INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multimodal Personality Traits Assessment (MuPTA) Corpus: The Impact of Spontaneous and Read Speech

Elena Ryumina[1], *Dmitry Ryumin*[1], *Maxim Markitantov*[1], *Heysem Kaya*[2], *Alexey Karpov*[3]

[1]St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
[2] Department of Information and Computing Sciences, Utrecht University, The Netherlands
[3]ITMO University, St. Petersburg, Russia

{ryumina.e, ryumin.d, markitantov.m}@iias.spb.su, h.kaya@uu.nl, karpov_a@mail.ru

Abstract

Automatic personality traits assessment (PTA) provides high-level, intelligible predictive inputs for subsequent critical downstream tasks, such as job interview recommendations and mental healthcare monitoring. In this work, we introduce a novel Multimodal Personality Traits Assessment (MuPTA) corpus. Our MuPTA corpus is unique in that it contains both spontaneous and read speech collected in the midly-resourced Russian language. We present a novel audio-visual approach for PTA that is used in order to set up baseline results on this corpus. We further analyze the impact of both spontaneous and read speech types on the PTA predictive performance. We find that for the audio modality, the PTA predictive performances on short signals are almost equal regardless of the speech type, while PTA using video modality is more accurate with spontaneous speech compared to read one regardless of the signal length.

Index Terms: audio-visual resources, data annotation, multimodal paralinguistics, personality computing, big five traits

1. Introduction

Personality Computing (PC) is a multi-disciplinary field that combines both psychology and computer science to analyze human personality traits using various computational methods. Personality traits are believed to be relatively stable over time, and they are a key factor in shaping such human's individual patterns as thoughts, feelings, and behaviors [1]. Big Five model describes these patterns and comprises five personality traits, namely, Openness to experience (OPE), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), Neuroticism/Non-Neuroticism (NEU/NNEU).

The importance of PC lies in its strong relation to high-risk tasks such as job interview recommendation [2, 3], conversational interfaces [4] and mood disorders. Multiple studies in medical and social sciences, including [5, 6], have reviewed the association of personality traits with mood disorders, such as major depressive disorder. This meta-analysis indicated a strong connection between some mental illnesses and personality, of which all disorders had a configuration of low CON and high NEU values.

To date, all collected corpora for personality traits assessment (PTA) have exclusively included spontaneous speech only. While spontaneous speech can reveal an emotional tone and cognitive style through the analysis of word frequency and speech patterns [7], read speech can also provide valuable information on human's personality traits via non-verbal information. For example, the same phrase can be pronounced in a various way by different human beings with varying speech prosody [8]. Additionally, both types of speech can exert changes in human facial expressions and behavior.

In this paper, we consider PTA using read speech for the first time. We also compare which type of speech allows better assessment of the human personality. For that, we collected a novel Multimodal Personality Traits Assessment (MuPTA) corpus that contains both spontaneous and read speech.

2. Related work

2.1. Existing multimodal corpora

Automatic PTA can be performed by three communication modalities: audio (prosodic, energy-based and spectral features, voice quality, etc.) [9], visual (facial expressions, scene, aesthetic preference, etc.) [10, 11], and text (sentiment word and its meaning, etc.) [12]. A brief description and comparison of several existing multimodal corpora is presented in Table 1.

A review of multimodal corpora for PTA shows that: (1) most of the existing corpora are in English; (2) all the corpora were collected "in-the-Wild" or in office conditions; (3) corpora contain spontaneous speech on a fixed topic; (4) the personality traits annotation is made according to the results of self-evaluation, familiar- or third-party-evaluation; (5) there is an uneven gender distribution; (6) most of the speakers are young people under 30 y.o.

Thus, our MuPTA corpus differs from others in that it contains audio-visual recordings from 30 native Russian speakers with a uniform distribution per gender and age, and also includes both spontaneous and read speech.

2.2. State-of-the-art approaches

Several competitions focused on developing approaches for multimodal PTA were organized in prominent international conferences, including INTERSPEECH 2012 [21], CVPR 2017 [22], and ICCV 2021 [23]. Two corpora, namely FI v2 (ChaLearn First Impressions V2) [3] and UDIVA [18], were presented and used in the respective competitions, where competitors using a common protocol developed and tested their approaches. Regardless of the corpus used, there are several trends that have shown positive effects on the performance of the proposed approach. In this study, we develop a baseline approach using video (face) and audio modalities. Therefore, state-of-the-art (SOTA) approaches are only considered for these two modalities.

Audio modality. The use of log-Mel spectrograms [24, 25] to extract speech features from a signal prevails over other features such as hand-crafted features (e.g., openSMILE) [2, 26, 27], and raw audio signals [28]. Log-Mel spectrograms are used coupled with 2D Convolutional Neural Networks (CNN) [24, 25], Long Short-Term Memory Networks (LSTM) [25] or Fully Connected Neural Networks (FCNN) [24].

10.21437/Interspeech.2023-1686

Table 1: *Comparison of multimodal corpora: Fr – French, En — English, Spa – Spanish, Cat – Catalan, Ge – German.*

Corpus	Language	Evaluation	# Subjects	# Male/Female	Age (Range/Mean)	Duration (h)
ELEA [13]	Fr, En	Self	148	100/48	NA/25	10
Hire Me [14]	En	Self	62	17/45	NA/24	11
YouTube vlogs [15]	En	Third-party	442	208/234	[10,60]/NA	48
JOKER [16]	Fr	Self	37	23/14	[21,61]/35	8
MHHRI [17]	En	Self, familiar	18	9/9	NA	6
FI V2 [3]	En	Third-party	3060	1312/1748	[8,62]/24	41
MULTISIMO [9]	En	Self, familiar	49	24/25	[19,44]/30	4
UDIVA [18]	Spa, Cat, En	Self, familiar	147	81/66	[4,84]/31	90
RoomReader [19]	En	Self, familiar	118	51/65	[18,43]/24	8
DyCoDa [20]	Ge	Self	30	21/9	NA/22	10
MuPTA (ours)	Ru	Self	30	15/15	[19,86]/41	7

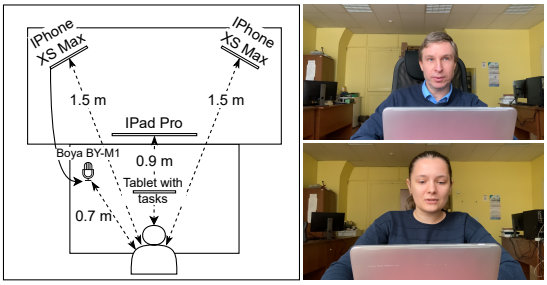

Figure 1: *Recording setup and sample frames from videos.*

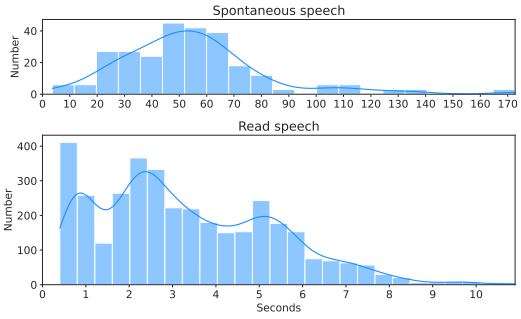

Figure 2: *Distribution of recordings' duration. The blue line is the distribution density.*

Video modality. Raw face images [2, 24, 25, 26, 27, 28] are mainly used as features, which significantly dominate over expert features such as image histograms [22], Local Gabor Binary Patterns [2], etc. The raw face images are used as input data for the 2D/3D/(2+1)D CNNs with the addition of spatial-temporal models (such as LSTM [25, 27] and Transformer [26]), Extreme Learning Machines (ELM) [2] and FCNN [28, 24].

It is challenging to determine best-performing unimodal deep models for audio and video modalities because most papers present results after combining audio, video (face and scene), and text modalities. Fusion of these modalities is usually done at the feature-level using Transformer [26, 24], Random Forest [2], Extra Tree Regressor [28], or LSTM [25, 27].

3. MuPTA corpus

3.1. Data collection

The MuPTA corpus contains data of 30 native Russian speakers. A comparison of the MuPTA corpus with existing multimodal corpora for PTA is presented in Table 1. The corpus was recorded using three devices: two Apple iPhone XS Max smartphones and one Apple iPad Pro tablet. The audio data were collected with a sampling frequency of 48 kHz, 16 bits per sample, mono format. We use the following video parameters: 4K resolution (3840×2160 pixels), frames per second (FPS) is 60 (for smartphones) and 30 (for tablet), the color coding is 24 bits per pixel. The recording setup was similar to the study [29] and sample video frames are shown in Figure 1. Each speaker completed three various tasks: (1) briefly introduced him/herself; (2) described what is happening in two complicated pictures; (3) read some scripted sentences out loudly (a list of 40 sentences was prepared for reading).

The phonetically balanced text presented in [30] was used to select utterances for reading. This text was carefully curated to investigate the speech patterns and variations of native Russian speakers with distinctive phonetic features. It allows developing a complete speech profile of the speaker. The text also contains dialogs with interrogative, exclamatory and affirmative sentences, that allow highlighting differences in personality traits. The number of words in sentences ranges from 1 to 22, the average number of words is 8 with a standard deviation (std) of 5. In total, 43 utterances were recorded from each speaker: 3 spontaneous (tasks 1 and 2, the latter having two sub-tasks) and 40 scripted sentences (task 3). Note that PTA by read speech was not studied before.

Figure 2 shows the duration distribution of the recorded phrases. The duration of spontaneous speech is at least 4 sec. The duration of read speech varies from 0.4 to 11.5 sec, 2.5 sec in average. In total, MuPTA consists of 4.1 hours of spontaneous speech and 3.3 hours of read one.

The data of each informant were recorded continuously by all three devices, so we have split the recorded files into phrases. Firstly, we annotated the start and end points of speech activity using Adobe Audition. We then synchronized the recordings of all three devices using Adobe Premier Pro. Finally, we have split all the recordings using obtained speech timestamps and shifts for each channel.

3.2. Data annotation

Each speaker (informant) have filled in a self-evaluation questionnaire of 60 questions [31]. This is a standard questionnaire

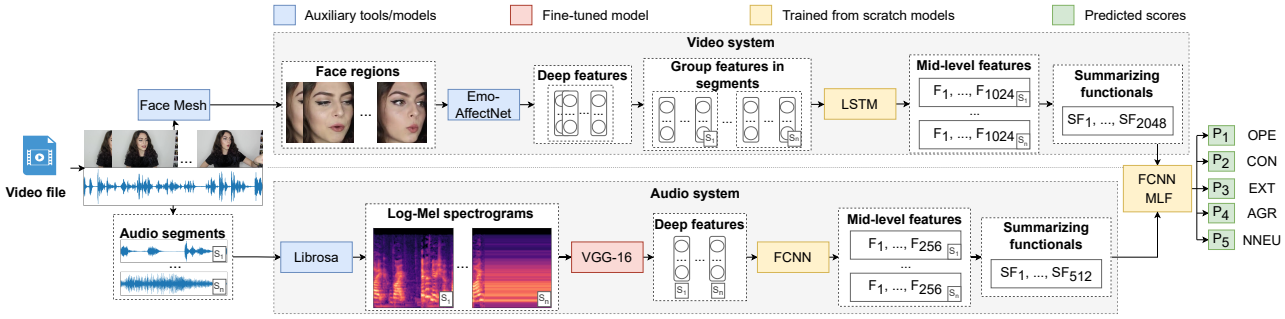

Figure 3: *Pipeline of the proposed audio-visual approach for personality traits assessment. MLF means mid-level feature fusion.* S_i *– audio/video segment,* $i = 1, ..., n$*,* n *- the number of 2-sec segments with 1-sec steps in the audio-visual signal.*

used to estimate big five traits. Adapted versions of questionnaires for Russian are presented in [32]. All questions were Likert type with scores in the range from 1 to 5. 60 questions evenly covered five traits, i.e., there were 12 questions for each trait. Hence, the maximum total score that can be obtained for one trait is 60. We normalized all the scores into [0, 1] range. Moreover, since this questionnaire provides the scores for the NEU negative trait, we have converted negative scores into positive ones (as in [3]) and got the NNEU positive trait. It allows scaling all the traits into a positive scale. Obviously, a self-evaluation can be biased, so in the future we are going to annotate the data by third-party evaluations as well, as in [15].

Each informant has provided the following own metadata: birthday, gender, marital status, education, and occupation. All the informants were asked to complete an informed consent form prior to data collection. Such metadata can be used for PTA as in [18]. The same metadata can also be used for fairness analysis and bias mitigation as well [3].

The collected data were partitioned into three subsets: Train (18 speakers, 11610 utterances), Development (6 speakers, 3870 utterances) and Test (6 speakers, 3870 utterances). This speaker-independent partitioning was mode taking into account balanced gender and age distributions.

4. Proposed approach

The pipeline of the proposed audio-visual approach for PTA using a mid-level feature fusion is shown in Figure 3. The approach integrates audio and video subsystems. The Video system receives downsampled frames (5 FPS) as an input. Mid-level NN-based features for both systems are extracted from 2 sec segments with 1 sec steps. For each set of mid-level NN-based features, we calculate both mean and std values, concatenate and pass them into a FCNN in order to estimate personality traits scores for the whole clip. In [18, 33], chunking was done with 1.5 and 2.5 sec segments for the PTA task. We chose a segment length of 2 sec, because the most utterances of read speech are between 2 and 2.5 sec.

We trained all the models using the Adam optimizer for 100 epochs and the Cosine Annealing Learning Schedule [34] with 5-rate restart cycles. Such hyper-parameters as the number of layers and units in them, a dropout probability, and a learning rate are selected experimentally by grid search.

Our approach differs from other SOTA approaches in that we: (1) downsample frames and segment the clips; (2) fine-tune two models to extract mid-level features at the segment-level; (3) use feature-level fusion at the clip-level to calculate the predictions. This strategy reduces the number of model parameters, making our approach suitable for real-time applications.

4.1. Audio system

The log-Mel spectrograms with 128 Mel filter banks were extracted by the open-source library Librosa [35] from each audio file. The size of the feature matrix for a 2-sec audio segment is 128×173. The features are padded with mean values in the case an audio segment is shorter than 2 sec. The extracted features were converted into images, resized to 224×224 pixels and repeated three times. So, we use input vectors of 224×224×3, which are then normalized to the range [0, 1].

We apply the pre-trained VGG-16 model [36] for extracting deep features from log-Mel spectrograms. This model has been successfully used in the PTA task [24, 25]. FCNN is used for the final regression task; it consists of three fully connected layers (FCL) with 512, 256, and 5 neurons, as well as a linear activation function for the last layer. In the training process, the learning rate ranged from 5e-5 to 5e-6.

4.2. Video system

We apply the Face Mesh model [37] from the MediaPipe library for detecting facial regions and 468 3D facial landmarks. We chose this model because of the richness of facial landmarks. Also, since the frame rates of the videos are different, each video file is downsampled to 5 FPS to keep the same processing conditions for LSTM networks. For a 2-sec video segment, the number of frames is 10. The last frame is repeated as many times as necessary if the video segment is shorter than 2 sec.

It is known that personality traits are determined by the sequence of emotional and behavioral reactions of different people to the same stimuli [38]. Inspired by this fact and the research [39], we apply the open-source Emo-AffectNet model [40] for extracting 512 deep emotional facial features. This model's performance is confirmed in the emotion recognition task recently [40]. The size of the feature matrix for a 2-sec video segment is 10×512.

A single-hidden-layer LSTM model is used to extract mid-level features from the videos. This model comprises one LSTM layer with 1024 units and one FCL with 5 neurons with a linear activation function. We trained this model with a learning rate ranging from 5e-3 to 5e-4.

4.3. Audio-visual system

The duration of audio-visual clips differs in the MuPTA corpus. For example, if the duration of a clip is 15 sec, it amounts to 16 (overlapping) segments. Hence we get $16 \times F$ feature ma-

Table 2: *Results of the proposed systems obtained on the MuPTA corpus. A, V, AV denote audio, video, and audio-visual systems (S), respectively.*

	TRAIT-WISE ACC					AVERAGE	
S	**OPE**	**CON**	**EXT**	**AGR**	**NNEU**	**ACC**	**CCC**
Development set							
A	.946	.905	.895	.852	.895	.903	.650
V	**.947**	**.914**	**.909**	**.864**	.906	**.908**	.626
AV	**.947**	.907	.889	.857	**.914**	.903	**.672**
Test set							
A	**.936**	.921	.849	.901	.869	.895	.574
V	**.936**	**.931**	.841	**.902**	.868	.895	.523
AV	.935	.915	**.876**	.898	**.871**	**.899**	**.614**

trix for the clip, where F is the number of mid-level features that varies depending on the systems. Then we aggregate mean and std values (summarizing functionals) of these features per clip. Thus, we extract feature vectors of 256×2 components for the audio system, and 1024×2 – for the video one. These vectors are concatenated into a joint vector and used as an input for FCL. Unlike SOTA approaches [2, 24, 25, 26, 27, 28], we perform feature-level fusion using a single FCL, which is a simple and effective way to fuse the modalities.

5. Experimental results and discussion

We evaluated the proposed approach for PTA using standard performance measures: Accuracy (ACC), which is calculated as 1 - Mean Absolute Error (MAE) as in [3], and Concordance Correlation Coefficient (CCC). While ACC reflects the error between predicted and ground truth scores, CCC indicates a correlation between them. CCC is more robust compared to the Pearson's Correlation Coefficient (PCC), as it also considers the difference in means [41]:

$$CCC = \frac{2 \cdot \sigma_{t,p}}{\sigma_t^2 + \sigma_p^2 + (\mu_t - \mu_p)^2}, \quad (1)$$

where μ_t and μ_p denote the averaged ground truth and predicted scores for all test clips, respectively; σ_t and σ_p – the respective standard deviations; $\sigma_{t,p}$ – the covariance between t and p.

Table 3: *Comparison of the CCC of systems in the case of spontaneous (SP) and read (RE) speech type (ST), as well as segment duration (seconds).*

S	**ST**	**2**	**5**	**10**	**20**	**30**	**50**	**Whole**
A	SP	.575	**.578**	.575	.575	.574	.574	.574
	RE	.575	.574	.574	–	–	–	.574
V	SP	.518	.529	.530	**.539**	.538	.537	.536
	RE	.504	.522	.522	–	–	–	.522
AV	SP	.601	.618	.619	**.632**	.631	.628	.628
	RE	.588	.612	.613	–	–	–	.613

The results obtained on the MuPTA corpus are presented in Table 2, which indicates that in the Development set, AGR trait is the most difficult to predict, whereas in the Test set, EXT and NNEU are the least performing dimensions. This is because the distribution of ground truth scores for these traits in the Development and Test sets differs from that in the Train set. The ACC measure shows that the visual system outperforms the audio system. However, according to the CCC measure, the audio predicted scores are more reliable than the visual ones. Combining both systems leads to an absolute increase of 2.2% in CCC value (.650 vs. .672) for the Development set and 4.4% (.574 vs. .614) for the Test set. It demonstrates that PTA is more reliable when the audio and video systems are fused.

The performance measures for spontaneous and read speech at various segment durations are compared in Table 3. To present the measures, we cut the clips to {2-50} sec, or use the whole clip. There are no read utterances longer than 10 sec in MuPTA. The audio system shows almost no difference in performance between two speech types at short signals (2 sec).

Unlike the audio system, the visual one shows better CCC performance with spontaneous speech compared to read speech regardless of the signal length. The best CCC performance is achieved with a signal length of 20 sec. Despite the fact that the audio system outperforms the visual one at all signal durations, the audio-visual system shows a bias towards the video modality and displays a similar performance pattern.

Thus, we can draw the following conclusions: (1) read speech is informative in the case of audio system, unlike spontaneous speech, that is more informative for video and multimodal systems; (2) the optimal signal length for multimodal PTA is 20 sec, in which case the audio system works well; the video and multimodal systems reach their best performances.

It should be noted that our proposed approach works in real-time. Processing a 15.3 sec clip on a CPU (using Intel i9) takes 9.1 sec with a frame resolution of 3840×2160, of which 7.9 sec are needed to process video data. With a change in frame resolution to 1280×720, the processing time reduces to 3.8 sec.

6. Conclusions

In this paper, we presented the Multimodal Personality Traits Assessment (MuPTA) corpus, which is the first corpus that contains both spontaneous and read audio-visual speech. In addition, we propose a real-time multimodal approach for personality computing, by which we compared which type of speech allows better assessment of the human's personality traits. As a result, we find that audio modality outperforms video modality in terms of performance, while multimodal fusion gives the best performance. In addition, the optimal signal length for video and multimodal systems is observed to be 20 sec, while for audio it is shorter (5 sec). Lastly, for video and multimodal systems, spontaneous speech is found to suit the PTA task better, while both speech types of a short length are almost equally performing for the audio modality. Both the source code of our approach and the MuPTA corpus can be found at the web-page[1].

In the future, we plan to improve our approach by incorporating modalities like text, video (scene), and metadata. Moreover, we plan to conduct a large scale cross-corpus and multilingual research on personality traits assessment.

7. Acknowledgements

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement No. 70-2021-00141.

[1] https://oceanai.readthedocs.io

8. References

[1] A. Weise and R. Levitan, "Investigating the influence of personality on acoustic-prosodic entrainment," in *INTERSPEECH*, 2022, pp. 3093–3097.

[2] H. Kaya *et al.*, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *CVPRW*, 2017, pp. 1–9.

[3] H. J. Escalante *et al.*, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 894–911, 2020.

[4] D. Fernau *et al.*, "Towards Automated Dialog Personalization using MBTI Personality Indicators," in *INTERSPEECH*, 2022, pp. 1968–1972.

[5] D. N. Klein *et al.*, "Personality and depression: Explanatory models and review of the evidence," *Annual Review of Clinical Psychology*, vol. 7, pp. 269–295, 2011.

[6] R. Kotov *et al.*, "Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis." *Psychological bulletin*, vol. 136, no. 5, p. 768, 2010.

[7] C. Snyder *et al.*, "Individual variation in cognitive processing style predicts differences in phonetic imitation of device and human voices." in *INTERSPEECH*, 2019, pp. 116–120.

[8] A. Zakeri and H. Hassanpour, "Whispernet: Deep siamese network for emotion and speech tempo invariant visual-only lip-based biometric," in *International Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 2021, pp. 1–5.

[9] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus," in *International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 2945–2951.

[10] C. Palmero *et al.*, "Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results," in *Understanding Social Behavior in Dyadic and Small Group Interactions*, 2022, pp. 4–52.

[11] Y. Yang *et al.*, "Personalized image aesthetics assessment with rich attributes," in *CVPRW*, 2022, pp. 19 861–19 869.

[12] Y. Mehta *et al.*, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2313–2339, 2020.

[13] D. Sanchez-Cortes *et al.*, "An audio visual corpus for emergent leader analysis," in *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road Mapping the Future, ICMI-MLMI*. Citeseer, 2011, pp. 1–6.

[14] L. S. Nguyen *et al.*, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1018–1031, 2014.

[15] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, 2012.

[16] L. Devillers *et al.*, "Multimodal data collection of human-robot humorous interactions in the Joker project," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 348–354.

[17] O. Celiktutan *et al.*, "Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2017.

[18] C. Palmero *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1–12.

[19] J. Reverdy *et al.*, "RoomReader: A multimodal corpus of online multiparty conversational interactions," in *International Conference on Language Resources and Evaluation (LREC)*, 2022, pp. 2517–2527.

[20] D. Dresvyanskiy *et al.*, "DyCoDa: A multi-modal data collection of multi-user remote survival game recordings," in *International Conference on Speech and Computer*, 2022, pp. 163–177.

[21] B. Schuller *et al.*, "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer speech & language*, vol. 29, no. 1, pp. 100–131, 2015.

[22] S. Bekhouche *et al.*, "Personality traits and job candidate screening via analyzing facial videos," in *CVPRW*, 2017, pp. 10–13.

[23] H. Salam *et al.*, "Fact sheet: Automatic self-reported personality recognition track," in *Understanding Social Behavior in Dyadic and Small Group Interactions Challenge*, 2021, pp. 1–4.

[24] T. Agrawal *et al.*, "Multimodal personality recognition using cross-attention transformer and behaviour encoding," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, vol. 5, 2022, p. 501–508.

[25] S. Aslan *et al.*, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, vol. 110, p. 104163, 2021.

[26] T. Agrawal *et al.*, "Multimodal vision transformers with forced attention for behavior analysis," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3392–3402.

[27] A. Subramaniam *et al.*, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *ECCV Workshops*, 2016, pp. 337–348.

[28] Y. Li *et al.*, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," *International Journal of Computer Vision*, vol. 128, no. 12, pp. 2763–2780, 2020.

[29] M. Markitantov *et al.*, "Biometric russian audio-visual extended masks (BRAVE-MASKS) corpus: Multimodal mask type recognition task," in *INTERSPEECH*, 2022, pp. 1756–1760.

[30] S. Stepanova, "The phonetic properties of russian speech: implementation and transcription," Ph.D. dissertation, 1988.

[31] C. Soto and O. John, "The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power." *J. of Personality and Social Psychology*, vol. 113, no. 1, pp. 117–143, 2017.

[32] S. Shchebetenko, "The best man in the world: Attitudes toward personality traits," *Psychology. Journal of Higher School of Economics*, vol. 11, no. 3, pp. 129–148, 2014.

[33] D. Giritlioğlu *et al.*, "Multimodal analysis of personality traits on videos of self-presentation and induced behavior," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 337–358, 2021.

[34] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *ICLR*, 2017, pp. 1–16.

[35] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Python in Science Conference*, vol. 8, 2015, pp. 18–25.

[36] O. Verkholyak *et al.*, "Ensemble-within-ensemble classification for escalation prediction from speech," in *INTERSPEECH*, 2021, pp. 481–485.

[37] I. Grishchenko *et al.*, "Attention mesh: High-fidelity face mesh prediction in real-time," in *CVPRW on Computer Vision for Augmented and Virtual Reality*, 2020, pp. 1–4.

[38] R. R. McCrae and P. T. C. Jr., "The five-factor theory of personality," *Handbook of personality: Theory and research*, vol. 3, pp. 159–181, 2008.

[39] F. Gürpinar *et al.*, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 43–48.

[40] E. Ryumina *et al.*, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing*, vol. 514, pp. 435–450, 2022.

[41] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.