

Verbal and Nonverbal Coherence as an Index for Second Language Proficiency

Katsunori Kotani

Kansai Gaidai University
16-1 Nakamiyahigashino-cho, Hirakata,
Osaka, Japan
kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi

Ryukoku University
1-5 Yokotani, Seta Oe-cho, Otsu,
Shiga, Japan
yoshimi@rins.ryukoku.ac.jp

Abstract

Recent multimodal learner corpora incorporate visual information, such as posture, co-speech gestures, and facial expressions, alongside linguistic features. While multimodal information has contributed to modeling second language (L2) proficiency, little attention has been paid to cross-modal relationships. To address this gap, the present study investigates whether affective coherence, the alignment between textual sentiment and facial expressions, is associated with L2 proficiency. The results indicated that affective coherence did not significantly differ across proficiency levels from B1 to C2 under the Common European Framework of Reference for Languages (CEFR). Coherence measured at the valence level (positive, neutral, and negative) exhibited a systematic relationship with proficiency levels, whereas coherence based on discrete categories (e.g., happiness and sadness) did not. These findings suggest that multimodal learner corpora should incorporate not only information from multiple modalities but also cross-modal measures that capture the alignment between verbal and nonverbal communicative behaviors.

1 Introduction

Multimodal corpora have been used in affective computing research to integrate acoustic, semantic, and visual cues for analyzing verbal and nonverbal behaviors (Hyun et al., 2024; Peechatt et al., 2024). For example, laughter detection has been modeled by combining acoustic properties such as pitch and intensity with visual markers including Facial Action Units (AUs), together with semantic information derived from speech transcriptions (Hyun et al., 2024). Similarly, Peechatt et al. (2024) demonstrated that acoustic features such as fundamental frequency F_0 (Hz) can be enhanced through fusion with physiological and visual modalities. In the context of L2 research, nonverbal behaviors have recently been incorporated into analyses

of L2 communication (Hunston, 2022; Liu et al., 2023). While traditional multimodal learner corpora (Franceschi, 2018; Bonsignori and Cappelli, 2021) relied on manual annotation of nonverbal behaviors, recent studies automatically extract visual cues for analysis of L2 communication (Saeki et al., 2021; Mawalim et al., 2025). Consequently, recent multimodal learner corpora have made it possible to analyze L2 communication from both verbal and nonverbal perspectives, including posture, co-speech gestures, and facial expressions, all of which play essential roles in successful communication (Kendon, 2000; Bavelas and Gerwing, 2007; Cohn and Schilperoord, 2024).

Despite these advances of a multimodal learner corpus, a research gap remains regarding the systematic alignment between verbal and nonverbal behaviors. This alignment is conceptually grounded in affective coherence, a phenomenon in which an individual's internal emotional state is synchronized with outward facial expressions (Durán et al., 2017). In L2 communication, affective coherence between semantic content and nonverbal behaviors may contribute to comprehensibility (Gullberg et al., 2008; Wheeler and Saito, 2022). Such synchronization can be viewed as an aspect of pragmatic competence, as proficient learners are expected to display facial expressions consistent with the semantic content of their speech (Wu and Roche, 2025). For example, proficient learners may display smiles in positively valenced narratives. While Wu and Roche (2025) examined L2 arousal states using physiological measurements, facial expressions potentially provide a more accessible representation of both arousal and valence (i.e., positivity and negativity) (Zhang et al., 2024a).

The present study addresses a research gap regarding cross-modal relationships for multimodal learner corpus research. Specifically, it quantifies the alignment between verbal and nonverbal behav-

iors as a cross-modal annotation layer. Among various nonverbal behaviors, facial expressions were selected because they provide observable cues more directly associated with affective states than other cues such as postures and gestures. For the verbal modality, textual sentiment was selected because it directly reflects the semantic and affective content of discourse. In contrast, acoustic cues may be influenced by L2-specific factors such as Foreign Language Anxiety (FLA) (Szyszka et al., 2026; Wang et al., 2025). We hypothesize that affective coherence varies systematically across L2 proficiency levels. This hypothesis is examined through the following research questions:

1. What is the relationship between affective coherence and L2 proficiency?
2. To what extent can affective coherence serve as an index of L2 proficiency levels?
3. Can facial expression diversity serve as another index for L2 proficiency levels?

By addressing these questions, this study offers two primary contributions to computational L2 research. First, contrary to our hypothesis, affective coherence did not significantly differ across proficiency levels. Second, the observed distributions suggest that affective coherence differs between proficiency levels for positive, neutral, and negative valence categories.

2 Previous studies

This section first outlines theoretical perspectives relevant to affective coherence in L2 communication and then reviews three related research areas: (a) affective coherence between emotion and facial expression, (b) multimodal information fusion, and (c) the role of nonverbal behavior in analyzing L2 communication. Although these studies provide foundations for the present research, little attention has been devoted to whether textual sentiment and facial expressions exhibit systematic alignment in L2 communication. Furthermore, the extent to which such alignment varies across proficiency levels remains unclear.

2.1 Theoretical Basis for Affective Coherence in L2 Communication

Cognitive theories of L2 communication suggest that the coordination of verbal and nonverbal behaviors may vary across proficiency levels. In particular, Cognitive Load Theory (Sweller, 2010) posits

that working memory resources are limited during communication. Because L2 learners allocate attentional resources to lexical retrieval, grammatical encoding, and speech monitoring, the coordination of verbal and nonverbal behaviors may be more difficult than in first-language (L1) communication. This account aligns with findings of reduced nonverbal expressivity in L2 communication (Caldwell-Harris, 2014; Pavlenko, 2012).

2.2 Emotion–Facial Expression Coherence in L1 Communication

A useful baseline for examining affective coherence in L2 communication can be found in studies of L1 communication. Durán et al. (2017) conducted a meta-analysis of affective coherence between emotional experiences and facial expressions. Emotional states were measured through self-report instruments, whereas facial expressions were assessed using the Facial Action Coding System (FACS; Ekman et al., 2002) based on six basic emotions: anger, surprise, disgust, happiness, fear, and sadness (Ekman, 1971).

Durán et al. (2017) observed weak-to-moderate correlations between emotional experiences and facial expressions, ranging from $r = .11$ for fear to $r = .41$ for sadness (Durán et al., 2017). These findings suggest that affective coherence remains limited and variable even among L1 speakers under controlled experimental conditions. This empirical baseline provides a useful reference for examining affective coherence in L2 communication.

2.3 Multimodal Information Fusion

Recent affective-computing research has employed information fusion integrating visual, acoustic, and textual modalities for emotion recognition and sentiment analysis (Wang et al., 2022). According to Wang et al. (2022), most studies have focused on visual–acoustic, textual–acoustic, or visual–acoustic–textual fusion, while relatively little attention has been paid to direct interactions between textual and visual modalities.

This limitation was addressed by Li et al. (2024) with a multimodal fusion framework between textual and visual information. However, the visual modality in their study captured general visual information rather than facial-expression-based communicative signals. More recently, Wang and Chen (2026) developed a sentiment-analysis framework for complex affective phenomena such as sarcasm by integrating textual, acoustic, and visual informa-

tion, including facial expressions. In their framework, visual and acoustic signals function as supplementary semantic contextual cues. For example, sarcastic utterances may be accompanied by facial displays such as angry smiles and acoustic patterns that differ from those associated with genuinely positive expressions. These studies demonstrate the growing importance of cross-modal interactions in multimodal communication research. However, such interactions remain largely unexplored in L2 research.

2.4 Multimodal Model for L2 Proficiency

Although L2 research has recognized the importance of nonverbal behaviors, little attention has been devoted to the alignment between textual sentiment and facial expressions. Recent studies (Mawalim et al., 2025; Saeki et al., 2021) have developed automatic L2 proficiency evaluation systems using nonverbal features. Their findings suggest that visual features provide useful information but exhibit substantially lower explanatory power than lexical and acoustic features when analyzed independently.

Saeki et al. (2021) classified six proficiency levels using lexical, acoustic, and visual features. Visual features included facial expressions, head movements, and eye gaze derived from Facial Action Units (AUs) using OpenFace 2.0 (Baltrušaitis et al., 2018). While their multimodal classification model achieved the strongest correspondence with human ratings ($r = .80$), cross-modal relationships were not explicitly examined. They reported comparatively low performance for the visual model ($r = .03$), compared with the acoustic ($r = .17$) and lexical ($r = .78$) models. The authors suggested that the limited contribution of visual features may reflect substantial individual variation in nonverbal behavior. For example, behaviors such as frowning, gaze shifts, or reduced expressiveness may reflect cognitive processing demands rather than proficiency itself.

Mawalim et al. (2025) classified binary and six-level L2 proficiency classification models using textual, acoustic, and visual features. Visual information was extracted through OpenFace (Amos et al., 2016) and represented by the occurrence frequencies of individual AUs throughout the speaking task. Their results indicated moderate predictive value for visual features. In the binary classification, the visual model achieved an F1 score of 62.87, compared with 81.21 for the acoustic model and 77.72

for the textual model. In the six-level classification task, the visual model yielded a correlation of $\rho = .31$, compared with $\rho = .67$ for the acoustic model and $\rho = .61$ for the textual model. Mawalim et al. (2025) further reported that the contribution of visual information varied across AU types, with lip-related AUs serving as particularly informative indicators. As in Saeki et al. (2021), relationships among modalities were not directly investigated.

3 Method

3.1 Data Collection

We manually compiled 68 video recordings featuring L2 English communication from the YouTube channels that publish the International English Language Testing System (IELTS) speaking test performances and their corresponding band scores, including: IELTS Daily, IELTS Insider, IELTS English Academy. These videos were selected from publicly available content on the YouTube platform. As of March 2026, the source videos remain publicly available on the platform.

To ensure the validity of the facial expression analysis, videos were included only if they exhibited a clear, front-facing view of the speaker's face. This manual selection process allowed for the exclusion of videos with significant background noise and multiple speakers.

3.2 L2 Communication Data

The dataset consisted of video recordings of 68 L2 English learners participating in a mock IELTS speaking test. To ensure data consistency across proficiency levels, this study focused on the Part II monologue task, where learners delivered a one-to two-minute speech on an assigned topic after a one-minute preparation period.

Following the standard IELTS–CEFR alignment (IELTS, 2015), IELTS Speaking scores of 8.5 or above were mapped to CEFR (the Common European Framework of Reference for Languages) C2, scores of 7.0–8.0 to C1, scores of 5.5–6.5 to B2, and scores of 5.0 or below to B1. Table 1 presents descriptive statistics for IELTS Speaking scores by CEFR proficiency group.

3.3 Facial Expression Recognition

Facial expressions were analyzed using Py-Feat 0.7.0 (Cheong et al., 2023). All videos were processed at 25 frames per second (fps) to ensure computational efficiency. The system assigned proba-

Group	<i>n</i>	Mean	SD
C2	17	8.68	0.25
C1	29	7.47	0.42
B2	14	6.04	0.46
B1	8	4.81	0.37

Table 1: IELTS Speaking scores by CEFR proficiency group.

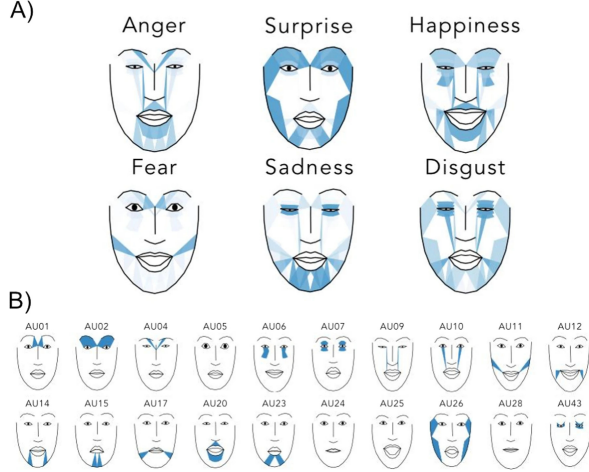


Figure 1: (A) Six basic facial expressions. (B) Individual Facial Action Units (AUs) detected. Reproduced from Cheong et al. (2023) under a CC BY 4.0 license.

bilistic confidence scores for the seven basic emotion categories: anger (ANG), disgust (DIS), fear (FEA), happiness (HAP), sadness (SAD), surprise (SUP), and neutral (NEU). As illustrated in Figure 1, it identifies individual AUs (B) and maps them to six basic facial expressions (A).

Frame-level scores were normalized to sum to 1.00 as shown in Table 2. Table 2 demonstrates sample scores of a learner at frames 0, 1926, 2005, and 3008 of a 120.32-second video (comprising 3,009 frames in total). This learner received a score of 4.0 in the mock IELTS speaking test.

3.4 Textual Sentiment Recognition

Before recognizing textual sentiment, the audio data were transcribed using the Whisper-large-v3 model with default settings (Radford et al., 2022), which is robust to accented speech (Ballier et al., 2023) and generates timestamped segments necessary for multimodal synchronization. Furthermore, unlike traditional transcription methods that may split segments solely based on silence, Whisper identifies meaningful spoken units that better align with facial expression frames. For instance, Whisper retains disfluencies (e.g., “um,” or repetitions;

Mateo et al., 2024), which may reflect cognitive or affective states.

Textual sentiments were recognized using Gemini-2.5-flash-lite (Google, 2024). Unlike traditional sentiment analysis relying on lexical and grammatical precision, this approach identifies textual sentiments even when learners’ utterances exhibit limited vocabulary or grammatical inaccuracies (Zhao et al., 2026; Zhang et al., 2024b).

The prompt used for this recognition is provided in Appendix A. For each transcribed segment, Gemini generated normalized probability scores (ranging from 0.00 to 1.00) across the same seven categories used in facial expression analysis. Table 3 presents sample segment-level textual sentiment scores for two specific utterances of the learner corresponding to Table 2: “He always” (1.74 seconds) and “He always help people or become a hero” (4.04 seconds). Within each segment, the category with the highest score was identified as the dominant category (DOM).

3.5 Alignment between Text Emotion and Facial Expressions

In alignment between textual sentiments and facial expressions, the former serves as a semantic reference for the intended affective content. This alignment establishes affective coherence as an index of L2 proficiency levels.

For the synchronization of the nonverbal and verbal datasets, facial expression data from Py-Feat were mapped to the corresponding timestamped text segments from Whisper (measured in seconds). This alignment was achieved by mapping the video frame indices to the start and end timestamps of each text segment. The total number of frames (N_f) within a specific text segment was determined by calculating the difference between the start and end timestamps (t_{start} and t_{end} , in seconds) as follows:

$$N_f = \text{round}((t_{end} - t_{start}) \times 25) \quad (1)$$

As illustrated in Tables 2 and 3, the utterance “He always” spans from 78.48 to 80.22 seconds, which corresponds to frames 1962 through 2005. The facial expression score for this segment (Segment 11 in Table 4) was calculated by taking the mean score for each of the seven categories across these 43 frames.

Frame	ANG	DIS	FEA	HAP	SAD	SUP	NEU
0	0.00	0.00	0.00	0.99	0.00	0.00	0.01
...	...	...	...	...	...	...	...
1962	0.00	0.00	0.00	0.08	0.00	0.04	0.88
...	...	...	...	...	...	...	...
2005	0.00	0.00	0.00	0.99	0.00	0.01	0.00
...	...	...	...	...	...	...	...
3008	0.00	0.01	0.02	0.68	0.00	0.11	0.19

Table 2: Sample frame-level facial expression scores

Segment	Timestamp (sec.)	Utterance	ANG	DIS	FEA	HAP	SAD	SUP	NEU	DOM
...	...	...	...	...	...	...	...	...	...	...
11	78.48 → 80.22	He always He always	0.00	0.00	0.00	0.00	0.00	0.00	1.00	NEU
12	80.22 → 84.26	help people or become a hero.	0.00	0.00	0.00	0.70	0.00	0.00	0.30	HAP
...	...	...	...	...	...	...	...	...	...	...

Table 3: Sample segment-level textual sentiment scores

Segment	Facial expression score									Utterance	Textual sentiment score								
	ANG	DIS	FEA	HAP	SAD	SUP	NEU	DOM	COS		ANG	DIS	FEA	HAP	SAD	SUP	NEU	DOM	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
11	0.00	0.01	0.00	0.67	0.00	0.01	0.31	HAP	0.42	He always	0.00	0.00	0.00	0.00	0.00	1.00	NEU		
										He always									
12	0.00	0.04	0.00	0.46	0.01	0.07	0.43	HAP	0.93	help people or become a hero.	0.00	0.00	0.70	0.00	0.00	0.30	HAP		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		

Table 4: Sample alignment of textual and facial emotion scores

3.6 Affective Coherence Analysis

Affective coherence was analyzed under two approaches: (i) Polarity-based coherence, and (ii) Category-based coherence. To test polarity-based coherence using Chi-square contingency tests across proficiency groups, the facial expression categories for each segment were classified into three classes: Positive (happiness, surprise), Negative (anger, disgust, fear, sadness), and Neutral. This classification was determined by the dominant emotion within the respective textual sentiment and facial expression categories of each learner. To control for Type I error across multiple pairwise comparisons, a Bonferroni correction was applied, establishing an adjusted significance threshold of $p = .02$. Due to low expected cell counts in specific categories (e.g., negative segments in B1 and B2), Fisher’s exact tests were also reported.

However, labeling based solely on the dominant category may obscure mixed or diffuse emotional states, when multiple facial expressions yield similar scores (see Table 4). In Table 4, while the dominant facial expression in the 11th and 12th segments is identified as Happiness, the one in Segment 12 represents a diffuse state where the score for Happiness (0.46) is nearly identical to Neutral (0.43). These examples indicate the limitations of labeling reduction.

To capture the nuances of mixed or diffuse categories, we calculated the Cosine Similarity between the textual sentiment $\mathbf{T}$ and facial expression $\mathbf{F}$ categories. This metric measures the angular alignment in a 7-dimensional affective space. Cosine Similarity (COS) is calculated as:

$$COS(\mathbf{T}, \mathbf{F}) = \frac{\sum_{i=1}^7 t_i f_i}{\sqrt{\sum_{i=1}^7 t_i^2} \sqrt{\sum_{i=1}^7 f_i^2}} \quad (2)$$

where t_i and f_i represent the scores for each category. COS ranges from 0.00 (no alignment) to 1.00 (perfect coherence).

For the statistical analysis, learners’ mean COS values were classified by proficiency group. The mean COS was calculated for each learner by dividing the sum of COS values by the total number of instances within a learner. These group-level distributions were then compared using an ANOVA (Analysis of Variance) test.

4 Results

4.1 Polarity-based Coherence

Table 5 summarizes the frequencies of facial expressions conditioned on textual emotion and CEFR proficiency. While Neutral facial expressions were the most frequent responses to all textual emotion categories, the C2 group showed the highest proportion of Positive facial expressions in response to Positive textual content (30.6%, $n = 33$ out of 108 segments), compared with the C1 group (21.3%, $n = 17$ out of 80); the B2 group (11.1%, $n = 4$ out of 36); and the B1 group (14.3%, $n = 3$ out of 21).

In contrast, alignment for the Negative textual emotion was comparable between the collapsed groups (C1-2 group: 28.3%, $n = 13$ out of 46; B1-2 group: 25.0%, $n = 2$ out of 8).

Text Emotion	Group	Facial Expression		
		Neg	Neu	Pos
Neu	C2	72	188	113
	C1	41	137	61
	B2	56	95	19
	B1	10	85	20
Pos	C2	15	60	33
	C1	10	53	17
	B2	17	15	4
	B1	0	18	3
Neg	C2	6	12	11
	C1	7	8	2
	B2	2	2	0
	B1	0	4	0

Table 5: Frequency of facial expressions conditioned on textual emotion and proficiency

To examine polarity-based coherence, we conducted three independent Chi-square contingency tests for each textual sentiment category (Negative, Neutral, and Positive).

The results no statistically significant association between textual sentiment polarity and facial expression polarity across any CEFR proficiency level or in the overall aggregated sample: B1 Level: $\chi^2(4, N = 140) = 3.58, p = .47$ (Fisher’s exact $p = .67$). B2 Level: $\chi^2(4, N = 210) = 3.51, p = .48$ (Fisher’s exact $p = .47$). C1 Level: $\chi^2(4, N = 336) = 9.47, p = .05$ (Fisher’s exact $p = .08$). C2 Level: $\chi^2(4, N = 510) = 2.95, p = .57$ (Fisher’s exact $p = .54$). Total Sample: $\chi^2(4, N = 1,196) = 3.73, p = .44$ (Fisher’s exact $p = .43$). Across all proficiency groups,

facial expressions were heavily skewed toward neutral displays regardless of textual sentiment (e.g., in the total sample, neutral facial expressions accounted for 505 out of 897 neutral text segments, 146 out of 245 positive text segments, and 26 out of 54 negative text segments). Although the C1 group exhibited a borderline trend ($\chi^2 = 9.47, p = .05$), the relationship did not reach significance under Fisher's exact test ($p = .08$), confirming the absence of a reliable categorical alignment between text emotion and facial expression.

4.2 Category-based Coherence

Table 6 summarizes the Cosine Similarity between textual sentiment and facial expression categories. This metric reflects the alignment degree between the 7-dimensional category distributions in an utterance.

Group	<i>n</i>	Mean (SD)
C2	17	0.63 (0.19)
C1	29	0.61 (0.21)
B2	14	0.60 (0.27)
B1	8	0.73 (0.15)

Table 6: Cosine Similarity between proficiency groups.

A one-way ANOVA was conducted to examine whether facial–semantic coherence differed across proficiency levels. The analysis showed no statistically significant difference in mean *COS* values across proficiency levels, $F(3, 25.7) = 1.19, p = .33$. Thus, facial–semantic coherence did not significantly differ across proficiency levels.

5 Discussion

The goal of this study was to investigate the extent to which L2 communication demonstrates affective coherence. The results indicate that the degree of affective coherence does not significantly depend on either learner proficiency or the granularity of the cross-modality analysis (polarity-based vs. category-based). Overall, the findings suggest that coherence between textual sentiment and facial expression are associated with the individual differences rather than against the assumption of perfect alignment between textual and visual modalities.

First, polarity-based coherence was analyzed by comparing the distribution of valence polarity (Positive, Neutral, and Negative) across the proficiency groups. The Chi-square tests demonstrate that facial expression polarity does not systematically cor-

respond to textual emotion polarity across CEFR proficiency levels. Across all CEFR tiers, learners maintained neutral facial expressions (677 out of 1,196 total segments; 56.6%), irrespective of whether the uttered segment was positive, negative, or neutral. This relying on neutral facial expressions suggests a general "flatness" in emotional display during L2 speech production.

The lack of statistical significance across all groups (Bonferroni-corrected $p > .02$) indicates the methodological limitations of reducing affective coherence to coarse, segment-level classifications that may include multiple valences. Furthermore, discrete categorical labeling forces subtle, continuous facial movements into rigid categories, which may hide mixed emotional states (e.g., subtle smiles occurring during neutral or negative speech). In sum, these findings demonstrate the necessity of continuous vector-based metrics, such as multidimensional Cosine Similarity, to capture fine-grained affective alignment in L2 communication.

Second, category-based affective coherence was analyzed using a one-way ANOVA of cosine similarity across seven emotional categories (anger, disgust, fear, happiness, sadness, surprise, and neutral) across CEFR proficiency levels. The analysis revealed no significant effect of proficiency level, $F(3, 25.7) = 1.19, p = .33$. This result suggests that category-based affective coherence was not significantly associated with L2 proficiency.

However, the descriptive results revealed an interesting pattern, which we refer to as a "proficiency–coherence paradox." B1 learners showed the highest mean cosine similarity ($M = .73, SD = .15$), followed by C2 ($M = .63, SD = .19$), C1 ($M = .61, SD = .21$), and B2 ($M = .60, SD = .27$). Thus, higher L2 proficiency was not accompanied by greater observable verbal–facial affective coherence. One possible explanation is that more proficient learners may possess greater pragmatic or rhetorical flexibility in managing emotional displays. Hence, they can employ facial expressions more independently of their verbal affect. In contrast, less proficient learners may rely more heavily on facial expressions to compensate for their verbal expression. This reliance is supposed to result in higher apparent affective coherence. These interpretations remain speculative, because the statistical analysis does not support proficiency as a systematic determinant of affective coherence. Rather, the descriptive pattern raises the possibility that the relationship

between proficiency and verbal–facial coherence may be non-linear or mediated by communicative strategies.

6 Conclusion

As L2 research has employed multimodal learner corpora to analyze verbal and nonverbal behaviors such as gestures and facial expressions, previous studies (Saeki et al., 2021; Mawalim et al., 2025) have developed multimodal models of L2 proficiency. However, given the limited contribution of nonverbal features in these models, the present study reconsidered the role of nonverbal behaviors as complementary communicative resources that interact with verbal behaviors rather than as independent modalities.

We examined affective coherence between textual sentiment and facial expressions under the hypothesis that more proficient learners would exhibit stronger multimodal alignment. By quantifying this alignment through dominant-emotion frequency and cosine similarity, we identified what may be described as a *proficiency–coherence paradox*. While valence-based alignment (i.e., positive, neutral, and negative categories) exhibited a systematic relationship with L2 proficiency levels, a more fine-grained analysis based on seven facial-expression categories revealed greater divergence between textual sentiment and facial expressions among higher-proficiency learners.

Overall, these findings provide empirical support for the affective coherence as a cross-modal construct in multimodal learner corpora. By demonstrating that nonverbal behaviors can contribute to the modeling of L2 proficiency, this study argues the importance of examining relationships between verbal and nonverbal modalities rather than treating them as separate sources of information. Furthermore, the results suggest that cross-modal alignment warrants broader investigation in L2 research. Future studies should examine how affective coherence converges and diverges across proficiency levels and communicative contexts to provide a more comprehensive account of multimodal L2 communication.

Limitations

This study is subject to several limitations related to the dataset, nonverbal features, analytical tools, and coherence measures.

First, the dataset used does not guarantee consis-

tent experimental settings. The data were selected for their public accessibility where recording conditions (e.g., lighting, camera angle) inevitably vary. While these conditions enhance ecological validity, future research should investigate preprocessing and normalization techniques to mitigate the effects of variation in lighting, camera angle, and recording quality.

Second, the scope of nonverbal features was limited to seven discrete facial-expression categories. Although these facial-expression categories capture surface affective information, a more comprehensive model should incorporate other nonverbal cues, including head movements, gaze direction, hand gestures, and posture. Facial-expression analysis could also be extended to Facial Action Units (AUs) and facial landmarks. Tools such as Py-Feat are capable of recognizing 20 AUs and 68 facial landmarks, enabling more fine-grained analyses of facial behavior.

Third, the findings of this study are dependent on the performance of the automated tools used for multimodal analysis. For instance, transcription accuracy depends on Whisper, facial-expression recognition depends on Py-Feat, and textual sentiment classification depends on Gemini. Errors introduced at any stage of the analysis pipeline may affect the resulting coherence measures. Facial-expression recognition systems may exhibit demographic biases associated with factors such as gender, ethnicity, and skin tone (Rhue, 2018; Nagpal et al., 2019). Because the present study relies on automated facial-expression analysis, such biases may affect the estimation of facial-expression categories and, consequently, the resulting coherence measures.

Finally, the coherence measures were based on static comparisons within speech segments rather than on temporal dynamics. While this approach facilitates comparisons between textual sentiment and facial expressions, it does not capture how nonverbal behaviors evolve over time. Future research should employ time-series approaches to examine how facial expressions evolve during speech and how such temporal patterns relate to affective coherence in L2 communication.

Ethics Statement

This study used publicly available video data and open-source computational tools, and did not involve direct interaction with human participants.

We manually verified that all videos remained public as of March 2026 and complied with the platform's Terms of Service.

The computational tools and video data were used solely for academic research and were not employed for decision-making, profiling, or evaluation of individual participants.

To protect the privacy of the speakers in the video data, no personally identifiable information was collected, and all data were analyzed solely at an aggregate level for linguistic research.

Data processing was conducted exclusively by the authors; as no external annotators were employed, no issues regarding labor compensation or third-party data exposure arise.

Regarding the use of artificial intelligence, the Gemini 2.5 Flash-Lite API was utilized via Google Colab to generate textual sentiment scores. All theoretical frameworks, research questions, methodological decisions, and statistical interpretations were developed exclusively by the authors.

We acknowledge that automated textual sentiment and facial-expression recognition systems may exhibit limitations and biases. Accordingly, the outputs of these tools were treated as computational indicators for corpus analysis rather than as direct measurements of speakers' internal emotional states. Furthermore, the study focuses on the relative alignment between textual sentiment and facial expressions and does not make psychological, clinical, or educational judgments about individual speakers.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and constructive suggestions, which helped improve the quality of this paper. This work was supported by JSPS KAKENHI, Grant Numbers 22300299, 15H02940, 17K18679, 21K18373, and 23K25347.

References

- Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *Openface: A general-purpose face recognition library with mobile applications*. Technical report, CMU-CS-16-118, CMU School of Computer Science.
- Nicolas Ballier, Adrien Meli, Maelle Amand, and Jean-Baptiste Yunès. 2023. *Using Whisper LLM for automatic phonetic diagnosis of L2 speech, a case study with French learners of English*. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 282–292, Online. Association for Computational Linguistics.
- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. *Openface 2.0: Facial behavior analysis toolkit*. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.
- Janet Bavelas and Jennifer Gerwing. 2007. *Conversational hand gestures and facial displays in face-to-face dialogue*. In Klaus Fiedler, editor, *Social Communication*, pages 283–308. Psychology Press.
- Veronica Bonsignori and Gloria Cappelli. 2021. *Developing strategies for conceptual accessibility through multimodal literacy in the english for tourism classroom*. *Multimodal Communication*, 11(1):31–47.
- Catherine L. Caldwell-Harris. 2014. *Emotionality differences between a native and foreign language: theoretical implications*. *Frontiers in Psychology*, 5.
- Jin Hyun Cheong, Eshin Jolly, Tianakang Xie, Sophie Byrne, Matthew Kenney, and Luke Chang. 2023. *Py-Feat: Python facial expression analysis toolbox*. *Affective Science*, 4:781–796.
- Neil Cohn and Joost Schilperoord. 2024. *A Multimodal Language Faculty: A Cognitive Framework for Communication*. Bloomsbury Publishing.
- Juan I. Durán, Rainer Reisenzein, and José-Miguel Fernández-Dols. 2017. *Coherence between emotions and facial expressions: A research synthesis*. In José-Miguel Fernández-Dols and James A. Russell, editors, *The Science of Facial Expression*, pages 107–130. Oxford University Press.
- Paul Ekman. 1971. *Universals and cultural differences in facial expressions of emotions*. In James K. Cole, editor, *Nebraska Symposium on Motivation, 1971*, volume 19, pages 207–282. University of Nebraska Press, Lincoln, NB.
- Paul Ekman, Wallace V Friesen, and Joseph C Hager. 2002. *Facial Action Coding System: The Manual on CD ROM*. Research Nexus eBook, Salt Lake City, UT. Oclc: 176910557.
- Daniele Franceschi. 2018. *Simulating a lawyer-client interview: the case of italian EFL university learners*. *ESP Across Cultures*, 15:41–59.
- Google. 2024. Gemini 2.5 flash lite. <https://ai.google.dev/>. Large language model.
- Marianne Gullberg, Kees De Bot, and Virginia Volterra. 2008. *Gesture and some key issues in the study of language development*. *Gesture*, 8(2):149–179.
- Susan Hunston. 2022. *Corpora in Applied Linguistics*, 2nd edition. Cambridge Applied Linguistics. Cambridge University Press.

- Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. [SMILE: Multimodal dataset for understanding laughter in video with language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, Mexico City, Mexico. Association for Computational Linguistics.
- IELTS. 2015. Ensuring quality and fairness in international language testing: An overview of the International English Language Testing System (IELTS). Guide for educational institutions, governments, professional bodies and commercial organisations.
- Adam Kendon. 2000. [Language and gesture: unity or duality?](#) In David McNeill, editor, *Language and Gesture*, Language Culture and Cognition, pages 47–63. Cambridge University Press.
- You Li, Han Ding, Yuming Lin, Xinyu Feng, and Liang Chang. 2024. [Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis](#). *Artificial Intelligence Review*, 57.
- Jiajun Liu, Aishan Wumaier, Cong Fan, and Shen Guo. 2023. [Automatic fluency assessment method for spontaneous speech without reference text](#). *Electronics*, 12(8).
- Willy Mateo, Leonardo Eras, Giancarlo Carvajal, and Federico Domínguez. 2024. [Detecting speech disfluencies using open-source tools in automatic feedback systems for oral presentation training](#). In *Proceedings of the 16th International Conference on Computer Supported Education, Volume 2*, pages 213–20. INSTICC, SciTePress.
- Candy Olivia Mawalim, Chee Wee Leong, Guy Sivan, Hung-Hsuan Huang, and Shogo Okada. 2025. [Beyond accuracy: Multimodal modeling of structured speaking skill indices in young adolescents](#). *Computers and Education: Artificial Intelligence*, 8:100386.
- Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. 2019. [Deep learning for face recognition: Pride or prejudiced?](#) *Preprint*, arXiv:1904.01219.
- Aneta Pavlenko. 2012. [Affective processing in bilingual speakers: Disembodied cognition?](#) *International journal of psychology : Journal international de psychologie*, 47:405–28.
- Michael Peechatt, Cecilia Ovesdotter Alm, and Reynold Bailey. 2024. [MULTICOLLAB: A multimodal corpus of dialogues for analyzing collaboration and frustration in language](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11713–11722, Torino, Italia. ELRA and ICCL.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Lauren Rhue. 2018. [Racial influence on automated perceptions of emotions](#). *SSRN Electronic Journal*.
- Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa, and Tetsunori Kobayashi. 2021. [Analysis of multimodal features for speaking proficiency scoring in an interview dialogue](#). In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 629–635.
- John Sweller. 2010. Cognitive load theory: Recent theoretical advances. In Jan L. Plass, Roxana Moreno, and Roland Brunken, editors, *Cognitive Load Theory*, pages 29–47. Cambridge University Press.
- Magdalena Szyszka, Pauliina Peltonen, and Pekka Lintunen. 2026. [Unravelling the relationship between language anxiety and foreign language speech fluency in a monologue production](#). *Journal of Multilingual and Multicultural Development*, 47(1):323–337.
- Heyong Wang and Yuanhao Chen. 2026. [Tri-cross modal fusion network for multimodal sentiment analysis in short videos](#). *Signal Processing: Image Communication*, 144:117542.
- Qingyi Wang, Feifei Xu, Xianyang Wang, Shengjun Wu, Lei Ren, and Xufeng Liu. 2025. [How anxiety state influences speech parameters: A network analysis study from a real stressed scenario](#). *Brain Sciences*, 15(3):262.
- Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022. [A systematic review on affective computing: emotion models, databases, and recent advances](#). *Information Fusion*, 83-84:19–52.
- Page Wheeler and Kazuya Saito. 2022. [Second language speech intelligibility revisited: Differential roles of phonological accuracy, visual speech, and iconic gesture](#). *The Modern Language Journal*, 106(2):429–448.
- Mengjiao Wu and Jennifer M. Roche. 2025. [Emotion, proficiency, and arousal: exploring speech and physiological responses in chinese esl learners](#). *Frontiers in Human Neuroscience*, 19.
- Junyao Zhang, Wataru Sato, Naoya Kawamura, Koh Shimokawa, Budu Tang, and Yuichi Nakamura. 2024a. [Sensing emotional valence and arousal dynamics through automated facial action unit analysis](#). *Scientific Reports*, 14.
- Wenxuan Zhang, Yue Deng, Bing Liu, and Lidong Pan, Sinno and Bing. 2024b. [Sentiment analysis in the era of large language models: A reality check](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, and 1 others. 2026. [A survey of large language models](#). *Preprint*, arXiv:2303.18223.

Appendix

Prompt for Textual Emotion Recognition

Analyze the emotional intensity of the following sentence for each emotion provided.

Sentence: {SENT}

Emotions to evaluate (STRICT ORDER):

[anger, disgust, fear, happiness, sadness, surprise, neutral]

Rules:

1. Assign a score from 0.00 to 1.00 for EVERY emotion in the list.
2. Even if the score is 0.00, you **MUST** include it in the output.
3. Output the results on a single line.
4. Format: Emotion: Score separated by commas.
5. **DO NOT** change the order of emotions.
6. Use the exact sequence: [anger, disgust, fear, happiness, sadness, surprise, neutral].
7. The sum of all emotion scores **MUST** be 1.00.