIGN

In this section, we elaborate on the design details of the proposed user profiling system. It comprises three main components as illustrated in Figure 2. In the first component, we focus on what is nonlinguistic audio and how to detect voice activities from them. Secondly, we extract two kinds of conversational features (turn-taking behaviors and interruption patterns) in multiple levels for

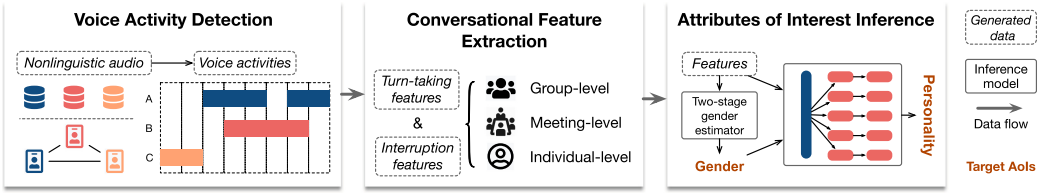



Fig. 2. An overview of the proposed user profiling system.

profiling AoIs. In the last component, we first use a two-stage classification model for gender detection followed by a gender-assisted multi-task learning model for personality recognition.

### 3.1 Voice Activity Detection

As explained, nonlinguistic audio is generated from spontaneous face-to-face conversation in naturalistic environments. Wearable devices like Open badge [19] and "sociometer" [31] are usually used for this purpose. Every participant of a meeting wears a badge powered by a button battery to collect the nonlinguistic audio. The microphone in the badge samples voice signals at 700 Hz and averages amplitude readings every 50 milliseconds. It generates 20 data samples in a *frame* (the time span of 1 second). The averaged amplitude readings generally reflect the fluctuation of voice volumes of badge wears. The advantage of nonlinguistic audio is not only privacy-preserving but also long battery lives which ensure an adequate recording capacity.

**Voice activity detection (VAD)** is to detect whether a participant speaks or not given the nonlinguistic audio of all participants of a meeting. It is not a simple binary problem where any nonzero audio signal could be regarded as voice activities. The main difficulties are twofold as explained in the Introduction. First, different levels of background noises, different ways of wearing the badge, and different natures of people's voices like the level of sound lead to varied forms of input signals. This variation poses a serious challenge to accurate VAD. Second, due to physical proximity, the recorded voice may not only come from the badge wearer himself (*local speech*) but also other nearby participants (*crosstalk*), which results in false-positive detections of voice activities.

Conventional ways [25, 33, 53] attempt to separate an individual's voice signals from others' voices because crosstalk imposes negative impacts on voice applications. More detailed information on traditional VAD methods could be found in Related Works (Section 5). However, according to our analysis, we observe an important phenomenon that when only one badge wearer speaks, his input audio signal is positively correlated with other peoples' badge signals due to crosstalk. This observation could be understood as follows. Given a set of people $\mathbf{P}$ in a meeting, the badge signal $\mathbf{S}_i$ of participant $i$ consists of three parts:

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

where $\mathbf{V}_i$ is the audio signal from participant $i$, $\phi_{ij} \in (0, 1)$ is an attenuation factor of audio signal over the distance between participant $i$ and $j$, and $\mu_d$ and $\mu_e$ are device and environmental noises, respectively. 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), which reveals an obvious linear correlation and the average value of

**ALGORITHM 1:** Bayesian voice activity detection

**Input** : **P**: a set of participants of a meeting
**F**: a set of their nonlinguistic audio frames
**Output**: Voice activities of all participants

/* Step 1: calculate probabilities of different cases */

**foreach** *frame* $k \in \mathbf{F}$ **do**
  $i \leftarrow \arg\max\left(\text{mean}(\mathrm{S}_p^k)\right), p \in \mathbf{P}$ ; // $i$ is the loudest
  // cor(.): Pearson Correlation
  $P_{i,j}^k(C) \leftarrow \text{cor}\left(\mathrm{S}_i^k, \mathrm{S}_j^k\right), i \neq j$ ; // crosstalk
  $P_i^k(\mathcal{L}) \leftarrow \frac{1}{|\mathbf{P}|-1} \sum_{j \in \mathbf{P}} P_{i,j}^k(C), i \neq j$ ; // local speech
  $P_j^k(\bar{\mathcal{L}}) \leftarrow P_i^k(\mathcal{L}) \cdot P_{i,j}^k(C), j \neq i$ ; // remain silent
  $P_j^k(\mathcal{L}|C) \leftarrow \frac{P_j^k(C|\mathcal{L}) \cdot P_j^k(\mathcal{L})}{P_{i,j}^k(C)} \approx 1/|\mathbf{T}| \cdot \sum_{t \in \mathbf{T}} P_{j,i}^t(C) \cdot (1 - P_j^k(\bar{\mathcal{L}}))/P_{i,j}^k(C), j \neq i, \mathbf{T} \subset \mathbf{F}$
**end**

/* Step 2: detect voice activities */

**foreach** *frame* $k \in \mathbf{F}$ **do**
  $i \leftarrow \arg\max\left(\text{mean}(\mathrm{S}_p^k)\right), p \in \mathbf{P}$;
  **foreach** $p \in \mathbf{P}$ **do**
    $\mathrm{A}_p^k \leftarrow 0$ ; // silent by default
    **if** $p == i$ **then**
      // compare($x$, $y$) $\leftarrow$ 1 if $x > y$, otherwise 0
      $\mathrm{A}_p^k \leftarrow \text{compare}(P_p^k(\mathcal{L}), P_p^k(\bar{\mathcal{L}}))$ ;
    **else**
      $\mathrm{A}_p^k \leftarrow \text{compare}(P_p^k(\mathcal{L}|C), P_p^k(\bar{\mathcal{L}}))$ ;
    **end**
  **end**
**end**

$\mathrm{S}_j^k$ is smaller than that of $\mathrm{S}_i^k$.

$$\begin{cases} \mathrm{S}_i^k = \mathbf{V}_i^k + \mu \approx \mathbf{V}_i^k \\ \mathrm{S}_j^k = \phi_{ij} \cdot \mathbf{V}_i^k + \mu \approx \phi_{ij} \cdot \mathbf{V}_i^k. \end{cases} \tag{1}$$

Based on the observation, we propose a Bayesian VAD algorithm. The main idea is that the correlation patterns within people's audio signals could help to identify a fraction of frames when only one person is likely to speak and others are likely to remain silent. Then based on the speaking and silent frames of a given individual, we could learn his vocal features including mean value and standard deviation and detect his voice activities in all frames. Lastly, we use the correlation again and rectify false activities caused by crosstalk.

Detailed steps are illustrated in Algorithm 1 which consists of two steps. The first step calculates the probability of the following four cases where $\mathcal{L}$ denotes local speech and $C$ denotes crosstalk.

—$P_i(\mathcal{L})$: probability that participant $i$ talks.
—$P_j(\bar{\mathcal{L}})$: probability that participant $j$ remains silent.
—$P_{i,j}(C)$: probability that $i$'s voice appears in $j$'s badge signal.
—$P_j(\mathcal{L}|C)$: probability that $j$ also talks when others talk.

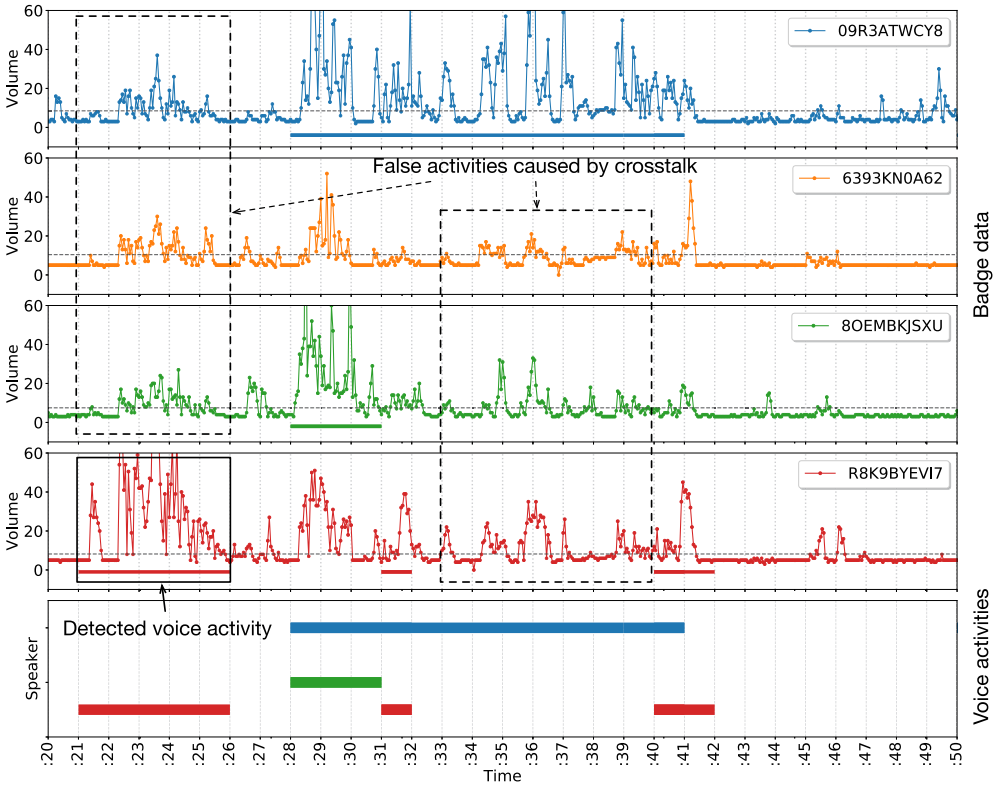



Fig. 3. An example result of Bayesian VAD of a meeting with four participants between 18:12:20 and 18:12:50.

For any frame, we find the loudest speaker (participant $i$) first as s/he is more likely to cause crosstalk (line 2). Then we calculate the probability of crosstalk from $i$ to $j$ using Pearson Correlation Coefficient. The larger the correlation, the higher probability that $j$'s signals are caused by $i$'s crosstalk (line 3). Also, we define the probability of $i$'s local speech as an average of the probabilities of his crosstalk to others. This means if other participants all have high correlations with $i$'s signals, $i$ is more likely to speak at that frame (line 4). On the contrary, the probabilities of others remaining silent are directly related to the probability of local speech of the loudest speaker $i$ and their probability of crosstalk from $i$ (line 5). Lastly, $P_j(\mathcal{L}|C)$ represents the probability of local speech of $j$ under the impact of crosstalk from $i$ (line 6). Specifically, we use the average correlation of $P_{j,i}(C)$ when $j$ is the loudest speaker to approximate $P_j(\mathcal{L}|C)$. The second step detects voice activities in all frames of the meeting. If the target person $p$ is the loudest speaker, we compare the distributions of $P_p^k(\mathcal{L})$ and $P_p^k(\bar{\mathcal{L}})$ since the impact of crosstalk could be ignored (line 13). If the probability of local speech is larger, then $p$ talks during frame $k$. Otherwise, we need to consider crosstalk and compare $P_p^k(\mathcal{L}|C)$ and $P_p^k(\bar{\mathcal{L}})$ (line 15). The complexity of the algorithm is linearly associated with the number of frames (the duration of a meeting), the size of a frame (fixed in 1 second in our setting), and the number of participants of a meeting.

Compared to other approaches, the advantages of the Bayesian VAD are threefold. First, the algorithm can adaptively learn the vocal features specific to given individuals. Second, it could identify situations when voice activity is caused by crosstalk. Last but not least, the Bayesian method avoids threshold setting tasks which are difficult in many cases. Figure 3 shows an example result of the

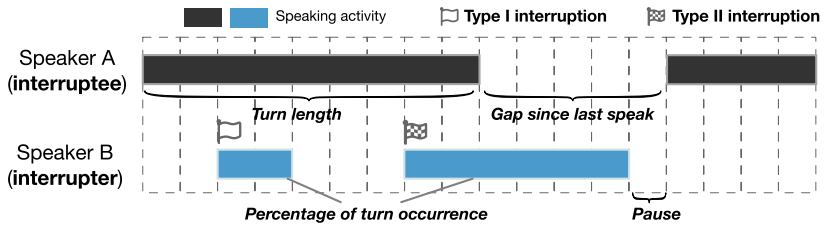



Fig. 4. Illustration of conversational features. Italic bold texts represent turn-taking features; the other bold texts represent interruption features.

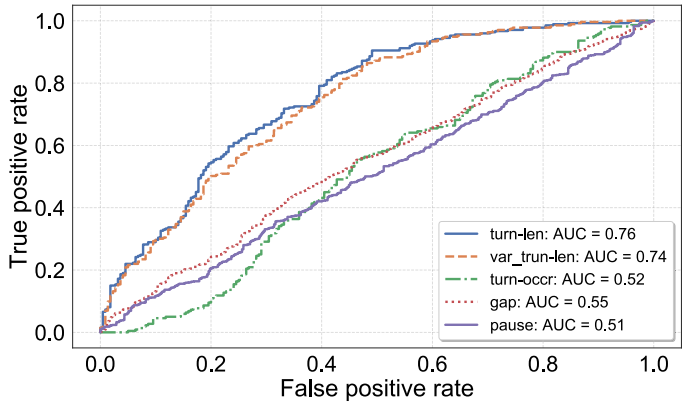



Fig. 5. ROC curves of different turn-taking features in gender identification.

proposed Bayesian VAD. The first four sub-figures reveal the badge data of four participants. It is clear that participants' badge signals within the box are likely to speak. However, these false activities are just caused by the crosstalk of the blue guy. As illustrated in the last sub-figure, we could see the proposed algorithm rectifies these false detections and detects voice activities for all participants effectively.

### 3.2 Conversational Feature Extraction

Based on the detected voice activities, conversational features are then extracted in this component. We define two kinds of conversational features or indicators, namely, turn-taking behaviors and interruption patterns.

***Turn-taking.*** As illustrated in Figure 4, turn-taking features contain turn length (how long a person's turn lasts, denoted as turn-len), the percentage of turn occurrence (how frequently a person speaks, denoted as turn-occr), pause between any consecutive turns, and the gap since the participant last speaks [37]. In addition, we also take the variance of turn length (var_trun-len) into consideration.

Through analysis of the data collected from the MIT Sloan Fellows program (see Section 4), we find that the average turn length of women (2.6 seconds) is shorter than that of men (3.2 seconds). In addition, only turn length and its variance are informative in identifying gender as shown in Figure 5. To measure the effectiveness of turn-taking features in gender identification, we exploit the **Receiver Operating Characteristic** (**ROC**) curve which is usually used to illustrate the diagnostic ability of a binary classifier as its discrimination threshold varies. The curve is created by

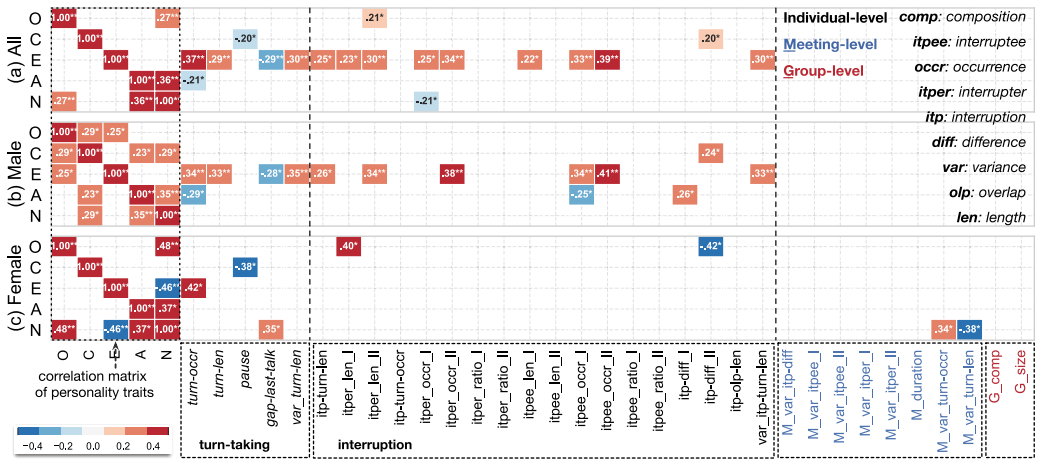



Fig. 6. Correlation of personality traits and conversational features (partial). (a) All participants; (b) Males only; (c) Females only. $*, p < 0.05$; $**, p < 0.001$.

plotting the **true-positive rate** (**TPR**) against the **false-positive rate** (**FPR**) at various threshold settings. Obviously, the turn-length-related features are more effective than others in identifying gender.

Batrinca et al. discussed a "particular speaking style" that "they (extrovert people) talk more, louder, faster and have fewer hesitations" [5]. "Talk more" herein could be captured by turn-taking indicators like more turn occurrence and longer turn length. In addition, turn length is also used for recognizing personality [39]. Figure 6 quantifies part of the correlation of turn-taking indicators with different personality traits. To derive the results, we tried several correlation coefficients including Pearson, Kendall, and Spearman. As there barely exist significant differences, we use the Pearson Correlation Coefficient as an example. For Figure 6(a), it is obvious that Extraversion (E) are correlated with most indicators since E is defined as active and talkative. For example, extroverts tend to have more turn occurrence ($\rho_{E,\text{turn-occr}} = 0.37$), longer turn duration ($\rho = 0.29$), larger variances of turn length ($\rho = 0.3$), and shorter gaps since last talk ($\rho = -0.29$). These results are generally consistent with Batrinca's findings [5]. Moreover, there is a negative correlation ($\rho = -0.21$) between Agreeableness (A) and turn-occr. Higher values of A indicate generousness and carefulness and thus might lead to smaller willing to take turns to speak. Lastly, Conscientiousness (C) negatively correlates with pause, which means higher values of C correspond to shorter pauses. People scoring high in C are described as efficient and organized. Usually, they could plan themselves in better ways and thus lead to shorter hesitations in a conversation.

Comparing (a), (b), and (c) of Figure 6, we find that different genders reveal distinct correlation patterns. For example, female extroverts have a stronger positive correlation in turn occurrence than men. However, there is no significant correlation between turn duration or gaps, which is different from that of men. In addition, some correlation, like the correlation between C and pause, only exists among a certain gender. It indicates that even for the same personality trait it might have different interpretability for different genders.

***Interruption.*** We define two roles of interruption as shown in Figure 4. An *interrupter* (itper) is a person who starts his or her turn before others' turns finish while an *interruptee* (itpee) is a person that is interrupted. According to literature, interruption is classified as cooperative and disruptive interruption [48, 57]. Cooperative interruption is mostly words of agreement and support

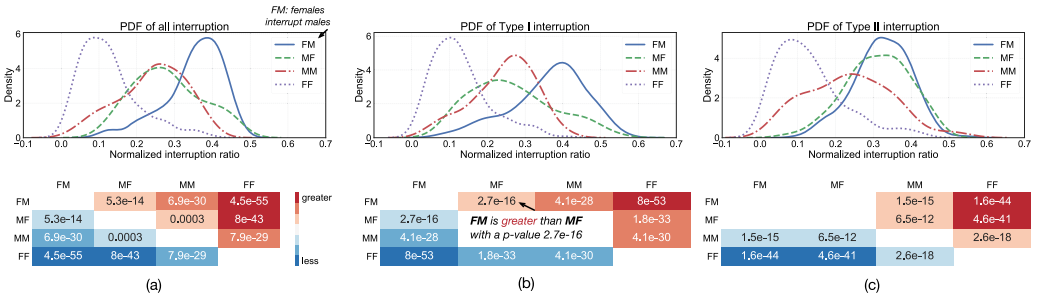
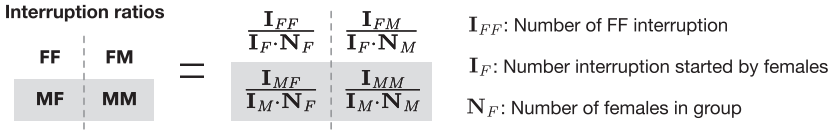



Fig. 7. Analysis of different types of interruption. (a) All interruption; (b) Type I interruption; (c) Type II interruption.

or anticipation of how other people's sentences and thoughts would end. Disruptive interruption, on the other hand, is having a tendency to take the floor or switch the topic. However, cooperative interruption and disruptive interruption are too complex and difficult to detect without conversational context. Therefore, to capture the latent difference, we devise two types of interruption as an alternative. *Type I interruption* could be regarded as a mixture of unsuccessful disruptive interruption and cooperative interruption, while the majority of *Type II interruption* is the successful disruptive interruption. The main difference between them is that interrupters of Type II managed to take the floor.

The majority of interruption indicators could be expressed as {role} × {len, occr, ratio} × {type}. For example, itper_len_I means the average length of Type I interruption when a participant acts as an interrupter. Indicator itpee_occr means the occurrence of interruption when a participant is interrupted. Indicator itp-diff represents the difference between itper_occr and itpee_occr.

After the analysis of the collected data, we notice that women interrupt men more frequently than vice versa, which is contrary to most existing findings in sociology studies [57, 58]. Notably, this finding is subject to a group of people with certain backgrounds. It might be unsuitable to generalize the conclusion without further study. 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. The normalized interruption ratio is a normalization of each ratio over their total sum. To show the relation of pairwise classes of interruption, we resort to the 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. After analyzing probability density functions of different classes of interruption with Mann-Whitney U test, we derive an interesting finding: the relations between four-class interruption are also different as illustrated in Figure 7. For all interruption as shown in Figure 7(a), the relationship of four-class interruption is FM > MF > MM > FF. For Type I interruption in Figure 7(b), the relationship mostly holds except there is no significant difference between MF and MM. The PDFs of Type II interruption as illustrated in Figure 7(c) indicate that there is no significant difference between FM and MF.

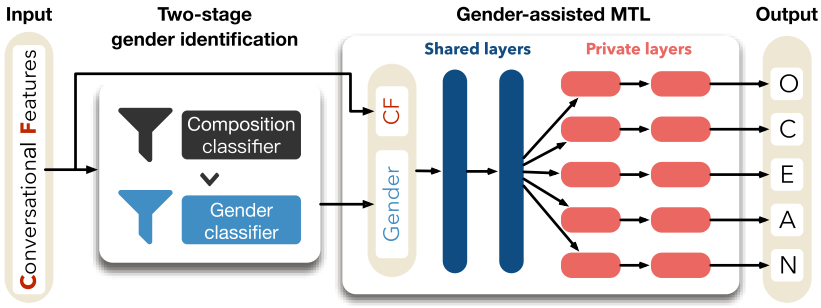



Fig. 8. Illustration of the proposed two-stage gender classification and gender-assisted MTL for personality recognition.

We also show the correlation of interruption indicators with personality traits in Figure 6. Similar to turn-taking indicators, E has the most correlation and different genders reveal different correlation patterns. From Figure 6(a), we could find that extroverts are more likely to have interruption especially being interrupted. Furthermore, the correlation between itper_occr_II ($\rho = 0.34$) is higher than that of itper_occr_I ($\rho = 0.25$). This is understandable since to have more turns extroverts have to interrupt other people more frequently, especially via Type II interruption. Meanwhile, the more turns extroverts take, the higher probability they get interrupted by other participants. This could be one of the potential reasons for the high positive correlation of itpee_occr_I and itpee_occr_II. We also observe the sparse correlation of other traits. For instance, Openness (O) has a positive correlation with itper_len_II ($\rho = 0.21$). The trait O is depicted as curious, which might develop more interests in others' opinions to have a more thorough discussion, as a result of which might derive a longer Type II interruption.

***Multi-Level Features.*** According to previous results, not only turn-taking and interruption indicators are important; we also found that gender is an important factor in recognizing personality traits. This finding is validated in a recent work [2].

As explained in the Introduction, an individual's conversational behaviors could be affected by emotional and environmental factors. For example, people behave differently in groups with different gender compositions. Interruption is more evenly distributed in same-gendered groups [29]. Current indicators remain at the individual-level, which makes it difficult to predict stable personality traits effectively. Therefore, we devise additional group-level indicators, including group size (G_size) and group gender composition (G_comp), to partially explain the dynamics of an individual's conversational behaviors. We also calculate the variances of some individual-level features as meeting-level features (features begin with "M" in Figure 6) including turn length and the occurrence of both types of interruption. These meeting-level indicators are intended to illustrate intra-group interactions and eventually capture behavior dynamics.

### 3.3 Inferring Attributes of Interest

In this component, we propose a gender identification model and use the inferred gender as an additional input for personality recognition as shown in Figure 8. First, we estimate the gender composition of the group (same-gender or cross-gender) and then incorporate this information in gender identification. Second, based on the inferred gender information, we develop a gender-assisted **Multi-Task Learning** (**MTL**) approach taking both personality trait correlation and gender differences into consideration.

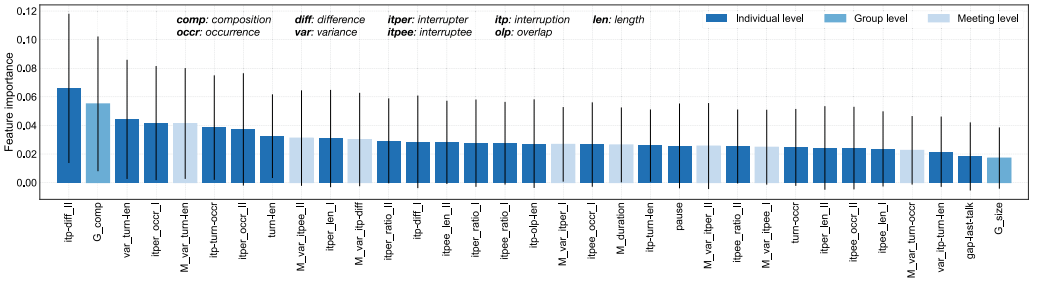



Fig. 9. Feature importance in gender identification (derived from a Random Forest consisting of 100 trees).

***Two-Stage Gender Identification.*** We observe that conversational features are closely related to gender, which is the foundation of gender identification. In addition, we also find gender composition is helpful in identifying gender information for the whole group. Therefore, we propose a two-stage classification method. The main idea is to infer the latent information of gender composition and treat it as an additional input feature for gender identification.

In the first stage, we use meeting-level indicators of each group to predict its gender composition. Each participant in the group meeting has two roles, interrupter and interruptee. We notice the variance of the difference between interrupter and interruptee in a meeting (M_var_itp-diff) is a good indicator of gender composition. A group with the same gender is prone to have smaller variance as interruption is more evenly distributed in same-gender groups [29]. In the second stage, we combine the selected features and the inferred gender composition as input to predict gender for each participant of the group. In both stages, we choose popular classification models like linear SVM and Random Forest.

We also show the importance of the features in Figure 9. A Random Forest of 100 trees is used to evaluate their importance on the task. Each bar represents the importance of a certain feature, along with its inter-tree variability. The result indicates that gender composition (G_comp) is one of the most important features for identifying gender.

***Gender-Assisted MTL Personality Recognition.*** Figure 6 discloses certain correlations among some personality traits. This motivates us to learn the five traits simultaneously with MTL. Jointly learning multiple tasks could improve a model by introducing an inductive bias that prompts the model to prefer some hypotheses over others [40]. For example, $\ell 1$ regularization is a common form of inductive bias which leads to a preference for sparse solutions. In MTL, the inductive bias is provided by the auxiliary tasks. With the inductive bias, models prefer hypotheses that explain multiple tasks, leading to a better generalization.

Comparing (b) and (c) of Figure 6, we notice that the correlation patterns have obvious differences. These inherent differences (aka gender differences) naturally exist [18], which are insightful in understanding human societies. For example, the significant positive correlation between N and O and the significant negative correlation between E and N are merely detected among women. In addition, men's conversational behaviors are mostly correlated with E, which is intuitively understandable since E is described as active and talkative. However, this intuition does not work for women. Therefore, we propose a gender-assisted MTL approach as illustrated in Figure 8. Generally speaking, we combine the gender-related information and conversational features using a hard parameter sharing MTL to incorporate both trait correlation and gender differences. It works by sharing the hidden layers between all tasks while keeping several task-specific output layers. The more tasks we learn jointly, the less is the chance of overfitting on the original task.

**ALGORITHM 2:** Joint learning process of multiple tasks

**Input** : $\mathbf{D}_m, 1 \le m \le 5$: Datasets of 5 tasks
$T$: the maximum iteration; $\alpha$: learning rate
**Output**: Model $f_m, 1 \le m \le 5$

```
foreach iteration t ∈ [0, T] do
    foreach task m ∈ [1, 5] do
        B_m ← split D_m into batches
    end
end
B̄ = Randomise(∪_{m=1}^{5} B_m)
foreach batch b ∈ B̄ do
    calculate loss L(θ) over batch b
    θ_t ← θ_{t−1} − α · ∇_θ L(θ)
end
```

There are $M = 5$ correlated tasks. $\mathcal{D}_m$ is the dataset for the $m$-th task with $N_m$ samples:

$$\mathcal{D}_m = \left\{x^{(m,n)}, y^{(m,n)}\right\}_{n=1}^{N_m}, \tag{2}$$

where $x$ and $y$ represent the training data and ground truth, respectively. Suppose $f_m(x;\theta)$ is the model for the $m$-th task; multi-task learning aims to minimize the linearly joint objective function:

$$\mathcal{L}(\theta) = \sum_{m=1}^{M}\sum_{n=1}^{N_m} w_m \mathcal{L}_m(f_m((x^{(m,n)};\theta), y^{(m,n)})), \tag{3}$$

where $\mathcal{L}_m(\cdot)$ is the cross-entropy loss function of the $m$-th task, $w_m$ is the weight of the $m$-th task, and $\theta$ are the parameters including both shared and private layers. The weights are assigned according to the levels of importance or difficulty. Mostly, all tasks have the same weight, namely, $w_m = 1$.

The learning process consists of two steps: joint training of multiple tasks and fine-tuning of every single task. For each iteration, a random task was chosen with gradient descent algorithms to update parameters as illustrated in Algorithm 2. Based on the parameters derived from multi-task learning, fine-tuning of every single task leads to better performances.

## 4 EXPERIMENTAL EVALUATION

In this section, we evaluate and discuss the performance of the proposed user profiling system. Experiment Settings include the setup of experiments, the collected dataset, baseline approaches, and evaluation metrics. Evaluation Results consist of the performance of voice activity detection, gender identification, and personality recognition. The effectiveness of the extracted multi-level features and the proposed gender-assisted multi-task learning model are also evaluated.

### 4.1 Experiment Settings

***Setup.*** We collected the nonlinguistic audio data from spontaneous face-to-face meetings of MIT Sloan Fellows class of 2016/17 for 4 weeks. 100 out of the 110 students enrolled in the study, including 31 females and 69 males. They come from 35 different countries with an average age of

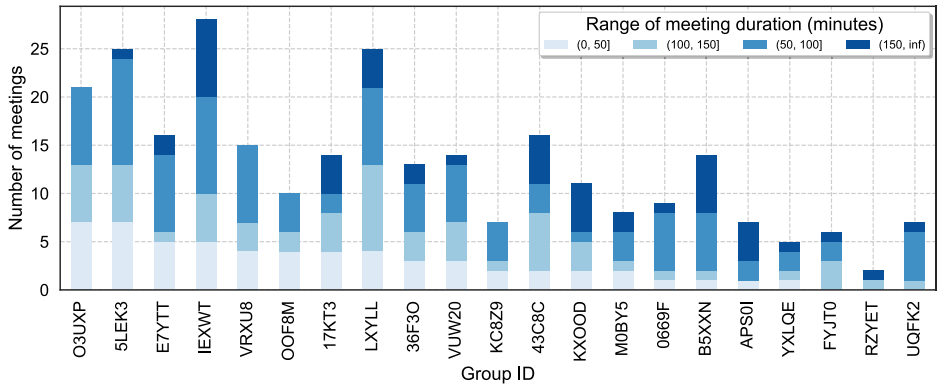



Fig. 10. Stacked histogram of number of meetings and meeting duration of all groups.

37.41 ± 4.45 years (mean ± standard deviation) and an average work experience of 13.78 ± 4.24 years. All participants gave written informed consent of their participation in the study.

As group collaboration is explicitly valued in the program, students are assigned to study groups of four or five students before the program starts. There are 21 study groups with five same-gender groups and 15 mixed-gender groups. During the whole program, all groups remain unchanged, and students meet regularly to work on the courses together. There are no requirements on how often and how long they should meet.

***Dataset.*** After the study, we collected 273 effective meetings from 21 groups with a total length of 438.25 hours. On average, each group had 13 meetings, but still, some groups had around 5 meetings as illustrated in Figure 10. Over half of those meetings last for more than 100 minutes. We also collected nonlinguistic audio and video recordings from four meetings with a total length of 1.1 hours. Those meetings are held in scenarios with different levels of background noises and different participants. Based on the video recordings, we manually annotate the voice activities of each participant to evaluate the performance of the proposed Bayesian VAD.

As introduced in Section 2, we exploit the TIPI [13] to get the ground truth of students' personalities. TIPI is a brief measure of the Big-Five personality traits (see Introduction). It contains two items for each of the five personality traits. Each item is rated on a seven-point scale ranging from 1 (disagree strongly) to 7 (agree strongly). Although TIPI is considered to be inferior to longer measures of Big-Five, it has been shown to be an adequate measure when brevity has higher priority [46]. As it is extremely difficult to collect such data from a large number of people and long tedious surveys would discourage volunteers from participating in the event. Considering this, TIPI is often a good tradeoff [46]. We further separate each score into three different levels: Low: 1–3, Mid: 3.5–5, and High: 5.5–7. Figure 11 demonstrates the distributions of five traits for both genders. We could notice that some traits (like C, N, and O) are biased. In addition, different genders have different distributions. For example, women have higher Openness than men.

***Baseline Approaches.*** As mentioned, to the best of our knowledge, there are no existing works exactly doing user profiling with nonlinguistic audio. Therefore, we use baseline approaches to validate our technical contributions which are threefold. The first contribution is the Bayesian voice activity detection algorithm which is parameter-free and could detect voice activities adaptively. The second contribution is the devised multi-level features. Compared to using only individual-level features, multi-level features could capture intra-group interaction and model contextual factors leading to more effective performance. The third contribution relies on the proposed

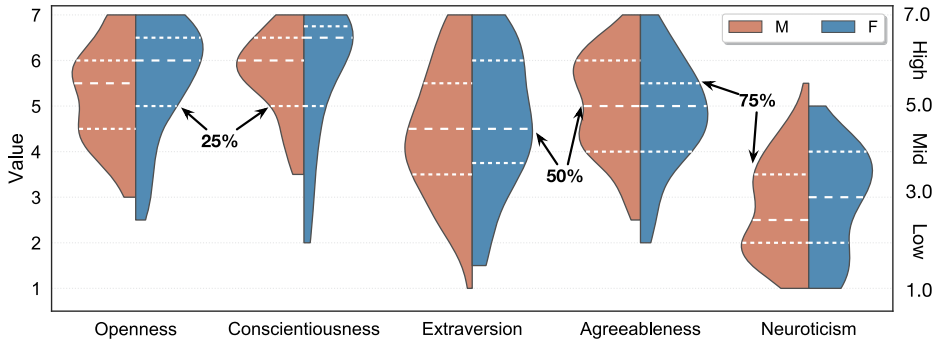



Fig. 11. Distributions of ground truth OCEAN data.

gender-assisted MTL model for personality recognition. Due to the existence of gender differences in conversational behaviors and personality, data from the same gender are more cohesive for learning.

For the first contribution, we use a threshold method used in the literature [20] as a baseline. It works in a straightforward manner as depicted in Equation (4). If the mean value of signals from user $i$ within frame $k$ is larger than $\delta$, then a voice activity of $i$ is detected.

$$\mathrm{A}_i^k = \begin{cases} 1 \text{ if } \mathrm{mean}(\mathrm{S}_i^k) > \delta \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

To validate the last two contributions and to demonstrate the effectiveness of gender identification and personality recognition, we have devised various baseline approaches. Figure 12 shows the detailed configurations in terms of target AoI, feature space, and learning models.

**Automatic personality recognition** (**APR**) could be solved with existing multi-label classification techniques, including Binary Relevance, Classifier Chains, and Label Powerset. These techniques work by combining classic classification models like K-Nearest Neighbor. Binary Relevance (BR) is the most straightforward technique, which treats each label or task as a separate multi-class classification problem. For example, BR K-Nearest Neighbor (BR_KNN) solves APR by applying KNN to each task separately. In **Classifier Chain** (**CC**), the first classifier is trained just on the input data and then each next classifier is trained on the input data and all the other previous classifiers in the chain. For **Label Powerset** (**LP**), the problem is transformed into a multi-class problem with one classifier trained on all unique label combinations.

We could verify the effectiveness of multi-level features through the comparison of methods using the individual-level features and multi-level features in both gender identification (GI vs. GM; BR_idl vs. BR) and personality recognition (GAMTL_idl vs. GAMTL). To evaluate the proposed gender-assisted MTL approach, we also compare the performance of different methods based on the same input features (BR vs. GAMTL vs. MTL vs. NN).

***Evaluation Metrics.*** Gender identification, personality recognition, and voice activity detection are classification problems in essence. Therefore, we use precision, recall, and F1-score to evaluate their performance.

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

| Truth → / Prediction ↓ | $X$ | $\tilde{X}$ |
|---|---|---|
| $X$ | tp | fp |
| $\tilde{X}$ | fn | tn |

$X$ Target label {low, mid, high}

$\tilde{X}$ Non-target label

| Code | Target AoI | Feature space | Learning models |
|---|---|---|---|
| GI | Gender | Individual-level | KNN, SVM, random forest, … |
| GM | Gender | Multi-level | KNN, SVM, random forest, … |
| {BR/CC/LP}_idl | Personality | Individual-level | KNN, SVM, random forest, … |
| BR/CC/LP | Personality | Multi-level | KNN, SVM, random forest, … |
| GAMTL_idl | Personality | Individual-level | Gender-assisted MTL |
| GAMTL | Personality | Multi-level | Gender-assisted MTL |
| MTL | Personality | Multi-level | MTL |
| NN | Personality | Multi-level | Neural networks for each trait |

*Evaluate multi-level features*

*Evaluate gender-assisted MTL*

Fig. 12. The detailed configurations of baseline approaches.

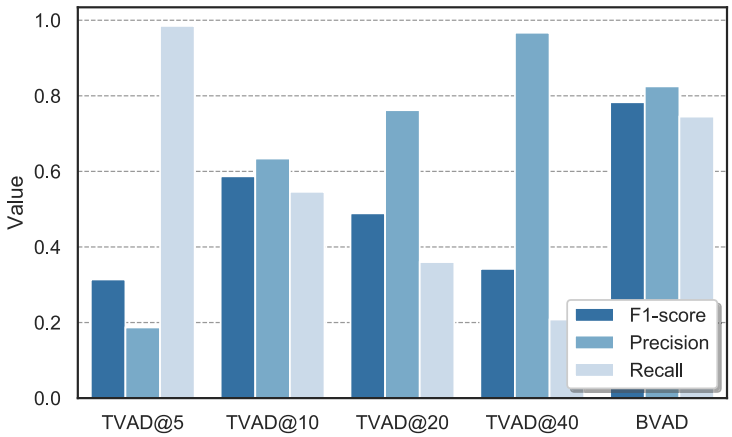



Fig. 13. Performance of Bayesian VAD (BVAD) and threshold VAD (TVAD). "TVAD@5" means $\delta = 5$.

Considering the imbalance in numbers of different classes, we use a weighted version of those metrics. The weighted F1-score is calculated with Equation (5) where $S_H$ is the number of true "high" instances and $\mathrm{F1}_H$ is the F1-score for the class "high." The weighted versions of precision and recall are derived in a similar way.

$$\mathrm{F1} = \frac{S_H \cdot \mathrm{F1}_H}{S_H + S_M + S_L} + \frac{S_M \cdot \mathrm{F1}_M}{S_H + S_M + S_L} + \frac{S_L \cdot \mathrm{F1}_L}{S_H + S_M + S_L}. \tag{5}$$

***Parameter Selection.*** Although there are no parameters in the proposed system, we have parameters for baseline approaches including the threshold $\delta$ which will be discussed later. For parameters of a certain model in different baselines like the number of neighbors in KNN, they share the same default settings as specified in scikit-learn [32].

## 4.2 Evaluation Results

Given the limited size of data samples, all the experimental results are derived from 10-fold cross-validation.

***Performance of Voice Activity Detection.*** Figure 13 illustrates the performance of **threshold VAD** (**TVAD**) and **Bayesian VAD** (**BVAD**). The proposed BVAD significantly outperforms TVAD by at least 33.4% and achieves an F1-score of 0.783. Both precision and recall of BVAD outperform that of TVAD owing to the capacity to capture individual vocal features and differentiate

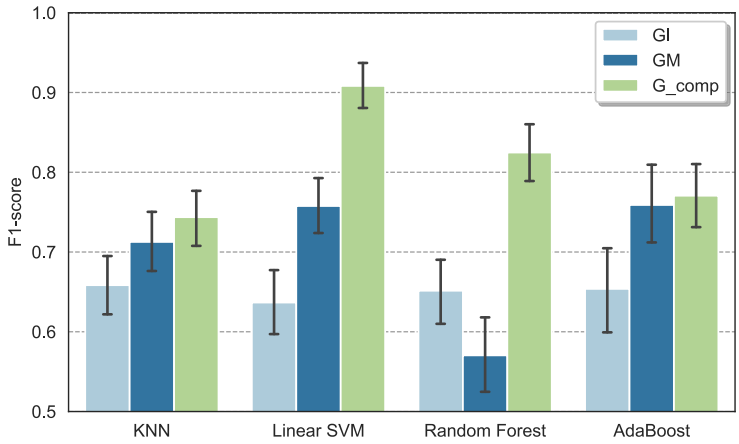



Fig. 14. Performance of gender composition detection (G_comp) and gender identification with different features.

crosstalk. When the threshold is small, crosstalk and noises could be recognized as voice activities resulting in false-positive detections. The detected voice activity could be caused by crosstalk if there are participants nearby with relatively louder voices due to the fact people have varied vocal features including loudness. Therefore, the precision generally increases with the threshold, while large thresholds will neglect voice activities from participants with relatively lower voices. As shown in Figure 13, TVAD@40 has poor recall due to a large number of false-negative detections.

***Performance of Gender Identification.*** We evaluate the performance of gender composition (G_comp) detection and gender identification on selected classification models including KNN, Linear SVM, Random Forest, and AdaBoost. As mentioned in parameter selection, all the parameters used are default settings in scikit-learn. The results are illustrated in Figure 14. For gender composition detection, as the number of groups is small, we repeat the 10-fold cross-validation process five times. It is clear that linear SVM outperforms other models and achieves an F1-score over 0.9. As explained, same-gender groups have evenly distributed interruption patterns. In such groups, the difference between a person being an interrupter and an interruptee is small.

We choose Linear SVM as the composition classifier and regard the inferred gender composition as group-level features for gender identification. As shown in Figure 14, except Random Forest, all models using multi-level features (GM) outperform models with individual-level features (GI). On average, GM outperforms GI by 7.7% in F1-score. Gender composition could partially address the instability of conversational behaviors and thus increase the interpretability of conversational features. As explained, human behaviors could be readily affected. With meeting-level and group-level features capturing intra-group interaction and external factors, we could explain the dynamics of conversational behaviors to a certain extent. The best performance is achieved on AdaBoost with an F1-score of 0.759. Therefore, we choose AdaBoost as gender classifier for personality recognition.

***Performance of Personality Recognition.*** Table 1 summarizes the recognition performance of four approaches in five personality traits. For all traits, the proposed gender-assisted MTL model with multi-level features outperforms other approaches. From high to low, the average F1-score of five traits are as follows: GAMTL (0.652) > MTL (0.620) > GAMTL_idl (0.600) > NN (0.571).

Table 1. Performance of Five Personality Traits

| Method\Trait | Openness (O) | | | Conscientiousness (C) | | | Extraversion (E) | | | Agreeableness (A) | | | Neuroticism (N) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| NN | 0.793 | 0.573 | 0.665 | 0.600 | 0.639 | 0.619 | 0.634 | 0.492 | 0.554 | 0.533 | 0.487 | 0.509 | 0.537 | 0.486 | 0.510 |
| MTL | 0.854 | 0.622 | 0.720 | 0.655 | 0.661 | 0.658 | 0.660 | 0.558 | 0.605 | 0.585 | 0.579 | 0.582 | 0.577 | 0.502 | 0.537 |
| GAMTL_idl | 0.724 | 0.695 | 0.709 | 0.595 | 0.701 | 0.644 | 0.574 | 0.588 | 0.581 | 0.577 | 0.500 | 0.536 | 0.680 | 0.426 | 0.524 |
| GAMTL | 0.828 | 0.706 | **0.762** | 0.682 | 0.659 | **0.670** | 0.608 | 0.604 | **0.606** | 0.663 | 0.639 | **0.651** | 0.709 | 0.475 | **0.569** |

Bold text represents the best performance among four methods on a certain trait. P: Precision, R: Recall, F1: F1-Score.

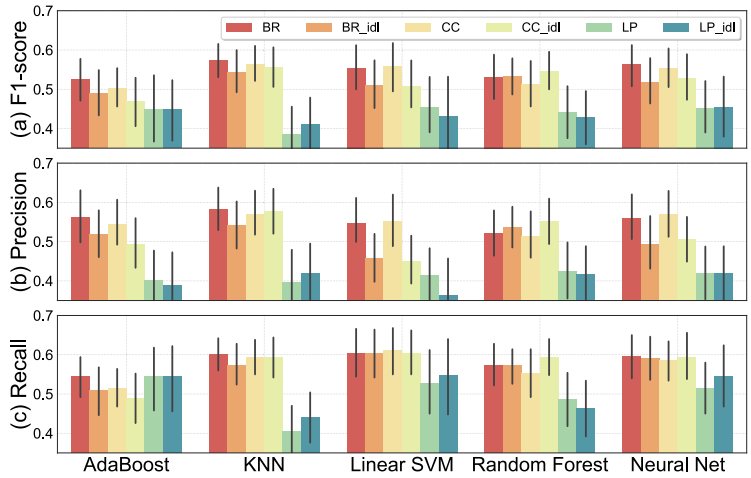



Fig. 15. Comparison of baseline approaches with and without additional levels of features.

The performance gains of GAMTL over GAMTL_idl range between 4.3% and 21.5%, which demonstrate the effectiveness of multi-level features. As explained, individual-level features could hardly reflect the intra-group interaction and social contexts, like the gender composition of the group, are also important in recognizing personality traits. Also, the average F1-score of GAMTL outperforms MTL and NN by 8.7% and 14.2%, respectively. It reveals the proposed gender-assisted structure is effective in improving recognition performance. The improvements are owing to the appropriate manipulations of gender differences in personality and the correlation between different traits.

***Effectiveness of Multi-Level Features.*** The effectiveness of multi-level features is evaluated on selected classifiers: AdaBoost, KNN, Linear SVM, Random Forest, and Neural Network (Multilayer Perceptron). The parameter settings for all models are consistent with different baselines. For example, KNN classifier uses the same parameter $k = 3$ for BR, CC, and so forth. The results are derived from repeated (five times) 10-fold cross-validation to ensure authenticity.

As clearly shown in Figure 15, most approaches using multi-level features outperform methods with single individual-level features. More specifically, multi-level features could improve the average F1-score by 7.49% and 5.73% for BR and CC approaches (except Random forest), respectively. We further find that the performance gains are mainly contributed by better precision. The utilization of meeting-level and group-level features could partially explain the dynamics of conversational behaviors. In addition, BR methods generally outperform other methods, especially LP approaches. This indicates when addressing the APR problem, multi-label techniques may not achieve satisfying results. A potential reason for the poor performance of LP is that unique label combinations will inevitably reduce the number of training data for each new class label.

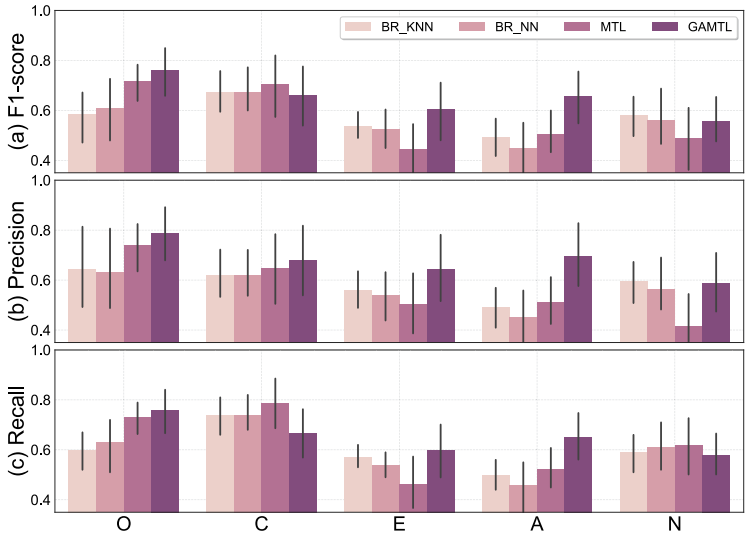



Fig. 16. Comparison of selected methods for evaluating trait correlation and gender difference.

**Effectiveness of Trait Correlation and Gender Difference.** Since the effectiveness of multi-level features has been confirmed, we select the top two baseline approaches (BR_KNN and BR_NN) to compare with the proposed **gender-aware MTL method** (**GAMTL**). The results are also obtained with repeated cross-validation.

All four methods have the same input but differ in the classifier and how they make use of the gender information. Figure 16 shows that GAMTL outperforms the other three methods by 12.93%–15.14% in F1-score on average. This demonstrates both trait correlation and gender difference are effective in improving the performance of personality recognition. By comparing MTL and BR_NN, we could further find that the effectiveness of trait correlation is relatively limited. A feasible explanation is that the correlations between tasks are relatively sparse and weak and thus have limited contribution to the performance gain. However, if we look at GAMTL and BR_NN, there is a significant improvement when combining trait correlation with gender information. We summarize the latent reasons could be twofold. On the one hand, correlation patterns of both genders are different; some associations cannot be revealed when treated as a whole. On the other, both genders have distinct conversational patterns, which are difficult to capture with one neural structure and one set of parameters.

***Discussion on Definitions of Trait Levels.*** Instead of using the proposed absolute thresholds, the existing approach [2] defines trait levels with relative thresholds derived from population norms. Both definitions have their physical meanings. Absolute thresholds aim to recognize personality in a global view while relative thresholds are expected to identify levels in a certain population.

The accuracy of [2] ranges between 37% and 44% for triple classification. The average F1-score of GAMTL with relative thresholds is 54%. The proposed method significantly outperforms the existing approach by at least 20%. However, there is a non-negligible performance decrease. A latent explanation is that when using relative thresholds, some tasks (like C), which are originally close to binary classification, become triple classification. The increased difficulty results in the decreases in recognition performance.

## 5 RELATED WORKS

Recent decades have witnessed the rapid development of human sensing using various modalities [9, 42, 43, 51, 52]. It also facilitates a wide range of applications [8, 22, 23, 31]. Here we discuss three research areas that are most related including voice activity detection, gender identification, and personality computing.

***Voice Activity Detection.*** Traditional VAD methods rely on multi-class classification. First, acoustic features are extracted from raw audio. Then classification models like Hidden Markov Model [53] or Gaussian Mixture Model [33] are utilized to detect voice activities. However, most valuable features could not be extracted from PS audio. In addition, it is difficult to adapt to scenarios without training data.

Another type of methods regards VAD as a blind source separation problem and solves it with **Independent Component Analysis** (**ICA**) [25]. However, ICA assumes stationary mixing of the signal, i.e., requires participants to remain motionless. Such a constraint is difficult to meet as people may walk around in real situations. Also, it is still non-trivial to separate speech and noise on the de-mixed signals, which is not resilient to different environments.

***Gender Identification.*** Voice-based gender identification relies 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) and phonetic differences. Various identification systems with different classification models and different types of features have been reported in the literature [1, 16, 35]. The most frequently used features are pitch [16] and first formant [35], which are closely related to voice sources and vocal tract, respectively.

***Personality Computing.*** Personality Computing addresses three fundamental problems [7, 14, 27, 49]. The recognition of the true personality of an individual ( APR), the prediction of the personality others attribute to a given individual (Automatic Personality Perception (APP)), and the generation of artificial personalities through embodied agents (Automatic Personality Synthesis).

Despite extensive efforts on personality computing, most attention is on APP rather than APR [2]. One of the potential reasons is that getting true personality via self-reports is more difficult than getting personality ratings from others. In addition, APR is more challenging. In a very comprehensive study, both self-reported and observer-rated personality scores are predicted from the essay and conversational data, using psycholinguistic, and prosodic feature sets. Models of observed personality achieved good results while no results above baseline are derived with models of self-reported personality [24].

APR has been studied with various modalities in literature. Mairesse mixed verbal and nonverbal cues [24]. The extracted features include mean, extremes and standard deviation of pitch, intensity, and speaking rate. The experiments aim to discriminate between individuals in the upper and lower halves of the observed scores of each trait. Pianesi and Sebe explored visual nonverbal features combine acoustic features like pitch and intensity to assess personality [5]. Their results show that C and N are the best recognizable traits.

A comprehensive set of features is extracted by An et. al. in [2, 3]. The features consist of linguistic features (like Linguistic Inquiry and Word Count (LIWC)) and acoustic features (like pitch). Psycholinguistic studies indicate that people choose words not only because of the linguistic meaning but also because of psychological conditions, such as emotion, personality, and relational attitude. Therefore, it is possible to detect personalities through text analyses associated with psycholinguistic techniques. In addition, previous researches have proved a variety of speech factors, such as fundamental frequency (pitch), voice quality, intensity, frequency, and duration of silent pauses, could reflect different personality traits.

## 6 CONCLUSION

In this work, we are the first to build a user profiling system from nonlinguistic audio to infer gender and personality. The effectiveness is verified with extensive experiments conducted with real study groups. Our main contributions are threefold. First, we proposed a Bayesian algorithm that could adaptively detect voice activities for nonlinguistic audio. Second, the extracted multi-level features and the proposed gender-assisted multi-task learning model are effective in user profiling. Multi-level features could capture intra-group interaction and model contextual factors leading to more effective performance. Also, due to the existence of gender differences in conversational behaviors and personality, data from the same gender are more cohesive for learning. Lastly, we analyzed face-to-face conversations in natural settings and provided evidences of gender differences in conversational behaviors and personality.

For future directions, there are two lines of research. The first direction is to improve the performance of conversational behavior-based user profiling by considering more contextual factors and developing more advanced signal processing and machine learning techniques. In real situations, there are many more contextual factors affecting conversational behaviors like language systems and the culture underneath. Capturing those factors would lead to a better understanding of conversational behaviors and thus increase the profiling performance. The second direction is to explore possibilities of profiling other AoIs like occupation, which is also of great interest in many applications.

## REFERENCES


[1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017. Multimodal gender detection. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM.

[2] Guozhen An and Rivka Levitan. 2018. Comparing approaches for mitigating intergroup variability in personality recognition. arXiv:1802.01405.

[3] Guozhen An, Sarah Ita Levitan, Rivka Levitan, Andrew Rosenberg, Michelle Levine, and Julia Hirschberg. 2016. Automatically classifying self-rated personality scores from speech. In *Interspeech*.

[4] Anthony O. Ahmed and Barbara Jenkins. 2013. Critical synthesis package: Ten-item personality inventory (TIPI)A quick scan of personality structure. *MedEdPORTAL* 9 (2013).

[5] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. 2011. Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *ICMI*. ACM.

[6] Rachel Bernstein. 2014. Communication: Spontaneous scientists. *Nature* 505, 7481 (2014), 121–123.

[7] Kseniya Buraya, Aleksandr Farseev, Andrey Filchenkov, and Tat-Seng Chua. 2017. Towards user personality profiling from multiple social networks. In *AAAI*.

[8] Yuanyi Chen, Jingyu Zhang, Minyi Guo, and Jiannong Cao. 2017. Learning user preference from heterogeneous information for store-type recommendation. *IEEE Transactions on Services Computing* (2017).

[9] Yuanyi Chen, Mingxuan Zhou, Zengwei Zheng, and Dan Chen. 2019. Time-aware smart object recommendation in social internet of things. *IEEE Internet of Things Journal* 7, 3 (2019), 2014–2027.

[10] P. T. Costa and Robert R. McCrae. 2010. The NEO personality inventory-3. *Odessa, FL: Psychological assessment resources* (2010).

[11] H. J. Eysenck. 1975. *Manual of the Eysenck Personality Questionnaire (Adult and Junior)*. Hodder & Stoughton.

[12] Adrian Furnham. 2008. Relationship among four Big Five measures of different length. *Psychological Reports* (2008).

[13] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann, Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* (2003).

[14] Ted Grover and Gloria Mark. 2017. Digital footprints: Predicting personality from temporal patterns of technology use. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM.

[15] André Hajek, Jens-Oliver Bock, and Hans-Helmut König. 2017. The role of personality in health care use: Results of a population-based longitudinal study in Germany. *PloS One* (2017).

[16] Yakun Hu, Dapeng Wu, and Antonio Nucci. 2012. Pitch-based gender identification with two-stage classification. *Security and Communication Networks* (2012).

[17] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. 2008. Paradigm shift to the integrative Big Five trait taxonomy. *Handbook of Personality: Theory and Research*, pp. 114–158. Guilford Press.

[18] Petri J. Kajonius and John Johnson. 2018. Sex differences in 30 facets of the five factor model of personality in the large public ($N$ = 320, 128). *Personality and Individual Differences* (2018).
[19] Oren Lederman, Dan Calacci, Angus MacMullen, Daniel C. Fehder, Fiona E. Murray, and Alex 'Sandy' Pentland. 2017. Open badges: A low-cost toolkit for measuring team communication and dynamics. arXiv:1710.01842.
[20] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. 2018. Rhythm: A unified measurement platform for human organizations. *IEEE MultiMedia* 25, 1 (2018), 26–38.
[21] Rui Li, Chi Wang, and Kevin Chen-Chuan Chang. 2014. User profiling in an ego network: Co-profiling attributes and relationships. In *WWW*.
[22] Wengen Li, Jiannong Cao, Jihong Guan, Man Lung Yiu, and Shuigeng Zhou. 2017. Efficient retrieval of bounded-cost informative routes. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (2017), 2182–2196.
[23] Wengen Li, Jiannong Cao, Jihong Guan, Shuigeng Zhou, Guanqing Liang, Winnie KY So, and Michal Szczecinski. 2018. A general framework for unmet demand prediction in on-demand transport services. *IEEE Transactions on Intelligent Transportation Systems* 20, 8 (2018), 2820–2830.
[24] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* (2007).
[25] S. Maraboina, Dorothea Kolossa, P. K. Bora, and Reinhold Orglmeister. 2006. Multi-speaker voice activity detection using ICA and beampattern analysis. In *2006 14th European Signal Processing Conference*. IEEE.
[26] Gerald Matthews, Ian J. Deary, and Martha C. Whiteman. 2003. *Personality Traits*. Cambridge University Press.
[27] Sarah Mennicken, Oliver Zihler, Frida Juldaschewa, Veronika Molnar, David Aggeler, and Elaine May Huang. 2016. It's like living with a friendly stranger: Perceptions of personality traits in a smart home. In *UbiComp*. ACM.
[28] Gelareh Mohammadi and Alessandro Vinciarelli. 2012. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing* 3, 3 (2012), 273–284.
[29] Anthony Mulac. 1989. Men's and women's talk in same-gender and mixed-gender dyads: Power or polemic? *Journal of Language and Social Psychology* (1989).
[30] Scott Nowson and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.
[31] Daniel Olguín-Olguín and Alex Pentland. 2010. Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering* (2010).
[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* (2011).
[33] Thilo Pfau, Daniel P. W. Ellis, and Andreas Stolcke. 2001. Multispeaker speech activity detection for the ICSI meeting recorder. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'01)*. IEEE.
[34] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *ICMI*.
[35] Kumar Rakesh, Subhangi Dutta, and Kumara Shama. 2011. Gender recognition using speech processing techniques in LABVIEW. *International Journal of Advances in Engineering & Technology* (2011).
[36] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality* (2007).
[37] Deirdre Reznik. 2004. Gender in interruptive turns at talk-in-interaction. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* (2004).
[38] Cecilia L. Ridgeway. 1992. *Gender, Interaction, and Inequality*. Springer.
[39] Giorgio Roffo, Cinzia Giorgetta, Roberta Ferrario, and Marco Cristani. 2014. Just the way you chat: Linking personality, style and recognizability in chats. In *International Workshop on Human Behavior Understanding*. Springer.
[40] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. arXiv:1706.05098.
[41] Gerard Saucier and Lewis R. Goldberg. 1996. The language of personality: Lexical perspectives. *The Five-Factor Model of Personality: Theoretical Perspectives*, pp. 21–50. Guilford Press.
[42] Jiaxing Shen, Jiannong Cao, and Xuefeng Liu. 2020. BaG: Behavior-aware group detection in crowded urban spaces using WiFi probes. *IEEE Transactions on Mobile Computing* (2020).
[43] Jiaxing Shen, Jiannong Cao, Xuefeng Liu, Jiaqi Wen, and Yuanyi Chen. 2016. Feature-based room-level localization of unmodified smartphones. In *Smart City 360*. Springer, 125–136.
[44] Jiaxing Shen, Oren Lederman, Jiannong Cao, Florian Berg, Shaojie Tang, and Alex Pentland. 2018. GINA: Group gender identification using privacy-sensitive audio data. In *ICDM*. IEEE.
[45] Tianyi Song, Xiuzhen Cheng, Hongjuan Li, Jiguo Yu, Shengling Wang, and Rongfang Bie. 2016. Detecting driver phone calls in a moving vehicle based on voice features. In *The 35th Annual IEEE International Conference on Computer Communications (IEEE INFOCOM'16)*. IEEE.

[46] Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. 2016. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2016).
[47] Angelina R. Sutin, Alan B. Zonderman, Luigi Ferrucci, and Antonio Terracciano. 2013. Personality traits and chronic disease: Implications for adult personality development. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* (2013).
[48] Deborah Tannen. 1991. *You Just Don't Understand*. Simon & Schuster Audio.
[49] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.
[50] Senzhang Wang, Jiannong Cao, and Philip Yu. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2020).
[51] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of acoustic gesture recognition. *IEEE Transactions on Mobile Computing* (2020).
[52] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID signal reflection for contact-free activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.
[53] Stuart N. Wrigley, Guy J. Brown, Vincent Wan, and Steve Renals. 2005. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing* 13, 1 (2004), 84–91.
[54] Lynn Wu, Benjamin Waber, Sinan Aral, Erik Brynjolfsson, and Alex Pentland. 2008. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. Available at SSRN 1130251 (2008).
[55] Xiangyu Xu, Hang Gao, Jiadi Yu, Yingying Chen, Yanmin Zhu, Guangtao Xue, and Minglu Li. 2017. ER: Early recognition of inattentive driving leveraging audio devices on smartphones. In *IEEE Conference on Computer Communications (IEEE INFOCOM'17)*. IEEE.
[56] Youqing Xu. 2009. Gender differences in mixed-sex conversations: A study of interruptions. (2009).
[57] Xiaoquan Zhao and Walter Gantz. 2003. Disruptive and cooperative interruptions in prime-time television fiction: The role of gender, status, and topic. *Journal of Communication* (2003).
[58] Don H. Zimmermann and Candace West. 1996. Sex roles, interruptions and silences in conversation. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4* (1996).

Received December 2020; revised May 2021; accepted June 2021