Journal on Multimodal User Interfaces (2023) 17:47–63
https://doi.org/10.1007/s12193-023-00401-0

ORIGINAL PAPER

Personality trait estimation in group discussions using multimodal analysis and speaker embedding

Candy Olivia Mawalim[1] · Shogo Okada[1] · Yukiko I. Nakano[2] · Masashi Unoki[1]

Received: 28 September 2021 / Accepted: 7 January 2023 / Published online: 8 February 2023

Abstract
The automatic estimation of personality traits is essential for many human–computer interface (HCI) applications. This paper focused on improving Big Five personality trait estimation in group discussions via multimodal analysis and transfer learning with the state-of-the-art speaker individuality feature, namely, the identity vector (i-vector) speaker embedding. The experiments were carried out by investigating the effective and robust multimodal features for estimation with two group discussion datasets, i.e., the Multimodal Task-Oriented Group Discussion (MATRICS) (in Japanese) and Emergent Leadership (ELEA) (in European languages) corpora. Subsequently, the evaluation was conducted by using leave-one-person-out cross-validation (LOPCV) and ablation tests to compare the effectiveness of each modality. The overall results showed that the speaker-dependent features, e.g., the i-vector, effectively improved the prediction accuracy of Big Five personality trait estimation. In addition, the experimental results showed that audio-related features were the most prominent features in both corpora.

Keywords Big five personality traits · Group discussion · Multimodal · Speaker individuality · i-vector

1 Introduction

The aspects of nonverbal communication have become important focuses in human–computer interaction (HCI) studies. This is because nonverbal aspects are naturally delivered in human-to-human communication. When we interact with other people, we consider not only what they are saying but also how they are speaking. If nonverbal aspects were not considered, communication would become very unnatural or robot-like.

✉ Candy Olivia Mawalim
candylim@jaist.ac.jp

Shogo Okada
okada-s@jaist.ac.jp

Yukiko I. Nakano
y.nakano@st.seikei.ac.jp

Masashi Unoki
unoki@jaist.ac.jp

[1] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi-shi, Ishikawa 923-1292, Japan

[2] Dept. of Computer and Information Science, Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, Tokyo 180-8633, Japan

The study of of the nonverbal aspects, e.g., personality, has attracted much attention in HCI. Personality extensively influences human life, in areas such as decision making, preferences, and reactions. It comprises the patterns of the habitual behaviors, emotions, and cognition of a person [34]. We could achieve a better understanding of ourselves and other people around us by understanding personality.

The integration between personality science and HCI studies has been emerging since the mid-2000s, and thus, the term personality computing (PC) was established as a research field [33,53]. Vinciarelli and Mohammadi [53] argued that three phenomena fuel PC from a technological perspective: (1) the availability of personal information in social networks, (2) the possibility of data collection via mobile technology on a daily basis, (3) the consideration of social and affective intelligence in the computing and machinery research. Subsequently, three major problems are addressed in PC, i.e., automatic personality recognition, perception, and synthesis tasks [53]. The features for these tasks are extracted from personality-expressive signals, such as behavioral modalities from various data sources [33].

The most popular and influential personality taxonomy is the Big Five personality trait system [31,34]. Since it is relatively stable in time as well as applicable across vari-

ous cultures and trait measures, the Big Five personality trait system is accepted in a wide range of areas, including in PC [33,34,53]. This measurement classification system comprises five traits:

1. Openness to experience (O): the degree of being curious and inventive;
2. Conscientiousness (C): the degree of being efficient and organized;
3. Extraversion (E): the degree of being energetic, active, and outgoing;
4. Agreeableness (Ag): the degree of being cooperative and compassionate;
5. Neuroticism (N): the degree of being sensitive and nervous.

In an earlier study, manual assessment was conducted by using a standardized factor analysis of personality description questionnaires to determine one's Big Five personality traits. However, this type of manual assessment is very costly and thus not applicable for HCI interfaces. Accordingly, automatic personality trait assessment studies have attracted great attention in recent years.

Several techniques have been proposed from the perspectives of various modalities for automatic personality trait estimation. For instance, personality detection studies based on facial expression analysis were reviewed in [17] using image processing techniques. Concurrently, studies on speech personality trait recognition have also progressed in the speech research community, especially since the Interspeech 2012 Speaker Trait Challenge was released [42]. Other approaches using language models have also been widely employed to estimate personality traits, such as those derived from conversations through social media [10,56]. Instead of focusing on one modality, several studies have also used multimodal analysis to infer personality traits [9,22,25,29].

Despite the growth in the number of automatic personality detection studies, the reliability of detection performance is still far from ideal. Most of the existing studies focused on inferring individually perceived personality traits in self-presentation scenarios [5,42], which is not ideal for representing personality. McCrae and Costa (1996) reported that personality shows the basic tendencies of a person, particularly in dealing with social interactions [31]. Manifesting personality traits in interactions is more meaningful than self-presentation.

In recent years, several studies have considered the automatic inference of personality traits from interaction processes, such as small group interactions [19,22,25,29]. Okada et al. [29] proposed a personality trait estimation method based on a co-occurrent multimodal event discovery approach using the audio-visual (AV) subset of a group meeting from the Emergent Leadership (ELEA) corpus (ELEA-AV). Subsequently, the study of Kindiroglu et al. [19] demonstrated a multidomain and multitask approach for predicting the extraversion and leadership traits in the ELEA corpus. Additionally, prior work in [25] focused on personality trait estimation by using multimodal features and communication skills indices for datasets with multiple discussion types.

Many transformer-based methods and various types of multimodal fusion techniques have been proposed for solving various computing tasks [21]. Most of the methods required a large-scale dataset which is difficult to fulfill for analyzing a social interaction, such as the main task addressed in this study. With a relatively smaller size of data, we focus on how to handle individuality features and how to mitigate the issue of individual differences in more diverse group discussion corpora (different language and environment settings).

This study aims to address two novel points. First, we investigate the relationships between the state-of-the-art speaker individuality feature extracted from speech, namely, the identity vector (i-vector), and the Big Five personality traits. Our hypothesis is that speaker individuality is interrelated with personality. Second, we investigate the effectiveness of multimodal features regardless of the selected language. In this study, we consider two group discussion datasets, including the Multimodal Task-Oriented Group Discussion (MATRICS) corpus (in Japanese) and the ELEA-AV corpus (in European languages), to infer the Big Five personality traits. By the end of this paper, we will discuss the following key questions:

1. Is the speaker individuality feature effective for inferring the Big Five personality traits?
2. What are the effective multimodal features for estimating the Big Five personality traits for the MATRICS and ELEA-AV corpora?

The rest of this paper is organized as follows. Section 2 describes works that are closely related to this study. In Sect.3, we introduce the utilized multimodal corpora. Subsequently, we describe the employed feature representation approach in Sect. 4. The experimental settings and results are summarized in Sect. 5. In Sect. 6, we discuss the results and answer the key questions in this study. Finally, this paper is concluded in Sect. 7.

2 Related work

Automatic personality computing is useful for many HCI applications because it can model the relationships between stimuli and the outcomes of social perception processes. In other words, an automatic personality computing method

Table 1 Dataset descriptions

Dataset	Description	
MATRICS	*Participants*	
	#number	40 (29 male, 11 female)
	Occupation	University student
	Recorded data	
	#sessions	30
	duration	∼9 h
	#utterances	20,339
	#participants per session	4
	#task type(s)	3 (in-basket, case study with prior, case study without prior)
	#recordings	120 (30 sessions × 4 participants)
	Available data	Audio, language, visual and motion (head, eye, and body movements), CS indices,actual and perceived Big Five traits
ELEA-AV	*Participants*	
	#number	102
	occupation	
	Recorded data	
	#sessions	27
	Duration	
	#utterances	
	#participants per session	Mixed (3 ∼ 4 people)
	#task type(s)	1 (winter survival task)
	#recordings	112 (27 sessions × 3 ∼ 4 participants)
	Available data	Audio, motion (head, eye, and body movements), leadership and dominance indices, actual and perceived Big Five traits

models or estimates how our responses or impressions towards others are based on every observable action performed by the subject. Efforts on personality trait analysis with the consideration of multimodality are relatively extensive. For instance, a study on personality trait recognition in social interactions using audio and visual features was conducted by Pianesi et al. [35]. Their study aimed to automatically predict personality traits obtained from self-reported questionnaires. In [1], Aran et al. investigated video blogs in a small group meeting to predict personality traits, especially extraversion traits. Another work from Jayagopi and Gatica-Perez [18] attempted to propose a solution for predicting group performance and personality traits by mining typical behavioral patterns. Subsequently, a mining approach for extracting co-occurrent events from multimodal time-series data for personality estimation was also proposed by Okada et al. [29]. Batrinca et al. [6] conducted a comparative analysis to observe the difference between the personality trait recognition accuracies obtained for a human-machine interaction (HMI) scenario and a human-human interaction (HHI) scenario.

In addition to the studies mentioned above, several studies specifically focused on improving Big Five personality trait prediction. For instance, Fang et al. [16] used three nonverbal features, including intrapersonal features, dyadic features, and one-vs.-all features, to predict the Big Five model. Lin et al. [22] developed a Big Five predictor based on the use of an interaction mechanism in bidirectional long short-term memory (BLSTM) to model the vocal behaviors of participants. In the prior study [25], communication skills and task types were considered for estimating the Big Five personality traits.

Our current work differs significantly from the existing studies in terms of the utilized features and dataset dependency. In most studies, low-level features were extracted for the estimation process. We consider the transfer learning technique by extracting higher-level features using state-of-the-art pretrained speaker embedding models (i-vector and x-vector extractors [13,47]). To ensure the effectiveness of our proposed system regardless of the selected language, we use two different language corpora, i.e., a European language corpus and a Japanese corpus (Table 1).

3 Multimodal data corpora

In this study, we utilized two multimodal data corpora, i.e., the MATRICS corpus and ELEA-AV corpus. Figure 1

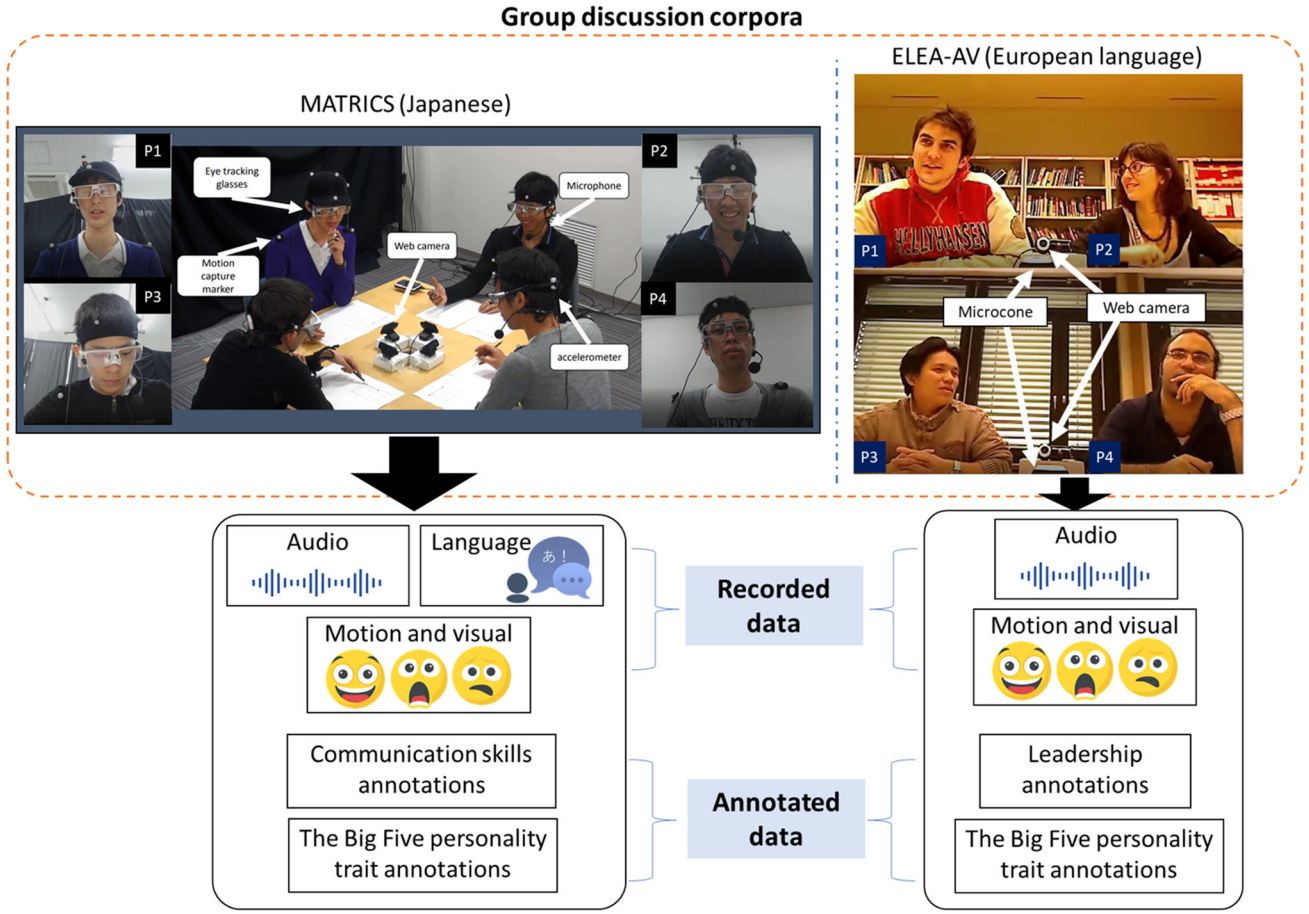

Fig. 1 Overview of the utilized multimodal group discussion corpora

presents an overview of these corpora. The MATRICS corpus was used as the main dataset for analyzing the effectiveness of each modality. In addition, the ELEA-AV corpus was used to analyze the speaker individuality features as audio-related features despite the different nature of this dataset.

3.1 MATRICS corpus

The MATRICS corpus is a Japanese group discussion dataset introduced in [28]. Forty participants were involved in ten uniformly distributed discussion groups (four participants in each discussion group). The MATRICS corpus consists of multimodal raw data, i.e., audio data, video data, and head motion data. In addition, reliable manual transcriptions and assessments of the Big Five personality traits and communication skills are also available. The audio data were recorded via an Audio-Technica HYP-190H hands-free head-worn microphone. In contrast, the video data were recorded using two SONY HDR-CX630V cameras that captured two opposite angles of the group interaction overview. The head motion data were recorded by ATR-Promotions WAA-010 accelerometers.

The assessment of Big Five personality trait scores in the MATRICS corpus was obtained from a survey, while the communication skills were annotated by 21 human resource management experts using the recorded video data. The communication skills annotations presented in [30] contained five different indices, including listening attitude (LA), smooth interaction (SI), aggregation opinions (AO), communicating one's own claim (CC), and logical and clear presentation (LP). The overall total score was also calculated as the total communication (TC) index. Each annotator assessed all the communication skills indices of each participant from the given segmented video sessions. The reliability of the assessment was confirmed by the level of agreement among the annotators with Cronbach's alpha (α) and the Pearson correlation coefficient (ρ), except for LA (with $\alpha < 0.85$ and $\rho = 0.59$).

Unlike the other group discussion datasets with only one discussion task available per group, such as the ELEA corpus [38], the MATRICS corpus consists of three different tasks for each discussion group. These tasks are distinct in terms of freedom and the scope of the given prior information regarding the conversation structure. The freedom levels of task-1,

task-2, and task-3 are ordered from low to high, whereas the amount of given preliminary information is ordered from more to less. The details of the discussion topic for each task are described as follows:

1. task-1 (in-basket): the selection of an invited guest for a school festival;
2. task-2 (a case study with prior information): preparation of a food and beverage booth at a school festival;
3. task-3 (a case study without prior information): arrangement of a two-day travel itinerary in Japan for a foreign friend.

3.2 ELEA-AV corpus

In addition to using the MATRICS corpus, we used the AV subset from the ELEA corpus [38] to check the effectiveness of speaker individuality features. This subset includes recordings from 27 group meetings with 102 participants. Each recording has a length of 15 minutes. The task in the ELEA corpus is known as a winter survival task. In this task, the participants had to order 12 different items to bring with them as if they were the survivors of an airplane crash that occurred in winter.

This corpus originally aimed to analyze emergent leadership in group discussions. Nevertheless, this corpus also provided both self-assessed and perceived Big Five personality trait scores for each participant. Therefore, the Big Five estimation model could be constructed using this corpus. We aimed to verify whether speaker individuality features, as audio-related features, could be practical in more general cases (regardless of the different characteristics of the MATRICS and ELEA-AV corpora).

4 Feature representation

In this study, we extracted three modality groups (i.e., audio, language, and motion & visual groups) and communication skills indices as the inputs for Big Five estimation. Table 2 shows a summary of the multimodal features.

4.1 Audio-related features

In prior work [25], audio-related features were obtained by OpenSMILE [15], which was configured for perceived speaker traits in the Interspeech 2012 Speaker Trait Challenge proposed by Schuller et al. [42]. Unlike prior work, we aimed to thoroughly analyze the effectiveness of audio-related features specifically for Big Five personality trait estimation in group discussions. Accordingly, five categories of audio-related features were extracted in this study, including speaker identity features, spectral-related features, voice-related features, energy-related features, and turn-taking features.

Speaker identity features—We aimed to investigate whether the features related to speaker identity could contribute to the performance of an automatic Big Five personality trait estimator. Accordingly, we extracted the i-vector and x-vector features in this study. The i-vector subspace modeling approach introduced by Dehak and Shum [13] has become the state-of-the-art technology in speaker recognition systems. In the i-vector approach for speaker recognition [12,13], a low-dimensional vector that is extracted using joint factor analysis (JFA) represents a speech segment. This approach has been reported to reduce high-dimensional sequential speech data to a lower-dimensional fixed-length vector representation that contains more relevant information. Figure 2 shows the simplified block diagram of the i-vector extraction process.

In the former i-vector modeling approach, the assumption of a Gaussian feature distribution was made; however, this is not always applicable in practice. Thus, a DNN model was developed to address this issue [45]. Subsequently, to improve the robustness of the i-vector obtained with the DNN model, the process of obtaining an i-vector from a DNN with embedding layers was proposed by Snyder et al. [46,47]. This i-vector is also known as an x-vector [47]. The architecture of the x-vector extractor is shown in Fig. 3. We utilized the pretrained VoxCeleb [27] i-vector and x-vector models provided by David Snyder that are available in the Kaldi toolkit [37,47]. These pretrained models were constructed using Mel-frequency cepstral coefficients (MFCCs) as their input features.

Before extracting an i-vector or x-vector using the pretrained models, we selected the "long" utterances (utterances with lengths of more than 3 s) of each speaker in a session (one instance). The speaker individuality vector for an instance was then defined as the average of the individuality vectors derived from all "long" utterances. This preprocessing step was conducted to assure the reliability of the extracted vector. Figure 4 shows the PCs of the x-vectors extracted from five speakers (MATRICS corpus) in three-dimensional space.

Spectral-related features—MFCCs are widely used as standard features in speech processing domains, including emotion and speaker trait recognition [40–42]. MFCCs represent the spectral envelope of a signal (timbral information) [50] and were reported to have the ability to separate the impacts of the source and filter of the input speech. An MFCC can be obtained by mapping the Fourier power spectrum of a signal onto the Mel scale [48]. Subsequently, the discrete cosine transform was performed for the Mel log powers was performed and resulted in the Mel spectrum, in which the

Table 2 Summary of the multimodal features used for Big Five trait estimation

Modality	Feature	Variables
Audio (A)	i-vector	400-dimensional vector
	x-vector	512-dimensional vector
	MFCC	MFCC with its delta and delta-delta
	LPC	Mean of 10th-order LP coefficients
		Deviation of 10th-order LP coefficients
		Range of 10th-order LP coefficients
	LSP	Mean of LSPs obtained from 10th-order LPs
		Deviation of LSPs obtained from 10th-order LPs
		Range of LSPs obtained from 10th-order LPs
	F0	Mean of F0 trajectory
		Deviation of F0 trajectory
		Range of F0 trajectory
		Minimum value of F0 trajectory
		Maximum value of F0 trajectory
	PI	Mean of sound energy
		Deviation of sound energy
		Range of sound energy
		Minimum value of sound energy
		Maximum value of sound energy
	ST	Total speaking length
		Total count of utterances
		Average length of utterances
Language (L)	PoS	Bag of PoS tags, including nouns, verbs, new nouns, interjections, and fillers)
	DT	12 dialog act tags
		3 speech act tags
		2 semantic tags
Motion and Visual (M)	HM	Mean of movement
		Deviation of movement
		Mean of movement while speaking
		Deviation of movement while speaking
		Difference of movement while speaking
	AU	Mean of action units
		Deviation of action units
	PS	Mean of pose movement
		Deviation of pose movement
		Range of pose movement
	GZ	Mean of gaze movement
		Deviation of gaze movement
		Range of gaze movement
Communication	CS	6 CS indices (LA, SI, AO, CC, LP, and TC)

amplitude refers to the corresponding MFCC. Figure 5 shows the block diagram of deriving the MFCC of an input signal.

In addition to MFCCs, the first- and second-order frame-based MFCCs (delta and delta-delta, respectively) are also considered prominent features in several applications. The following equation shows the mathematical expression of a delta coefficient (d_t) for a frame t given that the coefficients (c_{t+n} and c_{t-n}) with have typical N values of 2.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (1)$$

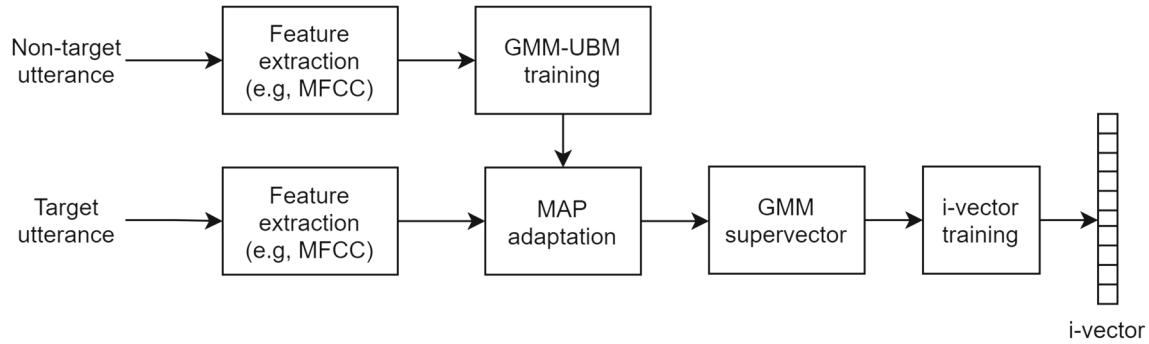

Fig. 2 Simplified block diagram of the i-vector extraction process

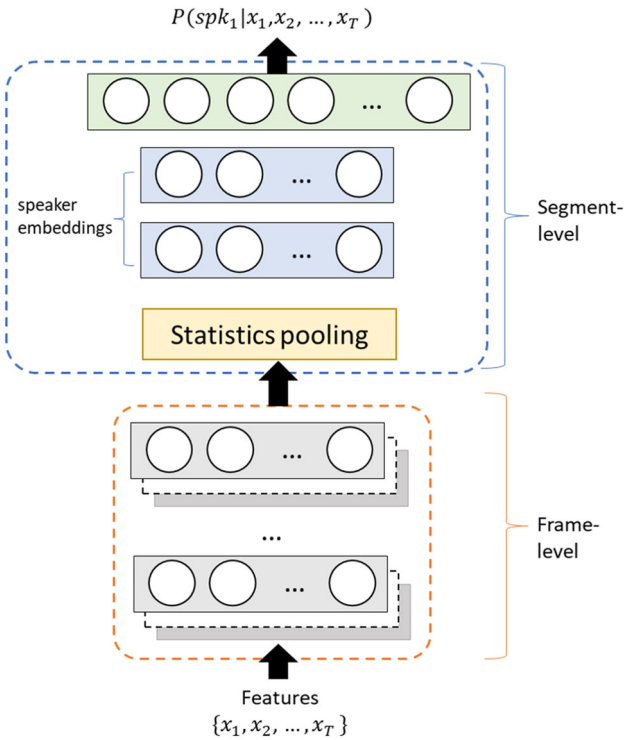

Fig. 3 A deep neural network (DNN) with an embedding layer architecture as an x-vector extractor [47]

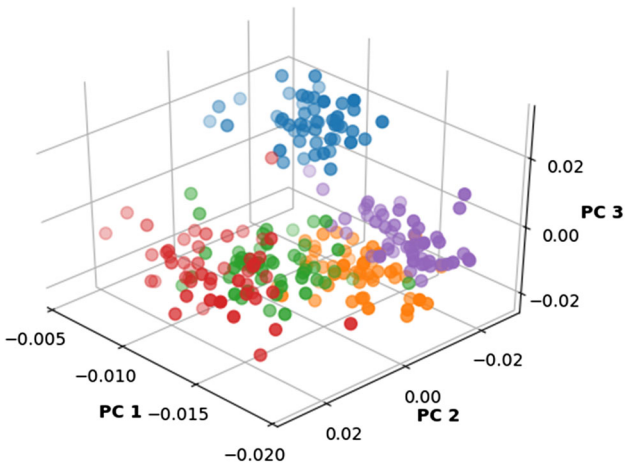

Fig. 4 Three-dimensional principal components (PCs) of the x-vectors of five speakers extracted from fifty speech utterances in the MATRICS corpus. The colors represent the speaker identity labels

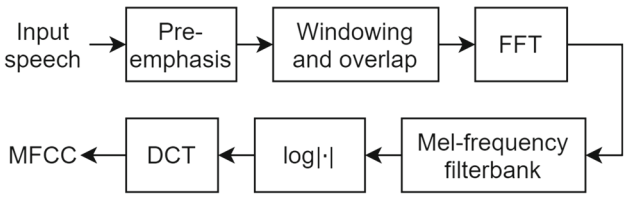

Fig. 5 Block diagram of MFCC derivation

In this study, we extracted MFCC features with delta and delta-delta using a speech processing toolkit (SPTK [52]) to infer Big Five personality traits. In general, it was suggested that the first 8-13 MFCCs represented the shape of the spectrum. Furthermore, the higher-order coefficients were related to the finer spectral details, such as pitch and tone. However, using a large number of cepstral coefficients results in more analytical complexity. Therefore, the first 12 to 20 MFCCs are typically used for optimal speech analysis [26]. We used the first 12 coefficients and both delta and delta-delta as the spectral-related features.

Voice-related features—We extracted the statistical properties of the fundamental frequency (F0), linear predictive coefficients (LPCs), and line spectral pairs (LSPs) by SPTK as voice-related features. Before extracting these features, we conducted preprocessing on the raw audio data via the selection of "long" utterances (more than 3 s), downsampling to 16 kHz, and framing with a 30-ms length and a 50% overlap. This preprocessing step was conducted to capture better information related to voiced speech.

The F0 trajectory estimation was acquired using the robust algorithm for pitch tracking (RAPT) [49] in SPTK. The LPC and LSP features were obtained using tenth-order linear predictive coding, which is commonly used for mimicking speech production systems [3]. The LPCs and LSPs were

useful for estimating speech formants. For this reason, we extracted these features only from the "long" voiced utterances.

Energy-related features—This feature set was derived from sound energy (further named PI). The sound energy was represented by statistical properties calculated in the frame-based unit.

Turn-taking features—This feature set is represented by three speaking turn (ST) feature variables for participants: (i) the total speaking length (the total duration of the speaking utterances in a session), (ii) the total utterance count (the number of utterances in a session), and (iii) the average utterance length (the total utterance count in a session divided by the total duration).

4.2 Language-related features

We utilized two language-related feature sets, i.e., a bag of part-of-speech tags (PoS) and dialog tags (DTs). The PoS feature set was extracted via manual transcription using MeCab [20], a Japanese morphological analysis toolkit. On the other hand, the DT feature set was obtained by the method introduced in [30]. This feature set consisted of 12 dialog act tags ("conversational opening", "open question", "suggestion", "backchannel", "open opinion", "partial acceptance", "acceptance", "rejection", "understanding check", "other question", "WH-question", and "y/n question") from Dialog Act Markup in Several Layers (DAMSL) [11] and Meeting Recorder Dialog Act (MRDA) [43], three speech act tags ("plan", "agreement", and "disagreement"), and two semantic tags ("fact description" and "reason").

4.3 Motion and visual features

In the MATRICS corpus, the motion and visual features can be categorized into two groups. The first group includes the features obtained from the head movements recorded by accelerometers. The statistical properties of the head movements were calculated (as shown in Table 2) [30]. Head movement refers to the norm of the three-dimensional head acceleration ($|a_t|$) at a particular time t (where $a_t = \{x_t, y_t, z_t\}$). The movements performed while speaking were calculated by joining the head activity data with the speaking time data via manual transcription for each participant. This feature set was also normalized using z-score normalization. We further referred to this feature set as head motion (HM).

The second group includes the face-related features extracted by using OpenFace [4], the state-of-the-art facial behavior analysis toolkit. We extracted action units (AUs), head pose (PSs), and eye gazes (GZs) by inputting the raw video data that captured the face of each participant while having a discussion. Figure 6 shows the example of the face-related features extraction by OpenFace. AUs are significantly related to human emotions as paralinguistic information [23,51]. A PS captures the position and rotation of a head in three-dimensional space (X, Y, Z, R_x, R_y, R_z). This feature set was reported as a prominent cue in social event analysis [54]. Last, GZs show the eye movements that contribute to social and emotional communications, especially for tracking the attention directions of participants [14,54,55]. In this study, we extracted GZs using the facial landmark detection model [57].

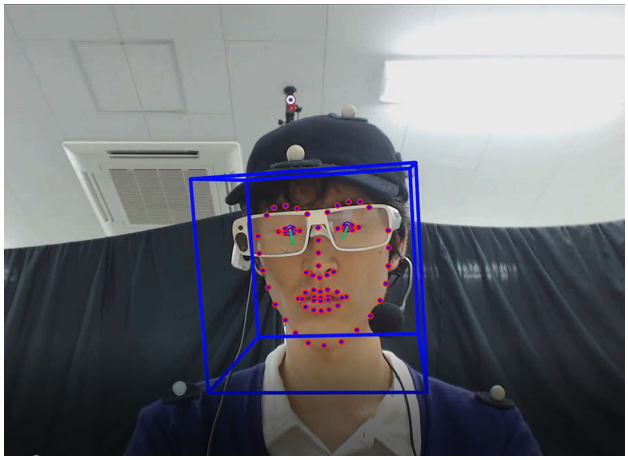

Fig. 6 Example of face-related feature extraction by using OpenFace from one of the video clips of the MATRICS corpus. Blue bounding box shows the head pose. Red points show the facial points. Green lines show the eye gazes

In the ELEA-AV corpus, there are three groups of motion- and visual-related features. The first group is referred to as visual activity features, which capture body activity (bMotion) and head activity (hMotion) features. These features were extracted by the body tracking, head tracking and optical flow [29]. The second group is based on motion energy images (MEIs) [7]. MEIs were obtained by integrating different images of the whole recorded clip. Since the MEIs changed on a time-series basis, the segmentation of time-series MEI data according to categorical patterns followed the procedure described in [29]. The third group of motion- and visual-related features in the ELEA-AV corpus is the visual focus of attention (VFOA). These features employed a probabilistic framework to estimate head locations and poses on the basis of a state-space formulation [39]. The VFOA features that we employed followed those utilized in [29].

4.4 Communication skills and leadership indices

As mentioned in Sect. 3, the communication skills (CS) indices in the MATRICS corpus were obtained by manual assessment from 21 experts in human resource management. Subsequently, the leadership (Ld) indices included

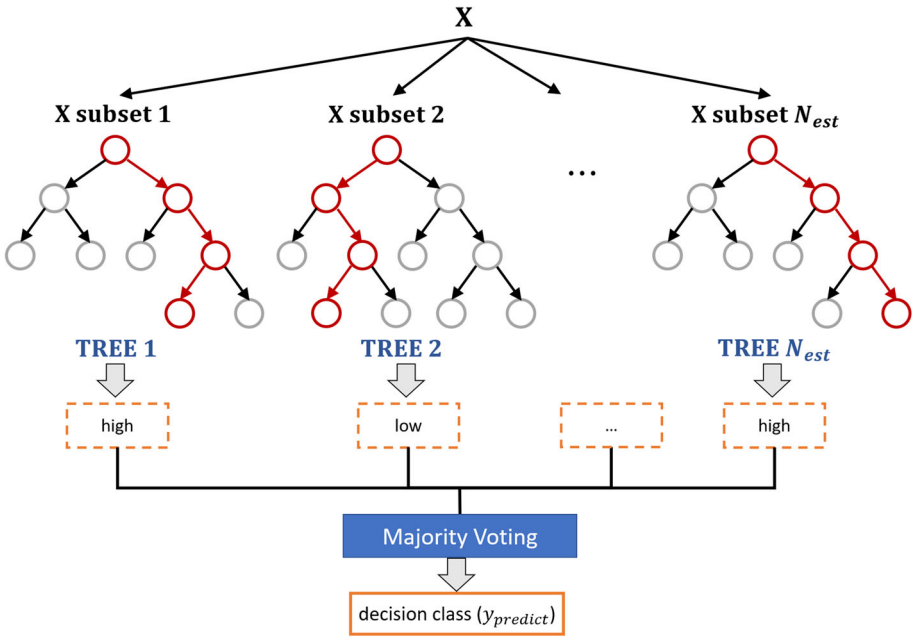

Fig. 7 Block diagram of our experimental process

in the ELEA-AV corpus were related to individual impressions about dominance and leadership. These indices were determined by other participants in the meeting as perceived interaction scores. Five Ld items were included: perceived leadership, perceived dominance, perceived competence, perceived liking, and dominance ranking. More details on the CS and Ld indices are described in [28,38], respectively. As a preprocessing step for these features, we applied z-score normalization to both the CS and Ld indices.

5 Experiment

In our preliminary study [25], one of the objectives was to clarify the effectiveness of verbal and nonverbal features and CS indices for estimating the Big Five personality traits. We extracted audio-related features in the same manner as the baseline system in the Interspeech 2012 Speaker Trait Challenge [42] designed for estimating perceived speaker traits from single speaker utterances. In contrast, we aimed to thoroughly study which audio-related features are more suitable for estimating the self-assessed speaker traits of each participant in a group discussion, as provided in the MATRICS corpus. Self-assessed speaker traits are more robust than perceived speaker traits, regardless of the speech content and environment. Since the sizes of the group discussion corpora are relatively limited, we also considered performing transfer learning by using the state-of-the-art speaker individuality features for estimating the Big Five personality traits. Figure 7 shows the main ideas of our experimental process.

The experiment in the current study aimed to investigate the effectiveness of (1) speaker individuality features (i-vector and x-vector) (Sect. 5.2.1); (2) nonverbal behaviors, e.g., face gestures (Sect. 5.2.2); and (3) a combination of modality groups (Sect. 5.2.3) for Big Five personality trait estimation in both the MATRICS and ELEA-AV corpora. Accordingly, we conducted unimodal analysis followed by multimodal analysis by considering each modality group. An ablation test was also conducted to study the importance of each modality group. In this study, the experiment was conducted as a binary classification task (similar to [2,29]). The input was the combination of features explained in Sect. 4, and the targets were the Big Five personality trait scores, i.e., neuroticism (N), extraversion (E), openness (O), agreeableness (Ag), and conscientiousness (C). As mentioned above, the Big Five scores were obtained from a self-assessed questionnaire, which is usually more accurate but more difficult to predict than the perceived Big Five scores used in prior studies [22,29].

5.1 Experimental settings

In the prior study [25], the support vector machine (SVM), random forest, Naïve Bayes, and decision tree algorithms were investigated for predicting the Big Five personality traits in the MATRICS corpus. The results showed that the random forest classifier could obtain the most reliable estimation accuracy for most traits and, therefore, suitable to generalize a prediction model. A random forest is an ensemble learning algorithm that generates a set of decision trees from the given data samples, randomly selects its subsets, and chooses the best solution among the subset predictions by voting. This algorithm can reduce overfitting issues and result in a robust and high-performance model [8]. Figure 8 shows

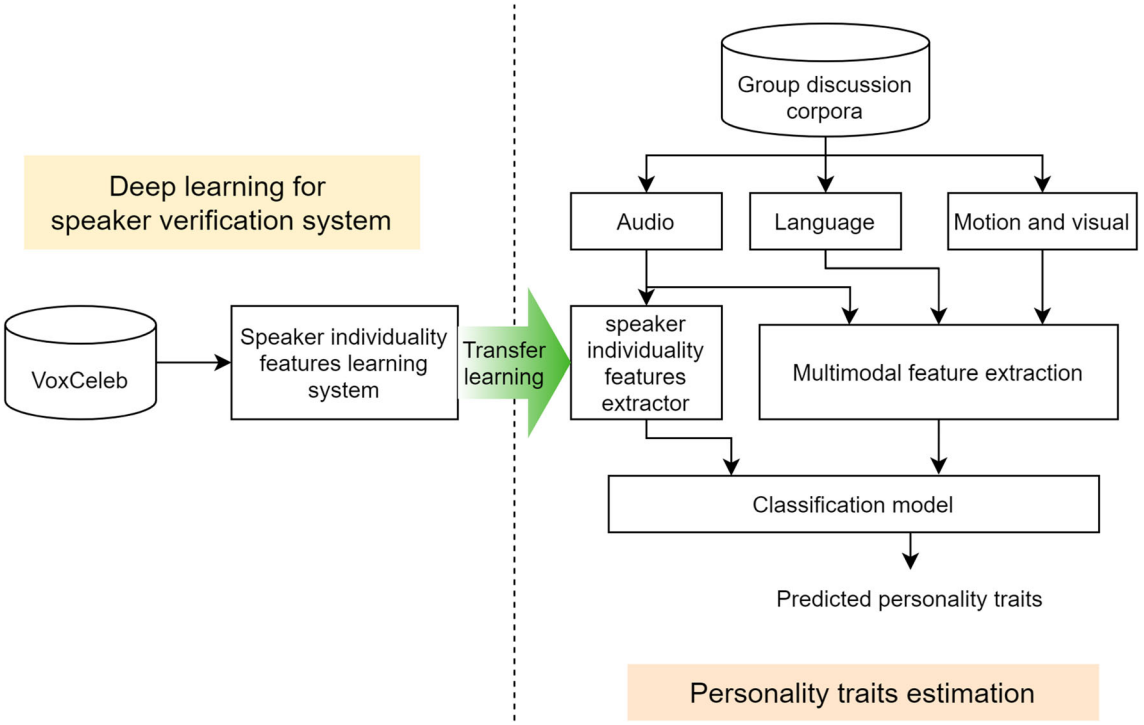

Fig. 8 Illustration of the random forest algorithm

Table 3 The combinations of modality groups used for multimodal analysis

Modality	Features	
	MATRICS	ELEA-AV
Unimodal	A	A
	L	M
	M	Ld
	CS	
Bimodal	A+L	A+M
	A+M	A+Ld
	A+CS	M+Ld
	L+M	
	L+CS	
	M+CS	
Multimodal	A+L+M	A+M+Ld
	A+L+CS	
	A+M+CS	
	L+M+CS	
	A+L+M+CS	

an illustration of the random forest algorithm. We utilized the random forest algorithm in the ensemble module from scikit-learn [32] to build our classification model. Parameter tuning was applied for the number of estimators (N_{est}) and the maximum depth.

Fig. 9 Heatmap of the i-vector F1-score matrix for feature selection when estimating the Big Five personality traits

To achieve our goals, we conducted a comparative analysis on the basis of the obtained feature set. The feature set for unimodal analysis is shown in Table 2. Additionally, an ablation test was conducted with respect to the modality groups for multimodal analysis. Four modality groups were involved, including the audio-related modality (A), language-related modality (L), motion- and visual-related modality (M), and communication-related modality (C). The combinations of these modality groups for multimodal analysis are listed in Table 3.

The feature selection procedure was conducted for each feature set, where the number of selected features was based on the best overall unimodal analysis result with default classifier parameters (no parameter tuning). This feature selection process was conducted only for feature sets with more than ten elements. A support vector regressor (SVR)

was used by fitting the training features and training outputs of this feature selection process. Figure 9 shows an example of i-vector feature selection analysis using several elements (ranging from $\{N_i/8, N_i/4, N_i/2, N_i\}$, where N_i is the number of i-vector dimensions (400)). Although a larger number of elements resulted in better accuracy for neuroticism, the estimates for other traits worsened. Therefore, we selected 100 as the number of features for the i-vector to compensate for the estimation of the other traits. Subsequently, to reduce the probability of imbalance issues, we also conducted late fusion for each modality group before merging it with the other modalities. The number of selected features from each modality group (except the CS and Ld groups) was uniform and selected from {5, 10, 20, 30}.

Following a previous study, [30], the utilized MATRICS corpus consisted of 107 out of 120 data samples due to some missing values recorded from accelerator data. Furthermore, for the ELEA-AV corpus, we used all 102 existing data samples. From the available data samples, we conducted leave-one-person-out cross-validation (LOPCV). As participant data were set as the testing data, the other participants' data were set as the training data. Thirty-fold cross-validation was carried out because there were 30 participants (3 people in each of the 10 discussion groups) in total for the MATRICS corpus. To evaluate the performance of the binary classification model, we used the F1-score metric, which considers the balance between the precision and recall of the estimation results.

5.2 Results

This subsection presents the results of our experiments, including those obtained from (1) unimodal and multimodal analyses for both the MATRICS and the ELEA-AV corpora and (2) a comparison with prior works [2,25,29].

To investigate the effectiveness of each feature set, we carried out a unimodal analysis to estimate the Big Five personality traits. After obtaining the most effective feature sets for each modality, we carried out a multimodal analysis. Tables 4 and 5 show the unimodal analysis and multimodal analysis results regarding the inference of the Big Five personality traits in the MATRICS corpus, respectively. In the same way, we also conducted unimodal and multimodal analyses to infer the Big Five personality traits in the ELEA-AV corpus. Tables 6 and 7 show the results of the unimodal analysis and multimodal analysis, respectively, for the ELEA-AV corpus.

5.2.1 Speaker individuality features for personality estimation

The Big Five personality trait estimation results in the MATRICS corpus using speaker individuality features (i-vector or x-vector) are shown in the first and second rows of Table 4. From this table, using speaker individuality features could effectively improve neuroticism trait estimation (F1-score $>$ 70%). The x-vector is also useful for estimating the extraversion trait (F1-score $>$ 65%). The comparison between the fusion of modality A (audio-related features) and A' (audio-related features without speaker individuality features) in Table 5 shows how speaker individuality affects the

Table 4 Big Five personality traits estimation results obtained for MATRICS corpus using single feature with LOPCV

Modality	Feature Set	F1-score (%)				
		N	E	O	Ag	C
Audio (A)	i-vector	**70.97**	58.13	53.05	55.13	48.46
	x-vector	**70.97**	67.18	45.54	56.02	58.81
	MFCC	59.79	53.69	47.62	**68.32**	52.30
	LPC	53.64	58.13	58.68	61.52	50.40
	LSP	50.95	60.14	63.54	62.05	54.25
	ST	55.05	50.25	63.49	56.35	63.46
	F0	52.36	**67.27**	57.91	54.27	56.95
	PI	68.21	58.86	63.56	54.33	69.09
Language (L)	PoS	66.11	52.19	61.68	58.63	57.91
	DT	51.34	50.18	57.01	53.27	69.18
Motion and visual (M)	HM	67.73	56.09	54.04	53.27	**69.89**
	AU	50.49	58.13	57.95	55.49	68.26
	PS	52.35	53.30	57.93	55.14	62.18
	GZ	54.94	53.27	**65.97**	59.64	56.13
Communication skills (C)	CS	62.45	59.76	55.87	60.89	61.66

Blue cells with bold captions represent the best prediction results

Table 5 Big Five personality trait estimation results obtained for the MATRICS corpus using a multimodal feature set with LOPCV

Modality	F1-score (%)				
	N	E	O	Ag	C
A+L+M+CS	62.62	67.56	57.91	56.41	60.71
A'+L+M+CS	64.49	44.89	53.20	49.84	47.30
A+L+M	61.70	67.56	63.55	49.52	57.01
A'+L+M	**65.44**	50.44	53.23	49.84	53.20
A+L+CS	53.25	63.66	54.17	51.90	63.48
A'+L+CS	56.92	45.79	39.90	48.65	51.49
A+M+CS	63.77	65.35	60.75	51.61	64.97
A'+M+CS	64.44	53.30	53.28	50.66	55.84
L+M+CS	44.19	52.24	65.19	50.55	59.60
A+L	58.17	**69.14**	53.28	55.03	58.88
A'+L	59.83	43.93	32.72	50.52	53.81
A+M	61.89	69.09	58.89	50.69	57.27
A'+M	64.50	55.16	43.58	52.28	58.88
A+CS	56.92	63.56	55.12	53.46	56.84
A'+CS	58.83	48.60	37.36	52.46	54.97
L+M	48.00	57.76	**68.20**	51.18	62.62
L+CS	38.57	47.68	64.92	49.13	65.81
M+CS	40.19	51.80	65.97	53.48	60.97
A	59.16	60.70	40.26	55.89	57.91
L	53.69	44.80	56.17	54.01	**66.60**
M	51.42	50.95	60.12	50.66	63.57
CS	62.45	59.76	55.87	**60.89**	61.66

A' denotes the low-level audio-related features (A without speaker identity features). Blue cells with bold captions represent the best prediction results

Big Five personality estimation in multimodal analysis. For most of the traits (except neuroticism traits), using speaker individuality features could improve the estimation results.

Similarly, we could see the Big Five personality trait estimation results in the ELEA-AV corpus using speaker individuality features in Table 6. Almost all of the personality trait estimations could achieve an F1-score of more than 60% (except conscientiousness trait). The best estimation using the x-vector could be achieved for the openness trait. When fusing with other modalities (as shown in Table 7), a noticeable improvement is shown in the estimation of openness and agreeableness traits. For instance, the estimation result using all modalities, including the x-vector, could achieve an approximately 8% higher F1-score than the one excluding the x-vector.

5.2.2 Nonverbal behaviors as features for personality estimation

We analyzed nonverbal behaviors, i.e., motion- and visual-related features, CS indices, and Ld indices, for Big Five personality trait estimation. The nonverbal features available in the MATRICS corpus are HMs, AUs, PSs, GZs, and CS. From Table 4, the best results for the openness and conscientiousness traits were achieved by the GZs and HMs, respectively. The nonverbal features available in the ELEA-AV corpus are bMotion, hMotion, MEIs, VFOA, and Ld. The highest F1-scores obtained during the single feature set analysis Table 6 were mostly achieved using nonverbal features, except for the openness trait. The estimation trait was best predicted by the Ld feature. The VFOA feature was best for predicting the agreeableness trait. In addition, the most effective feature set for the neuroticism and conscientiousness traits was the set of MEIs.

5.2.3 Multimodal features for personality estimation

On the basis of the unimodal analysis results, we used the prospective feature sets as one modality group. For instance, the feature sets for A included the x-vector, MFCC, F0, PI, and LSP. For the MATRICS corpus. Four modality groups were considered in this multimodal analysis. An ablation test was carried out to check the significance of each modality. Table 5 shows the results of the ablation test. These

Table 6 Big Five personality trait estimation results obtained for the ELEA-AV corpus using single-feature sets with LOPCV

Modality	Feature Set	F1-score (%)				
		N	E	O	Ag	C
Audio (A)	i-vector	61.33	66.64	62.78	61.97	59.42
	x-vector	61.75	64.77	**64.43**	64.61	49.26
	audio	58.78	60.83	55.61	57.81	57.44
	energy	50.98	59.20	58.79	61.82	57.03
	pitch	62.75	68.63	59.79	55.60	58.92
	MFCC	62.63	56.32	59.81	59.67	56.96
	LPC	58.78	60.08	55.87	64.26	56.96
	LSP	63.57	60.83	58.73	60.72	60.88
Leadership (Ld)	eLead	56.88	**69.63**	55.89	54.95	61.23
Motion and visual (M)	bMotion	50.53	56.91	55.27	62.44	60.50
	hMotion	55.86	52.55	55.82	61.73	59.63
	MEI	**64.74**	57.90	59.79	61.19	**62.92**
	VFOA	62.70	59.86	59.42	**65.47**	50.82

Blue cells with bold captions represent the best prediction results

Table 7 Big Five personality trait estimation results obtained for the ELEA-AV corpus using multimodal feature sets with LOPCV

Modality	F1-score (%)				
	N	E	O	Ag	C
A+M+Ld	59.82	57.41	**56.80**	64.71	49.28
A'+M+Ld	63.57	62.75	48.98	62.28	54.64
A+M	63.69	56.66	50.94	64.26	50.56
A'+M	63.14	58.74	49.83	62.42	52.50
A+Ld	64.78	58.82	54.55	63.50	43.17
A'+Ld	62.53	58.63	50.94	59.67	55.11
Ld+M	61.88	47.71	40.72	61.52	55.11
A	**67.51**	52.85	53.91	**65.58**	49.31
M	65.54	51.04	43.78	65.17	55.11
Ld	56.88	**69.63**	55.89	54.95	**61.23**

Blue cells with bold captions represent the best prediction results

results demonstrate that the multimodal analysis could only slightly improve the prediction results of the extraversion and openness traits in comparison with those obtained in the single feature analysis. Unfortunately, the prediction results of neuroticism, agreeableness, and conscientiousness obtained using multimodal analysis were worse than those obtained by using a single feature set. The best predictors for each Big Five trait (neuroticism, extraversion, openness, agreeableness, and conscientiousness) were A'+L+M, A+L, L+M, CS, and L, respectively. As an overall review, we can conclude that the A features are the most significant features for predicting the Big Five personality traits. Aside from A, the features related to motion and vision (M) are best for predicting the openness and conscientiousness traits.

Subsequently, Table 7 shows the multimodal analysis results for the ELEA-AV corpus. These results indicate that the multimodal analysis could slightly improve the estimation results of the neuroticism and agreeableness traits for this corpus. The best results were achieved by using the audio-related modality (A). In contrast, the Big Five personality trait inference model for extraversion, openness, and conscientiousness could not achieve better performance than that yielded by the model utilizing a single feature set.

5.2.4 Comparison with prior work

We carried out a comparative analysis with [25] for the MATRICS corpus and other related works [2,19,29] for the ELEA-AV corpus regarding the proposed features. For the MATRICS corpus, the evaluation was conducted using 10-fold cross-validation, and the dataset distribution was based on that contained in a prior study [25]. Table 8 shows the comparative results yielded by an ablation test in terms of the F1-score metric. The overall results of our current study were substantially better than those of the prior studies since the estimates of all traits were improved, with an F1-score increase of 8% on average. Significant improvement was achieved in terms of neuroticism and extraversion prediction (more than 10 %).

From Table 8, we could also conclude that the features related to A and M that we used in the current study were more suitable for Big Five estimation with the MATRICS corpus than the features used in the prior study. For instance, the F1-score for predicting the neuroticism trait using the A features was improved from 68 to 79%, whereas the results obtained using the M features improved from 60 to 65%. Furthermore, the best modality for estimating the neuroticism and conscientiousness traits in current work matched well with that in

Table 8 The Big Five personality trait estimation results obtained for the MATRICS corpus with 10-fold cross validation evaluation in the same manner with the prior work [25] (left) and current work (right)

Modality	F1-score (%)									
	N		E		O		Ag		C	
A+L+M+CS	64	↑ 76	55	↑ 76	53	↑ 68	66	↑ 74	56	↑ 69
A+L+M	53	↑ 78	63	↑ 77	59	↑ 69	62	↑ 75	58	↑ 65
A+L+CS	56	↑ 75	57	↑ 76	61	↑ 67	62	↑ 75	64	64
A+M+CS	55	↑ 76	60	↑ 78	57	↑ 69	63	↑ 74	59	↑ 66
L+M+CS	62	↓ 60	62	↑ 65	**64**	↑ 68	61	↑ 62	62	↑ 64
A+L	54	↑ 77	60	↑ 72	53	↑ 64	60	↑ 75	57	↑ 70
A+M	55	↑ 75	60	↑ 73	56	↑ **70**	62	↑ 72	63	↑ 68
A+CS	51	↑ 76	59	↑ **79**	59	↑ 65	64	↑ **77**	61	↑ 63
L+M	63	↓ 60	60	↑ 67	62	↑ 66	63	63	63	↑ 69
L+CS	58	↓ 56	62	↑ 64	62	↓ 57	63	↓ 54	64	↓ 50
M+CS	60	↑ 62	59	↑ 65	60	↑ 65	68	↓ 61	65	↑ 69
A	**68**	↑ **79**	55	↑ **79**	52	↑ 67	**74**	↑ 75	54	↑ 69
L	58	↓ 57	64	↓ 61	58	↑ 60	62	↓ 57	62	↑ 68
M	60	↑ 65	**64**	64	47	↑ 69	52	↑ 58	**66**	↑ **71**
CS	48	↑ 52	46	↑ 58	53	↑ 59	67	67	53	53

These results were obtained using the random forest algorithm with the optimal parameters. Red cells with bold captions represent the best overall prediction results. Blue cells with bold captions represent the best prediction results of each work. Green captions represent the improvement results. Meanwhile, red captions represent the declining results

Table 9 The Big Five personality trait estimation results obtained for the ELEA-AV corpus based on the current work and three prior works by Aran et al. [2], Okada et al. [29], and Kindiroglu et al. [19]

	Classifier	F1-score (%)				
		N	E	O	Ag	C
Proposed	RF	**68**	70	**64**	66	**63**
Aran et al. [2]	Ridge	52	67	55	59	52
	SVM	54	63	62	53	53
Okada et al. [29]	Ridge	56	70	58	**69**	60
	SVM	57	68	57	66	51
Kindiroglu et al. [19]	Ridge	59	**76**	52	56	61
	SVM	61	75	54	59	57
	RF	50	74	53	52	51

The three classifiers used in the corresponding works were a random forest, ridge regression, and a support vector machine (SVM). Red cells represent the best overall prediction results. Blue cells represent the best prediction results of each work

prior work (A and M, respectively). In the current study, the highest F1-score for the estimation of each Big Five personality trait was acquired by the following pairs: neuroticism (A), extraversion (A + CS or A), openness (A + M), agreeableness (A + CS), and conscientiousness (M).

Table 9 shows the comparative results obtained using various features proposed in the current work and three prior works [2,19,29]. The evaluation methods used in all of these works were based on LOPCV. These results show that for most of the Big Five traits (except for the agreeableness trait), the best results obtained by our proposed features could achieve better performance than those in prior works. Significant improvement was obtained by the using audio-related modality (A) for predicting the neuroticism trait (from 61% to 68%).

6 Discussion

In this section, we discuss the key information obtained in this study. We also discuss the prospective multimodal interfaces that utilize the results of our findings. Finally, the limitations and the future direction to address the remaining issues in this study will be discussed.

From the experimental results, as shown in Sect. 5.2, we can discuss two main points that answer the following key questions.

1. Is the speaker individuality feature effective for inferring the Big Five personality traits?
   On the basis of our experimental results, the speaker individuality feature, i.e., the i-vector or x-vector, could improve the prediction performance of the model several traits. For instance, as a unimodal feature, the vector could improve the prediction of the neuroticism and extraversion traits for the MATRICS corpus. On the other hand, it could also achieve accuracy values greater than 60% for the neuroticism, extraversion, openness, and agreeableness traits for the ELEA-AV corpus. These results suggest that the neuroticism and extraversion traits could be represented by the characteristics captured in the state-of-the-art speaker individuality feature from speech. We predicted that these results reflected that the speech characteristics representing speaker individuality were also related to several personality traits. For instance, it has been reported that prosodic features are highly related to speaker individuality [44]. As neuroticism represents the degree of being nervous and extraversion describes the degree of being energetic and active, the perceptions regarding the rising and falling patterns of the voice of a speaker affect the perceptions of these traits. In the case of the conscientiousness trait, our results show that speaker individuality and this trait do not share the same features.
2. What are the effective multimodal features for estimating the Big Five personality traits for the MATRICS and ELEA-AV corpora?
   As shown in Tables 5 and 7, most of the Big Five personality trait predictions obtained by using audio-related features (A) or combining them with another modality achieved the best accuracy for both MATRICS and ELEA-AV corpora. Subsequently, if we use the motion-related feature (M), we could improve the prediction accuracy for the conscientiousness trait. With the MATRICS corpus, we also analyzed the language-related feature (L) and CS indices. Although it was not as effective as M, the conscientiousness trait could also be reflected in the DT feature in L. As a unimodal feature, CS was not as effective for this task as other features. In the ELEA-AV corpus, the Ld indices were effective for predicting the extraversion trait.

Most well-known studies for personality trait estimation focused on self presentation scenarios. For instance, the Speaker Trait Challenge 2012 [42] and the ChaLearn Looking at People 2016 [36]. However, the findings from these studies might be limited because psychological science suggested that situations and social interactions are highly associated with personality states [31]. Only a few studies worked on predicting personality traits in social interactions, including [19,22,25,29]. This study specifically addressed the personality traits estimation using the speaker individuality and multimodal cues in multiple languages group discussion corpora (i.e., MATRICS and ELEA-AV).

As one of the primary key findings, the speaker individuality feature is considered beneficial for estimating neuroticism and extraversion traits in the European or Japanese language group discussion corpus. The neuroticism and extraversion traits are statistically significant in stimulating peoples' attitudes when receiving or making a call at public places [24]. Hence, the estimated personality can be utilized in a virtual

call center for giving customer-centric responses. Besides, we can also build an interface based on speaker embedding to detect the user's attitude. Similarly, the multimodal analysis results of this study could also be used for developing a virtual agent for group interactions that can respond appropriately to each participant based on the estimated personality traits. An appropriate response could lead to a smooth conversation.

While this study provides several key findings, the corpora used in this study might be considerably small-size, limited to the group discussion settings, and consist of only European and Japanese languages. The investigation of more diverse corpora will be considered as a future direction. In this study, we did not focus on analyzing the recent advanced machine-learning algorithms. Instead, we focus on mitigating individual differences from the relatively smaller size but more diverse group discussion corpora, which could be analyzed using classical machine learning algorithms. In future work, we will thoroughly consider how to model personality traits, and other internal properties based on the recent trends in multimodal machine learning [21].

7 Conclusion and future work

This paper analyzed the effectiveness of the state-of-the-art speaker individuality feature, namely, the i-vector, to predict the Big Five personality traits in two different group discussion datasets. Our experimental results showed that this feature could effectively estimate the Big Five personality traits in both datasets, i.e., MATRICS and ELEA. A significant improvement was obtained when predicting the neuroticism and extraversion traits. Subsequently, a multimodal analysis was also carried out to compare the effectiveness of each modality and psychological feature. The psychological features included CS and Ld indices. The results showed that the audio-related features contributed most significantly to this task. An improvement could be achieved by using motion-related features, especially for predicting the conscientiousness trait. Furthermore, the i-vector speaker embedding system could improve the estimation results of personality traits, even when only using one modality (audio-related).

In our future work, we will develop a multimodal interface based on speaker embedding for automatic personality trait estimation using multimodal features. For instance, an interface can give adequate personalized feedback to the user based on the estimated traits. Additionally, recent multimodal machine learning approaches, the relationship between personality traits and other internal properties, and the explainability of the multimodal cues will be thoroughly investigated.

Acknowledgements This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (No. 201605002), a Grant-in-Aid for Scientific Research (B) (No. 21H03463), and a JSPS KAKENHI grant (No. 22K21304). This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6).

Funding This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers 19H01120, 19H01719, 20J20580.

Availability of data and materials Not applicable.

Code Availability Not applicable.

Declaration

Conflict of interest The authors declare no conflict of interest.

References

1. Aran O, Gatica-Perez D (2013) Cross-domain personality prediction: from video blogs to small group meetings. In: Proceedings of the 15th ACM on international conference on multimodal interaction, association for computing machinery, ICMI'13, pp 127–130. https://doi.org/10.1145/2522848.2522858
2. Aran O, Gatica-Perez D (2013) One of a kind: inferring personality impressions in meetings. https://doi.org/10.1145/2522848.2522859
3. Atal B, Schroeder M (1979) Predictive coding of speech signals and subjective error criteria. IEEE Trans Acoust Speech Signal Process 27(3):247–254
4. Baltrusaitis T, Zadeh A, Lim YC, Morency L (2018) Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face gesture recognition (FG 2018), pp 59–66
5. Batrinca L, Mana N, Lepri B, Pianesi F, Sebe N (2011) Please, tell me about yourself: automatic personality assessment using short self-presentations. ICMI'11—proceedings of the 2011 ACM international conference on multimodal interaction, pp 255–262. https://doi.org/10.1145/2070481.2070528
6. Batrinca L, Mana N, Lepri B, Sebe N, Pianesi F (2016) Multimodal personality recognition in collaborative goal-oriented tasks. IEEE Trans Multimed 18(4):659–673. https://doi.org/10.1109/TMM.2016.2522763
7. Bobick A, Davis J (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267. https://doi.org/10.1109/34.910878

8. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
9. Celiktutan O, Eyben F, Sariyanidi E, Gunes H, Schuller B (2014) Maptraits 2014—the first audio/visual mapping personality traits challenge—an introduction: perceived personality and social dimensions. In: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI'14. Association for Computing Machinery, New York, pp 529–530. https://doi.org/10.1145/2663204.2668317
10. Celli F (2012) Unsupervised personality recognition for social network sites
11. Core MG, Allen JF (1997) Coding dialogs with the DAMSL annotation scheme
12. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798. https://doi.org/10.1109/TASL.2010.2064307
13. Dehak N, Torres-Carrasquillo P, Reynolds D, Dehak R (2011) Language recognition via i-vectors and dimensionality reduction. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 857–860
14. Emery N (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24:581–604. https://doi.org/10.1016/S0149-7634(00)00025-7
15. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia, MM'10. Association for Computing Machinery, New York, pp 1459–1462. https://doi.org/10.1145/1873951.1874246
16. Fang S, Achard C, Dubuisson S (2016) Personality classification and behaviour interpretation: an approach based on feature categories. https://doi.org/10.1145/2993148.2993201
17. Ilmini K, Fernando T (2016) Persons' personality traits recognition using machine learning algorithms and image processing techniques. Adv Comput Sci 5:40–44
18. Jayagopi D, Sanchez-Cortes D, Otsuka K, Yamato J, Gatica-Perez D (2012) Linking speaking and looking behavior patterns with group composition, perception, and performance. In: Proceedings of the 14th ACM international conference on multimodal interaction, ICMI'12. Association for Computing Machinery, pp 433–440. https://doi.org/10.1145/2388676.2388772
19. Kindiroglu A, Akarun L, Aran O (2017) Multi-domain and multi-task prediction of extraversion and leadership from meeting videos. EURASIP J Image Video Process. https://doi.org/10.1186/s13640-017-0224-z
20. Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 conference on empirical methods in natural language processing. Association for Computational Linguistics, Barcelona, pp 230–237. https://www.aclweb.org/anthology/W04-3230
21. Liang PP, Zadeh A, Morency LP (2022) Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. https://doi.org/10.48550/ARXIV.2209.03430
22. Lin YS, Lee CC (2018) Using interlocutor-modulated attention BLSTM to predict personality traits in small group interaction. In: Proceedings of the 20th ACM international conference on multimodal interaction, ICMI'18. Association for Computing Machinery, New York, pp 163–169. https://doi.org/10.1145/3242969.3243001
23. Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2015) Recognizing facial expression: Machine learning and application to spontaneous behavior. In: 2012 IEEE conference on computer vision and pattern recognition, vol 2, pp 568–573. https://doi.org/10.1109/CVPR.2005.297
24. Love S, Kewley J (2005) Does personality affect peoples Áttitude towards mobile phone use in public places? Springer, London, pp 273–284. https://doi.org/10.1007/1-84628-248-9_18
25. Mawalim CO, Okada S, Nakano YI, Unoki M (2019) Multimodal bigfive personality trait analysis using communication skill indices and multiple discussion types dataset. In: Meiselwitz G (ed) Social computing and social media. Design, human behavior and analytics. Springer, Cham, pp 370–383
26. Mitrovic D, Zeppelzauer M, Breiteneder C (2010) Features for content-based audio retrieval. Adv Comput 78:71–150
27. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. CoRR. arXiv:1706.08612
28. Nihei F, Nakano YI, Hayashi Y, Hung HH, Okada S (2014) Predicting influential statements in group discussions using speech and head motion information. In: Proceedings of the 16th international conference on multimodal interaction, ICMI'14. Association for Computing Machinery, pp 136–143. https://doi.org/10.1145/2663204.2663248
29. Okada S, Aran O, Gatica-Perez D (2015) Personality trait classification via co-occurrent multiparty multimodal event discovery. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, ICMI'15. Association for Computing Machinery, New York, pp 15–22. https://doi.org/10.1145/2818346.2820757
30. Okada S, Ohtake Y, Nakano YI, Hayashi Y, Huang HH, Takase Y, Nitta K (2016) Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In: Proceedings of the 18th ACM international conference on multimodal interaction, ICMI'16. Association for Computing Machinery, New York, pp 169–176. https://doi.org/10.1145/2993148.2993154
31. Oliver P, John RWR (eds) (2021) Handbook of personality: theory and research. The Guilford Press
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
33. Phan LV, Rauthmann JF (2021) Personality computing: new frontiers in personality assessment. Soc Pers Psychol Compass 15(7):e12624. https://doi.org/10.1111/spc3.12624
34. Philip J, Corr GM (eds) (2009) The Cambridge handbook of personality psychology. Cambridge handbooks in psychology. Cambridge University Press, Cambridge
35. Pianesi F, Mana N, Cappelletti A, Lepri B, Zancanaro M (2008) Multimodal recognition of personality traits in social interactions. https://doi.org/10.1145/1452392.1452404
36. Ponce-López V, Chen B, Oliu M, Corneanu C, Clapés A, Guyon I, Baró X, Escalante HJ, Escalera S (2016) ChaLearn LAP 2016: first round challenge on first impressions—dataset and results. In: European conference on computer vision
37. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlíček P, Qian Y, Schwarz P, Silovský J, Stemmer G, Vesel K (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on automatic speech recognition and understanding
38. Sanchez-Cortes D, Aran O, Gatica-Perez D (2011) An audio visual corpus for emergent leader analysis. In: Multimodal corpora for machine learning: taking stock and road mapping the future
39. Sanchez-Cortes D, Aran O, Jayagopi D, Mast M, Gatica-Perez D (2013) Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. J Multimodal User Interfaces 7:39–53. https://doi.org/10.1007/s12193-012-0101-0
40. Sato N, Obuchi Y (2007) Emotion recognition using mel-frequency cepstral coefficients. J Nat Lang Process 14:83–96
41. Schuller BW (2013) Intelligent audio analysis. Springer Publishing Company, Incorporated, Berlin

42. Schuller BW, Steidl S, Batliner A, Nöth E, Vinciarelli A, Burkhardt F, van Son R, Weninger F, Eyben F, Bocklet T, Mohammadi G, Weiss B (2012) The INTERSPEECH 2012 speaker trait challenge. In: INTERSPEECH 2012, 13th annual conference of the international speech communication association, Portland, Oregon, USA, September 9–13, 2012, ISCA, pp 254–257. http://www.isca-speech.org/archive/interspeech_2012/i12_0254.html
43. Shriberg E, Dhillon R, Bhagat S, Ang J, Carvey H (2004) The ICSI meeting recorder dialog act (MRDA) corpus. In: Proceedings of the 5th SIGdial workshop on discourse and dialogue at HLT-NAACL 2004. Association for Computational Linguistics, Cambridge, Massachusetts, USA, pp 97–100. https://www.aclweb.org/anthology/W04-2319
44. Shriberg E, Ferrer L, Kajarekar S, Venkataraman A, Stolcke A (2005) Modeling prosodic feature sequences for speaker recognition. Speech Commun 46:455–472. https://doi.org/10.1016/j.specom.2005.02.018
45. Snyder D, Garcia-Romero D, Povey D (2015) Time delay deep neural network-based universal background models for speaker recognition. In: 2015 IEEE Workshop on automatic speech recognition and understanding (ASRU), pp 92–97
46. Snyder D, Garcia-Romero D, Povey D, Khudanpur S (2017) Deep neural network embeddings for text-independent speaker verification. https://doi.org/10.21437/Interspeech.2017-620
47. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-Vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5329–5333
48. Stevens SS, Volkmann JE, Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. J Acoust Soc Am 8:185–190
49. Talkin D (2005) A robust algorithm for pitch tracking (RAPT). Elsevier Science BV
50. Terasawa H, Slaney M, Berger J (2005) Perceptual distance in timbre space
51. Yl Tian, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115. https://doi.org/10.1109/34.908962
52. Tokuda K, Oura K, Takenori Y, Tamamori A, Sako S, Zen H, Nose T, Takahashi T, Yamagishi J, Nankaku Y (2017) Speech signal processing toolkit (SPTK) version 3.11. http://sp-tk.sourceforge.net/
53. Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Trans Affect Comput 5(3):273–291. https://doi.org/10.1109/TAFFC.2014.2330816
54. Weidenbacher U, Layher G, Bayerl P, Neumann H (2006) Detection of head pose and gaze direction for human–computer interaction. In: Proceedings of the 2006 international tutorial and research conference on perception and interactive technologies, PIT'06. Springer, Berlin, pp 9–19. https://doi.org/10.1007/11768029_2
55. Wood E, Baltruaitis T, Zhang X, Sugano Y, Robinson P, Bulling A (2015) Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE international conference on computer vision (ICCV), pp 3756–3764
56. Xue D, Wu L, Hong Z, Guo S, Gao L, Wu Z, Zhong X, Sun J (2018) Deep learning-based personality recognition from text posts of online social networks. Appl Intell 48(11):4232–4246. https://doi.org/10.1007/s10489-018-1212-4
57. Zadeh A, Lim YC, Baltrušaitis T, Morency L (2017) Convolutional experts constrained local model for 3d facial landmark detection. In: 2017 IEEE International conference on computer vision workshops (ICCVW), pp 2519–2528