SpeakerAICoach: A Multi-Agent Mobile Presenter Training

Demonstration Track

Andrey Savchenko
Sber AI Lab
ISP RAS Research Center for Trusted Artificial Intelligence
Moscow, Russia
avsavchenko@hse.ru

Anna Slovyagina
Irina Bogatyreva
HSE University
Moscow, Russia

ABSTRACT

In this paper, we introduce SpeakerAICoach, a lightweight multimodal framework that evaluates and coaches presentation skills from recorded video. The system decomposes analysis into vision and speech agents that extract verbal and nonverbal cues (filler words, speech clarity, gesture dynamics, gaze, facial affect, clothing). Moreover, we introduce two state-of-the-art lightweight models for age, gender, and ethnicity estimation, which facilitate speaker demography analysis. An aggregator agent merges per-segment analytics, and an LLM agent generates tailored coaching suggestions. The mobile demo for Android showcases fragment-level feedback, interactive visualizations, and an extensible agent pipeline, which is suitable for on-device clients with server-side inference.

KEYWORDS

Multimodal analysis; presentation coaching; human-AI interaction

ACM Reference Format:
Andrey Savchenko, Anna Slovyagina, and Irina Bogatyreva. 2026. SpeakerAICoach: A Multi-Agent Mobile Presenter Training: Demonstration Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/http://doi.org/10.65109/QBWH4657

1 INTRODUCTION

Presentations are inherently multimodal: fluent speech, expressive gestures, stable gaze, and facial feedback all contribute to perceived quality. Automated coaching tools promise frequent, affordable feedback, but must contend with noisy environments, diverse user behavior, and the need to deliver actionable insight to non-expert users. Traditional monolithic systems struggle to evolve when individual modules need replacement or retraining [3, 16, 27]. Various tools exist to train presentation skills [22, 24], but most have limitations. Many focus only on visual aspects and require specialized hardware, such as Kinect [23], or computationally heavy models that are unsuitable for mobile devices. Others [13] analyze speech—tracking filler words, pacing, or emotion, but often ignore visual cues or other key factors like background noise, speech clarity, or speaker emotion. Systems such as Microsoft's Speaker Coach [5] and RoboCOP [26] offer partial solutions, while OpenOPAF [11] uses a multimodal approach but demands professional-grade equipment. Crucially, existing tools typically provide only a single overall score for an entire video, missing finer-grained feedback for individual segments, and rely on generic models that may not perform well for non-native speakers or those with diverse visual appearances.

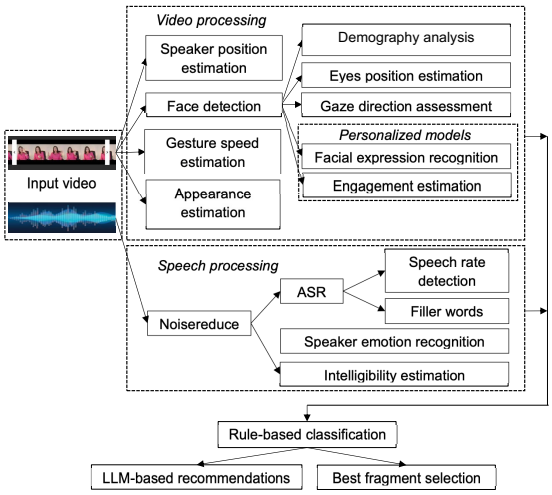

Figure 1: Architecture of SpeakerAICoach framework.

In this paper, we introduce SpeakerAICoach (Fig. 1), which reframes this task through a multi-agent systems (MAS) lens: each analytical function runs as an autonomous agent that processes fragments of the recorded talk and emits structured observations. This modular decomposition brings MAS virtues, agent autonomy, parallel execution, and clear responsibility boundaries to the task of multimodal behavior analysis. Mobile clients record sessions and invoke the agents in the backend; analysis is then aligned, aggregated, and presented to the user in an interactive interface that supports rapid iteration and self-improvement. We introduce two state-of-the-art lightweight models, based on the EfficientNet-B0 and MobileFaceNet architectures, for age, gender, and ethnicity estimation. Their usage for speaker demography analysis lets us provide personalized recommendations with LLMs by feeding the identified visual and acoustic presentation patterns. The source code[1] and demo video for our mobile tool[2] are publicly available.

2 PROPOSED SYSTEM

SpeakerAICoach is composed of a collection of autonomous agents that execute in parallel on segmented media. Agents interact indirectly by writing timestamped annotations to a shared fragment

[1]https://github.com/av-savchenko/Speaker-Trainer
[2]https://youtu.be/zKpVFS8b7d8

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/http://doi.org/10.65109/QBWH4657

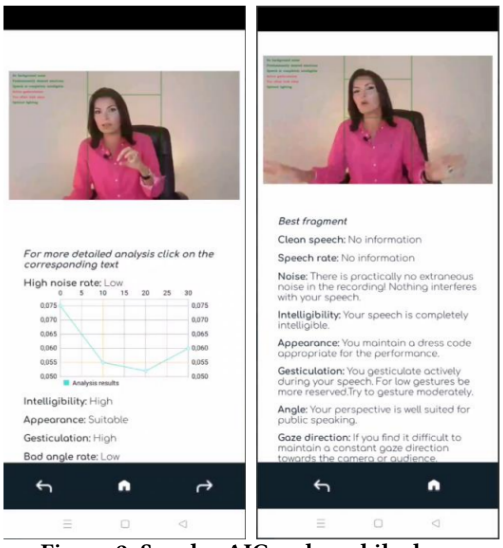

Figure 2: SpeakerAICoach mobile demo

timeline (1–2 seconds), allowing disparate modalities to be jointly interpreted. Each speech and vision processing agent encapsulates a learned model or deterministic pipeline. The design deliberately isolates modalities, allowing for the independent deployment of replacements (e.g., a new emotion model).

Audio and Speech Agents. The audio preprocessing agent reduces environmental noise using Noisereduce library [17] and computes quality indicators such as Short-Time Objective Intelligibility (STOI) index [7]. An automatic speech recognition (ASR) agent performs transcription with word-level timing, enabling the counting of filler words and the calculation of speaking rates. ASR models such as Whisper have been shown to generalize across accents and noisy conditions [14]. A speech emotion recognition agent uses the Aniemore library to classify vocal affect over each fragment, providing soft confidence scores that are later fused with facial affect estimates. Recognizing expressive intent in speech is known to improve behavior analysis.

Vision Agents. A pose and gesture agent tracks body keypoints and derives numerical descriptors of gesture amplitude, speed, and repetitiveness. By quantifying dynamics, the system can distinguish between engaging gesture use and distracting pacing. We analyze the low-level quality of gestures [8], such as the gesture speed [6]. A separate appearance agent (ResNet-34 model on the deepFashion dataset [10]) analyzes clothing and style, which can influence perceived professionalism in video-based presentations.

Facial analysis is handled by dedicated agents that rely on deep feature embeddings and are designed to operate efficiently on pre-extracted frames to predict facial expressions using the models from the EmotiEffLib [20] and demographic attributes such as age and gender. In this paper, we train new models for age prediction, gender recognition, and ethnicity classification. We started with EfficientNet-B0 [25] and MobileFaceNet [2] models trained on the VGGFace2 dataset [1] as described in [21]. Next, we fine-tune them on the Layer Age Gender Dataset (LAGENDA) [9] for simultaneous prediction of age and gender. The multi-task loss is used: we minimize the sum of weighted cross-entropy for age estimation and cross-entropy for gender recognition. Table 1 shows the accuracy

Table 1: Age/gender recognition results on UTK test set

Model	Gender accuracy (%)	Age MAE Argmax	Age MAE EV
DEX [15]	91.05	6.48	-
ResNet-50 (InsightFace) [4]	87.52	8.57	-
MobileNet-v1 [18]	90.09	7.07	-
MiVOLO [9]	92.04	5.55	-
ResNet50, CLAP2016 [12]	-	5.44	-
MobileFaceNet (Ours)	94.25	**5.39**	5.24
EfficientNet-B0 (Ours)	**94.65**	5.53	**4.96**

of gender recognition and MAE (Mean Absolute Error) for age estimation using the official test set from the UTKFace [28] dataset. We tested two options, if possible: obtaining the most probable age (using argmax for the model's outputs or direct age prediction for regression models) and computing the Expected Value (EV) for age class posterior probabilities.

To support ethnicity classification, we train a linear support vector machine classifier on top of facial embeddings extracted by the age/gender models on the ethnicity data from the UTKFace [28]. We achieved a test accuracy of 85.43% and a recall of 77.7%, which is comparable to the best-known competitors in similar settings [19]. The demographic models are converted to ONNX format to improve CPU inference speed.

Aggregation and Language Feedback. The aggregation agent aligns all fragment annotations along the timeline, normalizes scores, and computes high-level metrics such as average engagement and consistency of behavior. This structured representation feeds into a feedback LLM agent, which synthesizes coaching suggestions using an LLM that consumes structured summaries rather than raw signals, thereby preserving interpretability. To accommodate cases of varied presentation styles across different contexts and cultures, our framework includes algorithms for determining a person's facial attributes, such as ethnicity, gender, and age. These parameters are used to personalize the LLM prompt and make recommendations for improving the presenting style and presentation.

Mobile Demonstration. The speaker interacts with our tool through a mobile client (Fig. 2). The user records a presentation on an Android device. All analytical server agents begin processing immediately, producing fragment annotations in parallel. Progress is reflected on the mobile interface, which visualizes the expected processing time based on the agent's load.

Once processing is complete, the interface displays a timeline view with superimposed color-coded indicators from primary modalities. Users can scrub through the timeline; selecting a fragment plays back the corresponding video and displays aligned speech and gesture scores. Coaching recommendations are accompanied by contextual evidence, grounding suggestions in observable behavior.

3 CONCLUSION

Our SpeakerAICoach applies multi-agent principles to multimodal presentation coaching, using autonomous vision and speech modules for interpretable, fragment-level feedback and personalized assistance within a mobile workflow suitable for iterative practice. The Android demo showcases pipeline extensibility, parallel processing, and interactive visualization. The user can change the LLM to the model that is most suitable for their cultural context. Hence, our framework is suitable for all categories of users, regardless of their age, profession, or skill level.

ACKNOWLEDGMENTS

The work of A. Savchenko was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

REFERENCES

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th international conference on automatic face & gesture recognition (FG)*. IEEE, 67–74.

[2] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Chinese conference on biometric recognition*. Springer, 428–438.

[3] Can Cemal Cingi, Nuray Bayar Muluk, and Cemal Cingi. 2023. What Leads to Success in Presenting? Consider the Audience, Subject and Time You Have Available. In *Improving Online Presentations: A Guide for Healthcare Professionals*. Springer, 1–22.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[5] Delara Forghani, Moojan Ghafurian, Samira Rasouli, Chrystopher L Nehaniv, and Kerstin Dautenhahn. 2024. Evaluating people's perceptions of an agent as a public speaking coach. *Paladyn* 15, 1 (2024), 20240004.

[6] Zhong Zhu Huang, Zhi Quan Feng, Na Na He, and Xue Wen Yang. 2015. Research on Gesture Speed Estimation Model in 3D Interactive Interface. *Applied Mechanics and Materials* 713 (2015), 1847–1850.

[7] Jesper Jensen and Cees H. Taal. 2016. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24, 11 (2016), 2009–2022.

[8] P. Khajornphaiboon and S. Vungthong. 2019. Analyzing the introduction of TED Talks: A corpus–based analysis of discourse organization. *Journal of Humanities* 41 (2019), 52–70.

[9] Maksim Kuprashevich and Irina Tolstykh. 2023. MiVOLO: Multi-input transformer for age and gender estimation. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*. Springer, 212–226.

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[11] Xavier Ochoa and Heru Zhao. 2024. OpenOPAF: An Open-Source Multimodal System for Automated Feedback for Oral Presentations. *Journal of Learning Analytics* 11, 3 (2024), 224–248.

[12] Jakub Paplhám, Vojt Franc, et al. 2024. A call to reflect on evaluation practices for age estimation: comparative analysis of the state-of-the-art and a unified benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1196–1205.

[13] Daniil Pliushchenko and Mark Zaslavskiy. 2021. Public Speaking Web Trainer. *Conference of Open Innovations Association, FRUCT* 29 (2021), 485–490.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[15] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. 10–15.

[16] Heleen Rutjes, Martijn C Willemsen, and Wijnand A IJsselsteijn. 2019. Beyond behavior: the coach's perspective on technology in health coaching. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

[17] Tim Sainburg and Timothy Q. Gentner. 2021. Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions. *Frontiers in Behavioral Neuroscience* 15 (2021).

[18] Andrey V Savchenko. 2019. Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet. *PeerJ Computer Science* 5 (2019), e197.

[19] Andrey V Savchenko. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 119–124.

[20] Andrey V Savchenko. 2022. HSEmotion: High-speed emotion recognition library. *Software Impacts* 14 (2022), 100433.

[21] Andrey V Savchenko. 2024. Leveraging pre-trained multi-task deep models for trustworthy facial analysis in affective behaviour analysis in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 4703–4712.

[22] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2015. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. 539–546.

[23] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2016. Enhancing public speaking skills-an evaluation of the Presentation Trainer in the wild. In *Adaptive and Adaptable Learning: Proceedings of the 11th European Conference on Technology Enhanced Learning (EC-TE)*. Springer, 263–276.

[24] Shefaly Shorey, Emily Ang, John Yap, Esperanza Debby Ng, Siew Tiang Lau, and Chee Kong Chui. 2019. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. *Journal of medical Internet research* 21, 10 (2019), e14658.

[25] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[26] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. Robocop: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–24.

[27] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[28] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5810–5818.