AVI Challenge 2025: Assessing Personality Traits and Interview Performance from Asynchronous Video Interviews (AVIs)

Abstract

Asynchronous Video Interviews (AVIs) allow candidates to record responses to predefined questions using digital devices, offering flexibility and enabling remote assessments. Evaluating personality traits and interview performance through AVIs provides organizations with valuable insights into candidates' profiles and helps predicting future job performance. However, previously held challenges, whose data are mostly sourced from social media, have suboptimal construct and methodological validity, limiting their value for model development and practical applications. To address these gaps, we introduce the "AVI Challenge 2025", featuring a novel dataset of mock AVIs (3876 videos from 646 subjects) conducted in a simulated job application procedure. Interview questions were designed to reflect real-world selection contexts and activate personality traits based on Trait Activation Theory. Annotations for personality traits and job competencies were provided by trained evaluators and professional recruiters, ensuring methodological and ecological validity. Our challenge addresses limitations in prior challenges by (a) utilizing behaviorally anchored rating scales (BARS) for accurate personality assessment, (b) incorporating theory-driven tasks to enhance construct validity, and (c) aligning with real-life AVI practices for practical relevance. Additionally, ongoing data collection from Dutch speakers will make the dataset cross-cultural and multilingual, broadening its utility for future research and applications in global contexts.

ACM Reference Format:

. 2025. AVI Challenge 2025: Assessing Personality Traits and Interview Performance from Asynchronous Video Interviews (AVIs). In *Proceedings of ACM International Conference on Multimedia (ACM MM '25).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM MM '25, Dublin, Ireland

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

https://doi.org/10.1145/nnnnnnn.nnnnnnn

A Baseline method

To establish a baseline performance for the challenge, we propose a baseline assess framework utilizing classic multimedia deep learning methods. This framework aims to assist challenge participants in gaining a clear understanding of the expected performance standards and provide a reference point for evaluating their own models. As illustrated in Fig. 1, the proposed system processes single-speaker video clips to assess the personality traits for *Track 1*, and evaluate the interview performance for *Track 2*. The pipeline of the system can be divided into three modalities.

- ***Visual modality***: Refers to the imagery seen on screen, including movement, facial expressions, body language, and scene composition.
- ***Audio modality***: Includes all the sounds present in a video, such as dialogue.
- ***Verbal modality***: Represents the spoken words and language used in the video, often conveyed through dialogue or narration.

We process the input video in three subsequent steps. First, a feature extraction module operates on the input video to compute a set of features. Second, as these features are computed locally in the time axis, a temporal aggregation module is utilized to process the extracted features to describe the entire video, the output features of all modalities are then fused to provide a multi-modal feature vector. In the end, a regression module takes this feature vector as input and predicts the dependent variables. In the following, we detail each module along with its functionality.

A.1 Feature extraction

A.1.1 Visual modality. As humans, we communicate our emotions through our facial expressions. To exploit such pronounced information, the visual modality aims to describe the individual's observable behavior in the video. To this end, we leverage CLIP[1] (Contrastive Language-Image Pre-Training) neural network trained on a vast dataset of image-text pairs, for visual feature extraction, to detect subtle cues in appearance, body language, and visual context for personality prediction.

A.1.2 Audio modality. In addition to visually perceived behavior, audio features (i.e., voice characteristics) may carry

[1] https://huggingface.co/openai/clip-vit-base-patch32

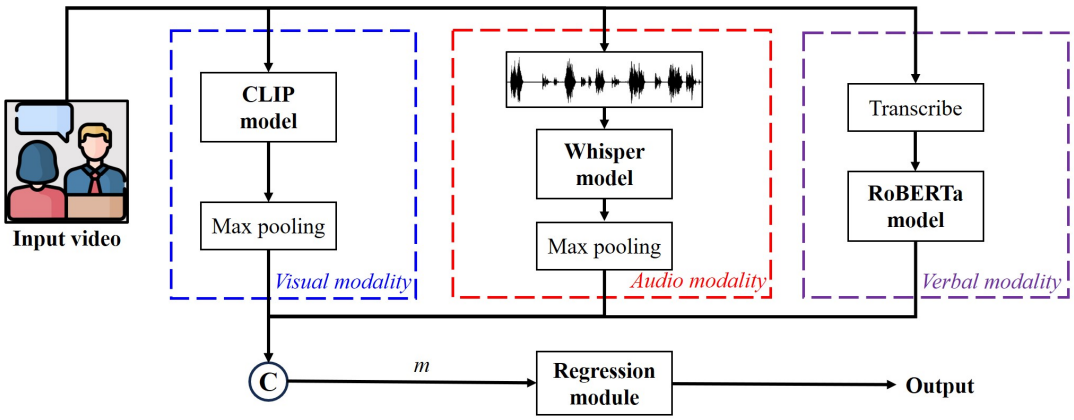

Figure 1. Pipeline overview: visual (blue), audio (red), verbal (purple) modalities, and 'C' stands for concatenation. During training, only the regression module requires personality annotations.

information on both self-rated and observer-rated personality traits. In this context, we use whisper[2], a general-purpose speech recognition model trained on a large dataset of diverse audio (680k hours) as well as various speech processing tasks including voice activity detection, multilingual speech recognition, and spoken language identification, etc., to extract audio feature.

A.1.3 Verbal modality. Spoken and written text is one of the most studied behaviors to infer personality traits. The most influential models of personality themselves (e.g., the Big Five and its successor, the HEXACO model) are based on the lexical tradition. In the proposed system, the verbal modality processes the transcription of individuals' speech using the RoBERTa[3] (Robustly Optimized BERT Pretraining Approach), an advanced version of the BERT (Bidirectional Encoder Representations from Transformers) model. Similar to BERT, RoBERTa is a transformer-based language model that employs self-attention to analyze input sequences and produce contextualized word representations within a sentence.

A.2 Modality fusion

To address the potential issue of varying time steps or temporal misalignment across modalities, the audio, visual, and verbal feature vectors are concatenated to form a unified multimodal feature vector, denoted as m in Fig. 1.

A.3 Regression

In the last block of the proposed system, the multimodal feature vector, m, is fed into a regression module, to predict the dependent variables, as shown in Fig. 1. In the system, regression is performed using the deep ensemble approach. Specifically, a deep ensemble of 32 multilayer perception (MLPs) with 32 and 8 hidden units are deployed. Each MLP is trained using random weights initialization and for 200 epochs minimizing the mean squared error.

[2]https://github.com/openai/whisper
[3]https://huggingface.co/FacebookAI/roberta-base

Table 1. Results of the proposed baseline system for personality assessment (*Track 1*).

	H	E	A	C
Proposed	0.1915	0.1105	0.2287	0.1877

Table 2. Results of the proposed baseline system for interview performance assessment (*Track 2*).

	Integr	Colleg	Soc	Dev	Hirea
Proposed	0.1456	0.2638	0.2379	0.1996	0.2664

A.4 Results

Table 1 presents the results of the proposed baseline system for personality assessment (Track 1). The table reports the system's performance across four personality traits: Honesty-Humility (H), Emotionality (E), Agreeableness (A), and Conscientiousness (C). The values represent the evaluation metric, Mean Squared Error (MSE), for each trait. These results reflect the baseline model's performance in predicting the respective personality dimensions.

Table 2 presents the results of the proposed baseline system for interview performance assessment (Track 2). The table displays the predicted values for five distinct dimensions of interview performance: Integrity (Integr), Collegiality (Colleg), Social versatility (Soc), Development orientation (Dev), and Overall hireability (Hirea). These results provide an insight into the system's effectiveness in evaluating various facets of interview performance, with each dimension contributing to the overall assessment of a candidate's suitability for hire.