Speech Improvement by Multimodal Analysis

Kazunori Minetaki[1] and I-Hsien Ting[2]

[1] Creative Management and Innovation Research Institute, and the Department of Business, Kindai University,3-4-1 Kowakae Higashi-Osaka City, Osaka Prefecture, Japan
[2] Social Networks Innovation Center, National University of Kaohsiung
700, Kaohsiung University Rd, Nanzih District, Kaohsiung 811, Taiwan

Abstract. This study examines whether targeted feedback can lead to improvements in nonverbal communication, specifically in speech delivery, by analyzing 3-minute videos of three management consultants. The analysis focused on facial expressions, utilizing the Facial Action Coding System (FACS) and vocal characteristics, as represented by Mel-Frequency Cepstral Coefficients (MFCCs). For five months, the authors provided individualized feedback focusing on facial expression and speech delivery, which involves key elements such as speaking pace, intonation, and pitch variation. While the sample size was limited to three individuals, the study leveraged a year's worth of accumulated video data. The findings indicate that feedback contributed to observable improvements in facial expressions, whereas vocal features derived from MFCCs remained largely unaffected. Furthermore, this study suggests that combining FACS with real-time speech content extraction offers a deeper understanding of how message content and emotional expression interact, and using the T5 model to generate feedback on speech content automatically improves the speech.

Keywords: FACS, MFCC, Multimodal

1 INTRODUCTION

In the contemporary workplace, nonverbal communication plays a pivotal role in shaping perceptions of professionalism, credibility, and emotional intelligence. Particularly in business and consulting contexts, the ability to convey confidence, enthusiasm, and approachability through facial expressions and vocal tone is often as critical as the content of the spoken message itself. As organizations increasingly prioritize soft skills in leadership development and client engagement, there is a growing need for objective tools to assess such nonverbal cues.

Recent advancements in affective computing have enabled the quantification of facial and vocal signals through technologies such as facial expression recognition (FER) and speech emotion recognition (SER). These methods facilitate the systematic analysis of emotional and interpersonal dynamics in communication settings, providing novel insights into speaker behavior and audience reception. However, much of the existing research has focused on emotional states such as anger, sadness, or happiness in everyday or clinical scenarios, leaving a gap in understanding how professional traits—such as confidence and cheerfulness—are expressed and perceived in formal workplace contexts.

This study addresses this gap by focusing on short, structured speech performances delivered by newly hired management consultants during their onboarding period. The goal is to examine how key expressive traits relevant to professional communication—namely smile, cheerfulness, and confidence—manifest through facial muscle movements and vocal features. Utilizing video-recorded speech data, we apply a multimodal analysis framework that integrates Action Units (AUs) from the Facial Action Coding System (FACS) and Mel-Frequency Cepstral Coefficients (MFCCs) to extract and quantify relevant signals. By developing AU-based composite indicators for each expressive trait, the study aims to contribute to a deeper understanding of nonverbal performance in early-career professionals and to inform future applications in training, feedback, and human-centered AI systems.

Mel-Frequency Cepstral Coefficients (MFCC) have long been a cornerstone in the field of audio signal processing, particularly within the domain of speech recognition. This feature extraction technique derives inspiration from human auditory perception by applying a mel-scale filter to spectral components, thus emphasizing frequencies most relevant to speech analysis. The computational process of MFCC entails several critical steps: pre-emphasis filtering to enhance high-frequency information, framing and windowing to mitigate spectral leakage, Fourier transformation to obtain frequency-domain representations, and mel-scale filtering followed by logarithmic compression to approximate perceptual loudness characteristics. Finally, the application of the Discrete Cosine Transform (DCT) ensures a compact representation of cepstral coefficients, facilitating effective pattern recognition in speech-related tasks.

The utility of MFCC extends beyond traditional speech and speaker recognition. It has found applications in fields such as emotion classification, biomedical signal analysis, and even environmental sound recognition, where spectral features are paramount. However, the advancement of artificial intelligence has increasingly necessitated the integration of multiple modalities to enhance interpretative accuracy. This brings us to the significance of multimodal analysis, an approach that synergistically combines information from diverse sources such as audio, visual, and textual inputs. By leveraging multimodal processing, systems can achieve superior robustness, mitigating the risks associated with noise and incomplete data. Additionally, the incorporation of multiple sensory streams into computational models mirrors human cognitive processes, thereby enabling machines to attain a more holistic understanding of context. Incorporating MFCC within a multimodal framework allows for the expansion of its applicability, particularly in human-computer interaction and affective computing. By fusing audio-derived MFCC features with visual cues such as facial expressions and physiological signals, researchers can develop more sophisticated models for emotion recognition and behavioral analysis. This multimodal approach enhances the interpretability and reliability of automated systems, ensuring that decisions are based on a comprehensive assessment of available data rather than isolated feature sets. Ultimately, the convergence of MFCC with multimodal methodologies represents a pivotal step towards achieving a richer and more nuanced understanding of human communication, thereby advancing the frontier of artificial intelligence in perceptual computing.

Over approximately one year, around ten 3-minute speech videos per individual were analyzed. Between late December 2024 and May 2025, the author provided individualized feedback to each consultant after reviewing their videos to support improvement. The primary objective of this paper is to evaluate the effects of these feedback sessions.

2 RELATED LITERATURE

Facial expression recognition (FER) has undergone significant evolution with advancements in deep learning methodologies. Early static FER systems predominantly relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), which provided a foundation for facial emotion analysis but lacked robustness in complex real-world scenarios. [1].

The advent of deep learning, particularly CNNs and Transformer-based models, has significantly enhanced facial expression recognition (FER) by enabling automatic extraction of spatial-temporal features. While benchmark datasets like AffectNet and RAF-DB support robust evaluation, challenges persist in terms of label consistency and demographic representation. FER applications now extend across healthcare, education, and HCI, with increasing attention to ethical and privacy concerns [2].

Deep learning methodologies, including CNNs and Transformer-based models, have advanced facial expression recognition (FER) by enabling robust spatial-temporal feature extraction, addressing challenges in real-world scenarios like demographic biases and label consistency, and expanding applications in healthcare, education, and human-computer interaction (HCI) while emphasizing ethical and privacy concerns [3].

Recent advances have also examined the role of facial expressions in public speaking and presentations. Zeng et al. [4] developed *EmoCo*, an interactive visual analytics system designed to evaluate emotion coherence across facial expressions, speech, and textual content in presentation videos. By analyzing TED talks, the system visualizes the alignment and inconsistency of emotional signals across modalities, helping identify moments where a presenter’s facial expression may contradict their verbal tone or semantic content. This approach provides valuable insights for improving nonverbal communication and emotional delivery in presentation settings.

Speech emotion recognition (SER) has evolved significantly with the advancement of deep learning. Early SER systems primarily relied on handcrafted acoustic features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and energy, which provided a foundation for emotion classification but lacked robustness in diverse conditions [5].

The introduction of deep neural architectures, particularly convolutional and recurrent neural networks, has led to substantial performance improvements. For instance, Fayek et al. [6] systematically evaluated various deep learning models and demonstrated that long short-term memory (LSTM) networks performed best for modeling emotional dynamics in speech.

Simultaneously, interest has grown in combining SER with facial expression recognition (FER) in multimodal emotion recognition systems. Dhall et al. [7] provided one of the earliest and most widely used benchmarks for evaluating audio-visual emotion recognition under real-world conditions. The EmotiW challenge highlights the advantages of integrating visual and acoustic modalities, particularly in spontaneous and noisy environments. Poria et al. [8] offer a comprehensive review of multimodal affective computing methods, categorizing approaches into early fusion, late fusion, and hybrid models.

Abdelwahab and Busso [9] explored the application of domain adversarial training methods to address challenges in speech emotion recognition (SER) when dealing with varying corpora. Their innovative approach aimed to mitigate the impact of discrepancies between training and testing data distributions, thereby enhancing the model's robustness across different datasets and improving overall SER performance.

Recent multimodal studies have demonstrated the effectiveness of integrating facial expression recognition (FER) with speech features, including MFCC, for emotion detection. For instance, Mamieva et al. [10] proposed an attention-based fusion model that combines independently extracted facial and speech features—including MFCCs—to enhance performance on datasets such as IEMOCAP and CMU-MOSEI. Their framework significantly outperformed unimodal baselines in recognizing spontaneous emotions in real-world recording conditions.

3 DATA ANALYSIS

From May 2024 to May 2025, three newly hired management consultants each delivered a three-minute speech, and video recordings of these speeches were collected from a consulting company. The author and consultants have a meeting every month from December 2024 to May 2025 to improve their speeches, focusing on facial expression and speech delivery, which involves key elements such as speaking pace, intonation, and pitch variation.

The present study is based on this dataset. For the analysis, facial expression features were extracted using Action Units (AUs) defined in the Facial Action Coding System (FACS), in combination with Mel-Frequency Cepstral Coefficients (MFCCs) for vocal analysis. AUs represent specific facial muscle movements and have been widely utilized in psychological and behavioral research as objective indicators of emotional and cognitive states. In this study, three expressive states—smile, cheerfulness, and confidence—were quantified through composite indicators constructed from AU combinations.

Smile was assessed using AU06 (cheek raiser) and AU12 (lip corner puller), which are both recognized as key components of a genuine, so-called Duchenne smile. Cheerfulness was captured by including AU25 (lips part) alongside AU06 and AU12, reflecting a more open and animated facial expression that is commonly associated with high energy and sociability. Confidence was evaluated using a broader set of AUs, specifically AU05 (upper lid raiser), AU06, AU07 (lid tightener), AU12, and AU23 (lip tightener). This combination was chosen to reflect a composed and focused facial configuration, conveying attentiveness, assertiveness, and emotional control—qualities typically attributed to confident individuals in professional contexts. By operationalizing these expressive states through AU-based metrics, this study aims to quantitatively assess nonverbal behavioral cues observed during speech, contributing to a deeper understanding of interpersonal communication in workplace settings.

This study presents an analytical framework for extracting and interpreting acoustic features from speech signals using Mel-Frequency Cepstral Coefficients (MFCC), a technique widely employed in speech and audio processing. The implemented methodology involves extracting audio from an MP4 file and converting it to WAV format, followed by the segmentation of the speech signal into 10-second intervals to facilitate a temporal analysis. The MFCC computation is performed with a dimensionality of 13 coefficients, which capture different spectral aspects of the signal, with higher weights assigned to coefficients MFCC2 to MFCC5 due to their correlation with speech expressiveness and emotional tone. The rationale for selecting 13 MFCC coefficients is grounded in their ability to represent the perceptual and articulatory characteristics of speech, wherein lower-order coefficients encapsulate broad spectral envelopes related to vocal resonance. In contrast, higher-order coefficients capture finer details of articulation. By leveraging a multimodal analytical framework that integrates auditory information with additional contextual cues, this approach aims to enhance the interpretability of expressive speech characteristics, paving the way for advanced applications in affective computing, human-computer interaction, and automated emotion recognition systems. The findings underscore the significance of employing weighted MFCC features for expressive speech analysis and demonstrate the potential of multimodal methodologies in refining the accuracy and robustness of computational models tasked with emotion and expressiveness detection in speech processing domains.

Fig. 1-9 illustrates the results of facial expression. Fig.1-3, 4-6, and 7-9 are the results of each participant. For ID1, both the smile score and cheerfulness score were initially at high levels; however, these scores declined over time, possibly due to stagnation in sales performance. In response, the author provided monthly feedback aimed at improvement. The effects of these interventions became evident in the speeches delivered in March and May 2025, where both scores returned to high levels. The confidence score for ID1 had been consistently high, and it remained stably elevated in March and May 2025. Notably, ID1 commented that they had become more conscious of their facial expressions by observing themselves in the mirror. Both ID2 and ID3 consistently demonstrated high levels of smile and cheerfulness scores. Although ID2's confidence score was initially moderate, it reached its highest level in February 2025. According to ID2, colleagues from other departments commented positively on their improved facial expressions, and ID2 feels more confident. ID3 exhibited consistently high confidence scores, which are presumed to stem from inherent personal traits rather than the direct effects of the author's interventions.

To statistically examine whether there were differences in median values among ID1, ID2, and ID3, the Kruskal-Wallis test was applied to the smile score, cheerfulness score, and confidence score. The results indicated that the median score for ID1 was lower than that for ID2 and ID3. The differences in medians were statistically significant for both the smile score ($p < .01$) and the cheerfulness score ($p < .05$). Regarding the confidence score, the median for ID2 was significantly lower compared to ID1 and

ID3. The Kruskal-Wallis test confirmed that this difference was statistically significant ($p < .05$). An analysis over the entire period revealed that, compared to ID2 and ID3, ID1 tended to display a more rigid facial expression during speeches; however, it did not convey a lack of confidence.

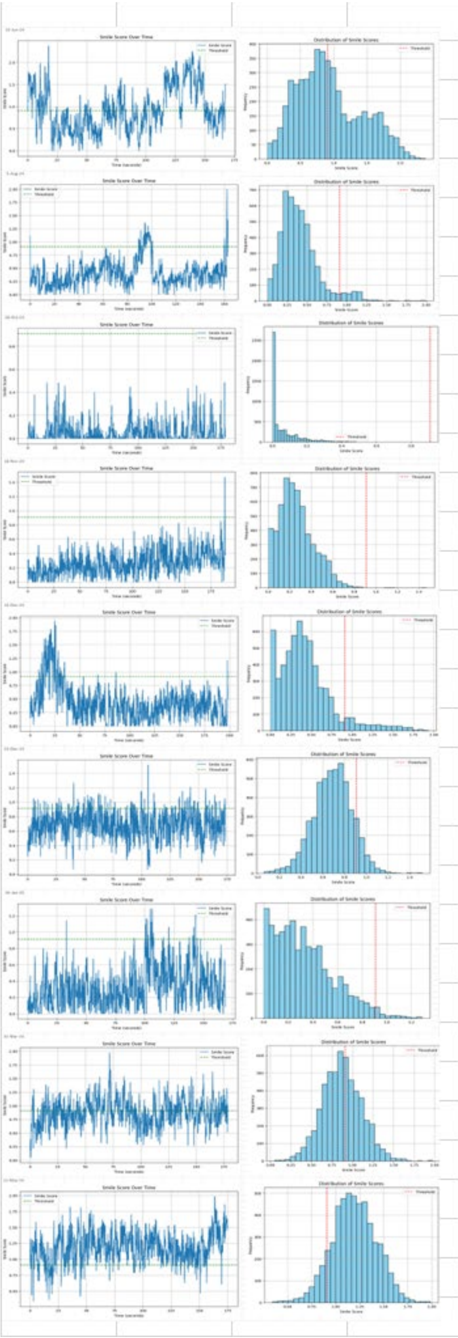
Fig.10-12 describes the results of MFCC. In contrast, MFCC-based vocal features did not show noticeable differences for any of the three participants when comparing the period after January 2025 to the earlier recordings. Thus, the effects of the author's feedback aimed at improving speech delivery were not observable in the acoustic domain.

4 DISCUSSION

Although analyses involving AU6, AU12, and AU25 have been reported in previous studies, the confidence score in this study was computed by integrating the specific characteristics of each action unit. The combination of AU05, AU06, AU07, AU12, and AU23 is generally associated with facial expressions that reflect heightened attention, emotional intensity, or controlled affect. Specifically, AU05 and AU07 are indicative of increased cognitive load or focused attention, and AU06 and AU12 are commonly linked to genuine smiling and social engagement. At the same time, AU23 suggests lip tension or suppression, implying emotional restraint or composure. As observed in previous studies, future analyses will focus on classifying confident facial expressions using deep learning techniques applied to image data.

MFCC can be applied in emotion detection, such as emotional agitation. For speech data analysis, we will also measure speaking rate and intonation patterns, aiming to provide feedback for improving speech delivery.

The authors recognize the need to increase the number of participants; however, gaining the cooperation of partner companies has proven to be challenging. In May 2025, a new onboarding program began, and three newly hired consultants will receive feedback over the course of one year. As a new initiative, an attempt is currently underway to convert speech data into text for further analysis. The following section introduces a part of this initiative.



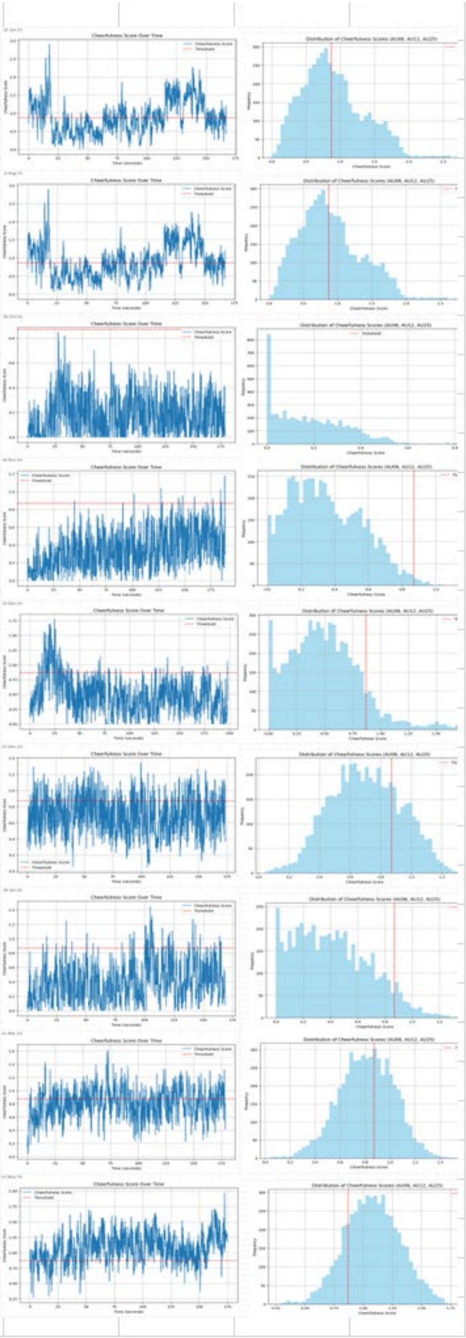
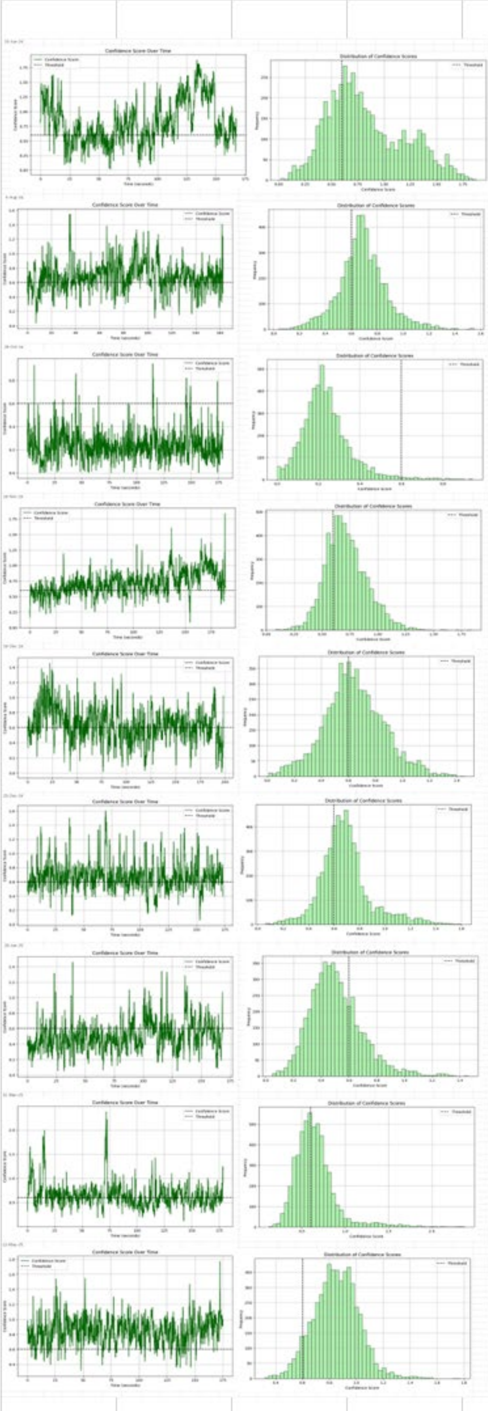


Fig.1 Smile score (ID1)

Fig.2 Cheerfulness score (ID1)



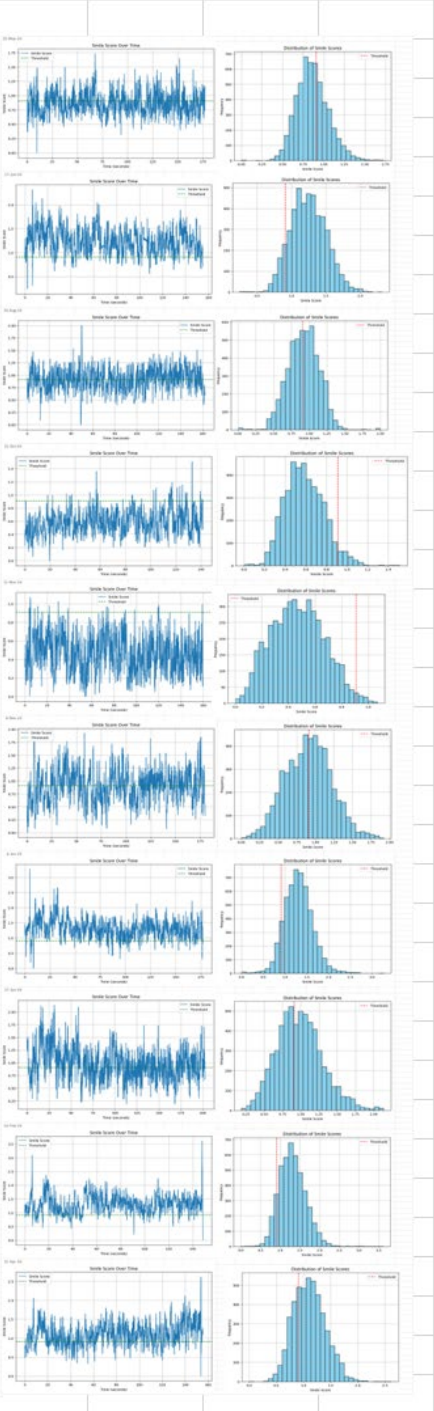
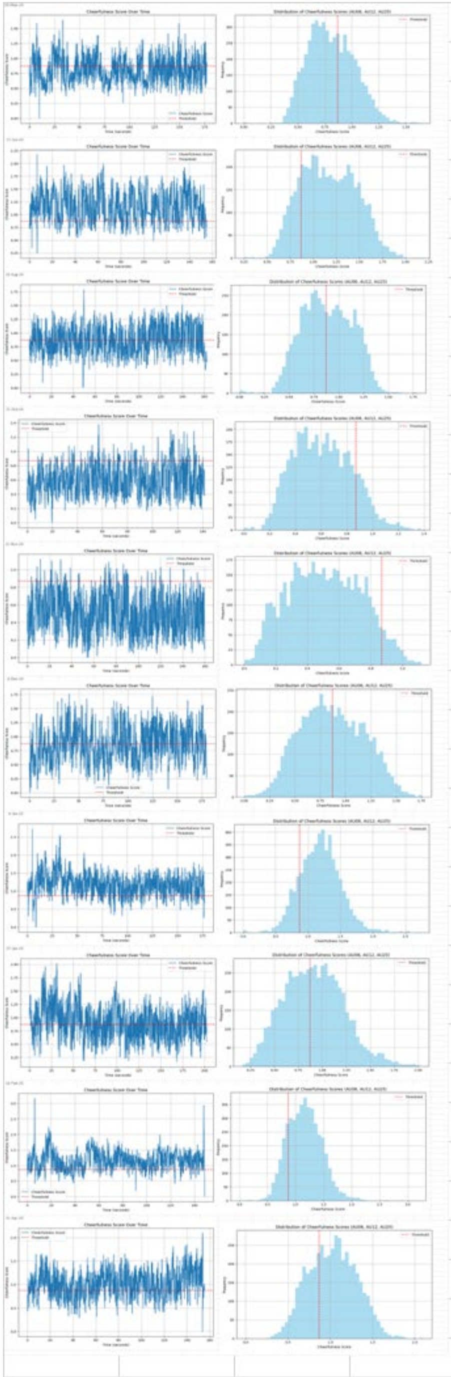


Fig.3 Confidence score (ID1)

Fig.4 Smile score (ID2)



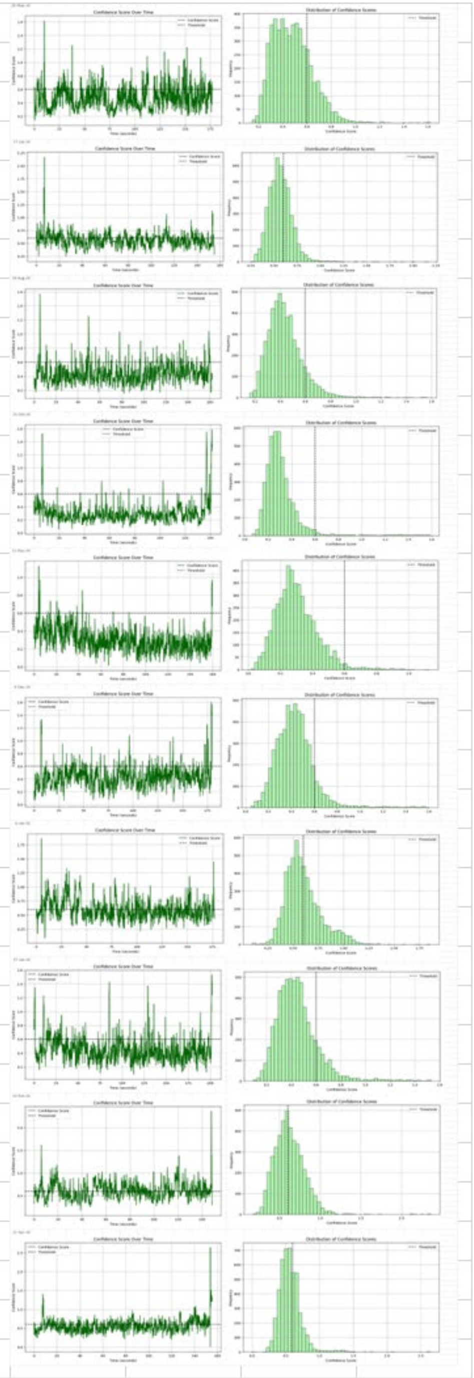
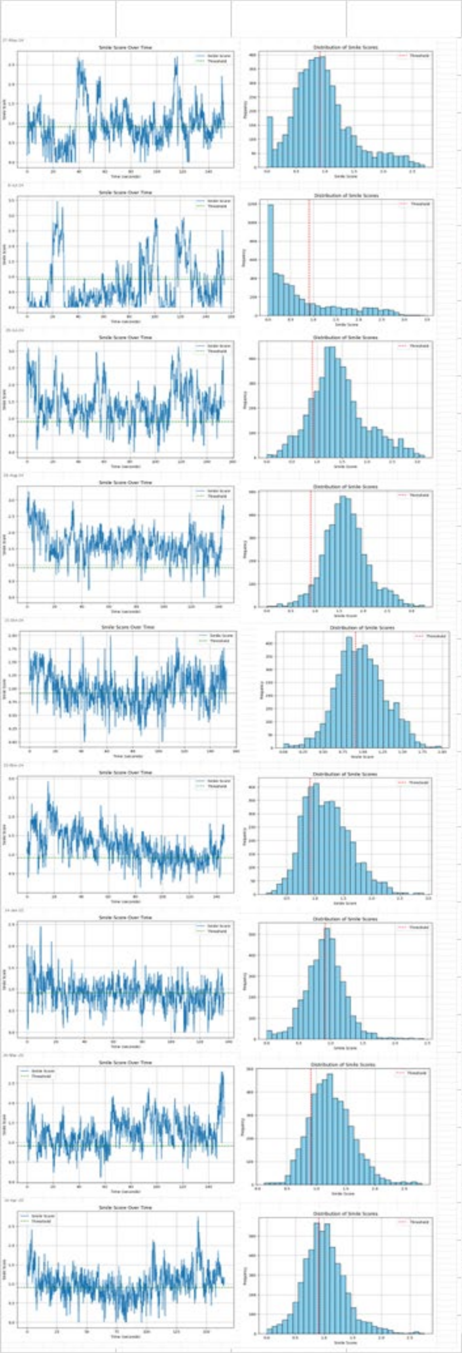


Fig.5 Cheerfulness score (ID2)

Fig.6 Confidence score (ID2)



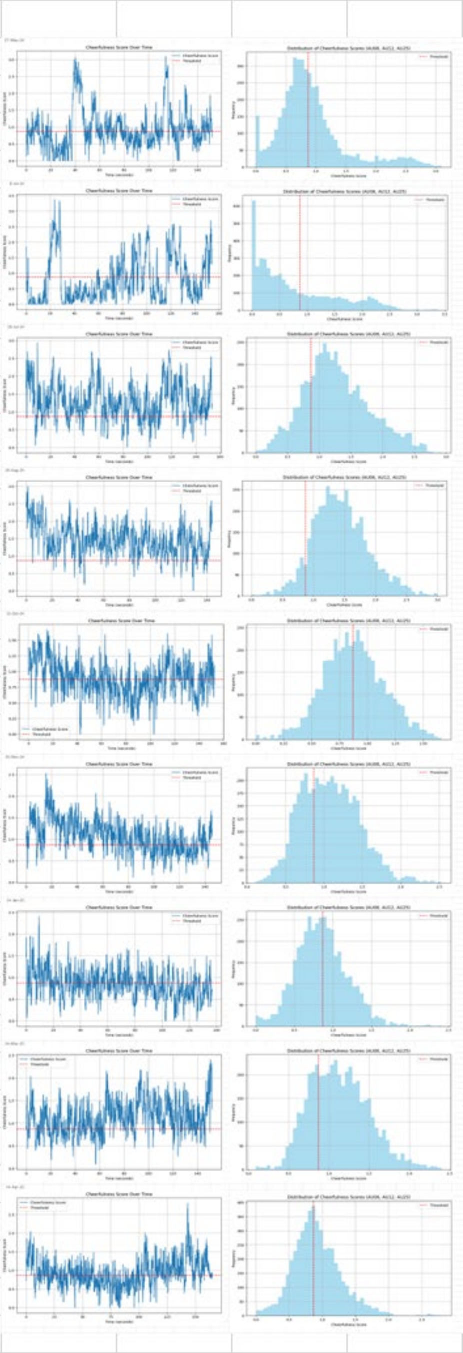
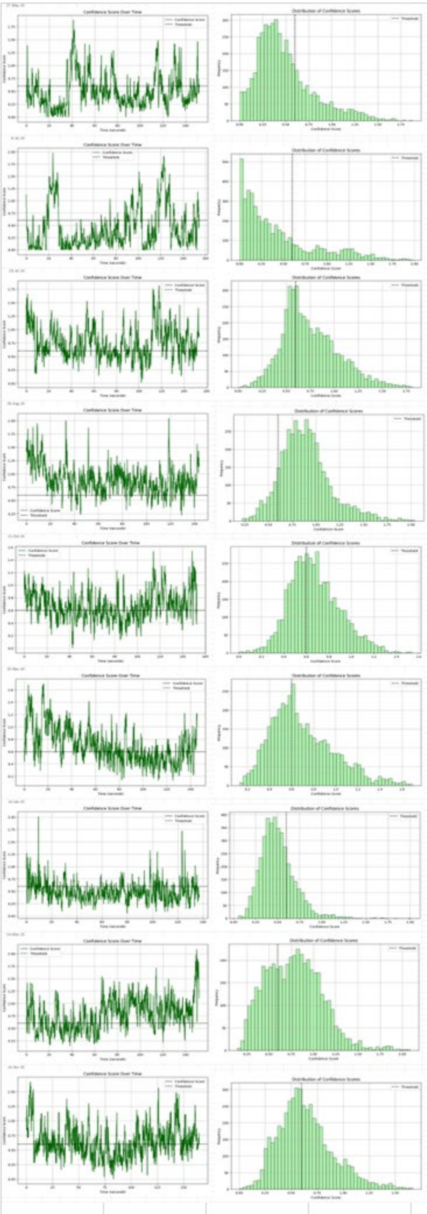


Fig.7 Smile score(ID3)

Fig.8 Cheerfulness score (ID3)



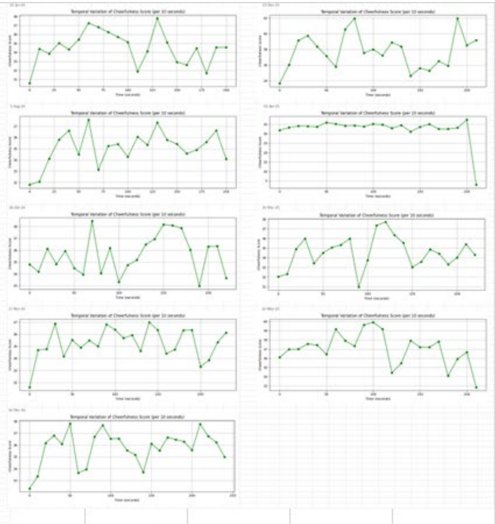


Fig.10 MFCC(ID1)

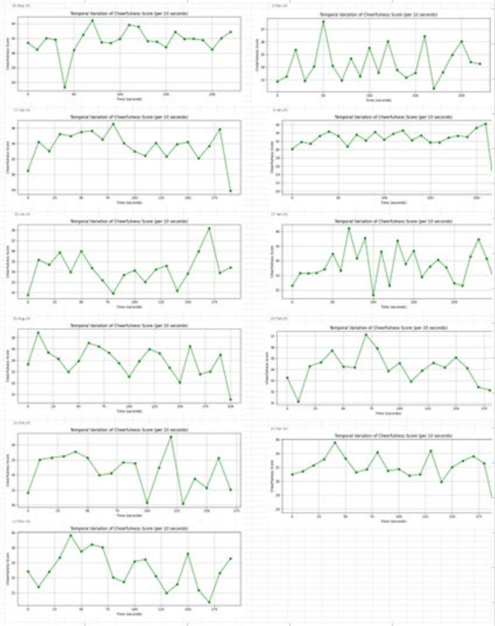


Fig9 Confidence score(ID3)

Fig.11 MFCC(ID2)

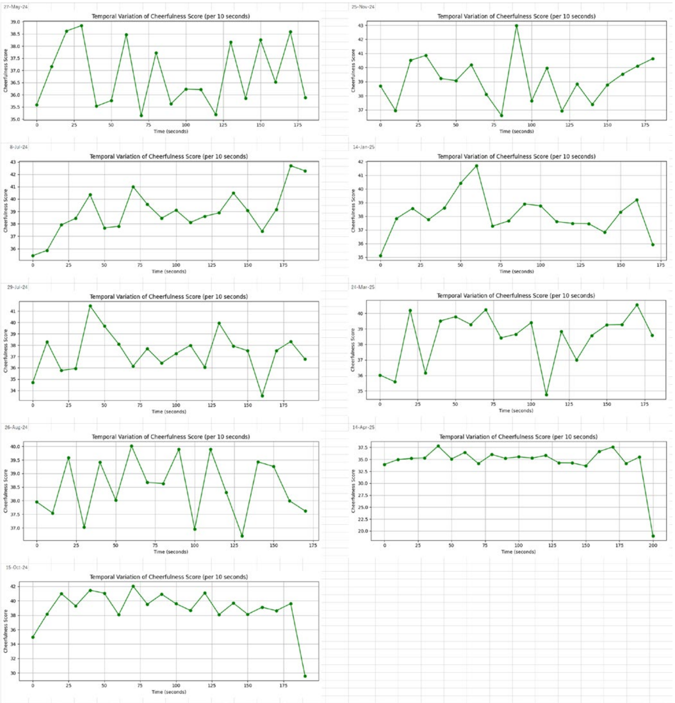


Fig.12 MFCC(ID3)

Table 1: Kruskal-Wallis Test Result: Facial expression

	smile score	cheerfulness score	confidence score
H statistic	9.6402	8.7916	8.5836
p-value	0.0081	0.0123	0.0137

5 FURTHER STUDY

Currently, authors are conducting an experimental approach that combines speech-to-text conversion with facial expression analysis. An example of this is shown in Fig. 13, where the content of speech is automatically displayed for 10 seconds during moments when the cheerfulness score is elevated. The displayed speech content appears to convey an uplifting message, suggesting that facial expressions may change depending on the message being conveyed. This approach has been in practice since May 2025, providing feedback on what types of topics should be discussed at length to improve facial expression conveyance.

Providing feedback to improve the content of speeches is also crucial. As a new initiative, Authors have begun using T5(Text-to-Text Transfer Transformer) to automatically generate feedback comments based on the text data of speeches. An example of this can be seen in Table 2, where the identified issue points to the abstract nature of the content. The T5 model has been trained, using a paired dataset consisting of past speeches and corresponding comments written by the authors. The trained model is now used to automatically generate feedback for the speech delivered by this year's new consultants.

"I went to the event mainly because my favorite artist was performing, but I actually got really into the other performances too!"

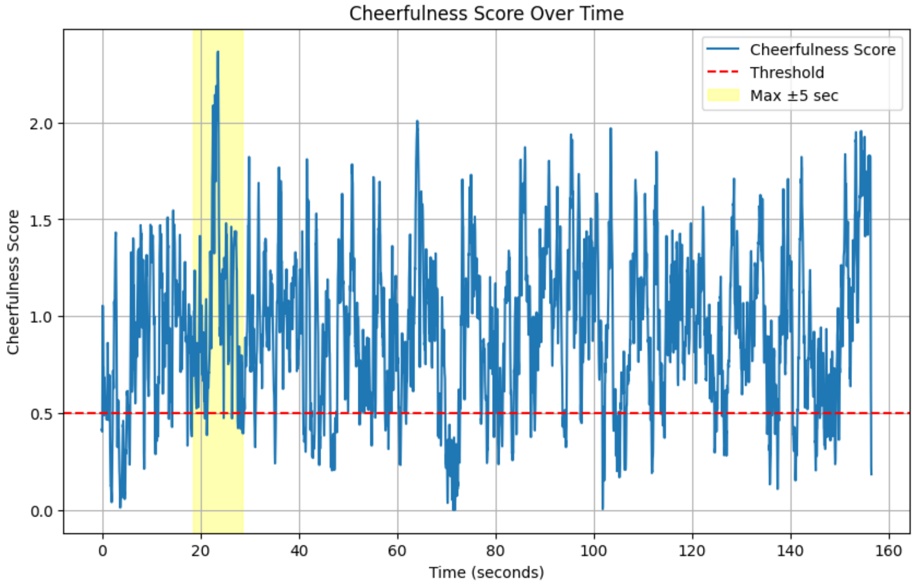


Fig.13 Display of speech content corresponding to peaks in the cheerfulness score

Table2 Feedback to speech by using T5（Text-to-Text Transfer Transformer）

Feedback on the speech of which issue is "professionalism"
This speech resonates strongly as it draws upon the speaker's personal experience, making it highly relatable. The focus on improving work processing speed is particularly valuable, as it directly contributes to enhancing the quality of work as a consultant and strengthening proposal capabilities for clients. The speaker clearly states the theme—"initiatives to increase processing speed"—at the beginning, which gives the speech clarity and direction. However, while the topic itself is valid, the content remains somewhat abstract overall. From the perspective of enhancing client proposal skills, the speech could benefit from more concrete examples or actionable strategies.

6 CONCLUSION

This study explored the integration of facial expression recognition (FER) and speech analysis techniques to assess the nonverbal communication traits—specifically smile, cheerfulness, and confidence—of newly hired management consultants during structured speech performances. By applying a multimodal analysis framework utilizing Action Units (AUs) from the Facial Action Coding System (FACS) and Mel-Frequency Cepstral Coefficients (MFCCs), the study quantified expressive behaviors that are critical to professional impression formation in workplace contexts.

The findings demonstrate that AU-based facial expression indicators, particularly for smile and cheerfulness, were responsive to individualized feedback provided over time, with noticeable improvement in some participants. Confidence levels, operationalized through a composite of multiple AUs reflecting focused and controlled expressions, remained consistently high in certain individuals and appeared less influenced by feedback. In contrast, MFCC-based vocal features did not show significant variation across time or participants, suggesting that the acoustic domain may be less sensitive to short-term feedback interventions focused on expressiveness.

Statistical analysis using the Kruskal-Wallis test confirmed significant differences in facial expression scores across participants, highlighting the utility of AU-based metrics in capturing interindividual variability. The study also introduces an innovative approach combining facial analysis with real-time speech content extraction, offering a deeper understanding of how message content and emotional expression interact.

Furthermore, the study initiated an experimental use of the T5 model to generate feedback on speech content automatically. Trained on past speech-comment pairs, the model offers scalable and consistent evaluations that can complement human feedback in communication training.

Although a small sample size limited the study, it demonstrates the feasibility and value of using multimodal analysis to assess and improve nonverbal communication in professional settings. Future research will expand the participant pool, enhance vocal feature analysis (e.g., speaking rate, intonation), and further develop automated feedback systems. These efforts aim to support the development of emotionally expressive and confident professionals through evidence-based, AI-enhanced feedback tools.

References

1. Wang, Yan, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, Zhongxue Gan.: A survey on facial expression recognition of static and dynamic emotions. arXiv preprint arXiv:2408.15777 (2024).
2. Kopalidis T, Solachidis V, Vretos N, Daras P： Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. *Information* 15(3):135 (2024).
3. S. Li, W. Deng: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, vol. 13, no. 3, 1195-1215(2022)

4. Zeng, H., Wang, X., Wu, A., Wang, Y., Li, Q., Endert, A., Qu, H: EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. IEEE Transactions on Visualization and Computer Graphics, vol. 26, no.1, 927-937(2020).
5. F. Eyben, M. Wöllmer, B. Schuller: open SMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proc. of the 9th ACM International Conference on Multimedia, Florence, Italy, 1459–1462 (2010).
6. Fayek, H. M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60–68(2017).
7. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion Recognition In The Wild Challenge 2013. Proceedings of the 15th ACM International Conference on Multimodal Interaction,509–516(2013).
8. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125 (2017).
9. Abdelwahab, M., Busso, C.: Domain adversarial for acoustic emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12), 2423–2435(2018).
10. Mamieva, D., Abdusalomov, A.B., Kutlimuratov, A., Muminov, B., Whangbo, T.K.: Multimodal emotion detection via attention-based fusion of extracted facial and speech features. Sensors 23(12), 5475 (2023).