Multimodal Affective Modelling in an Intelligent Tutoring System for Foreign Language Learning

Dionysios Koulouris[1], **Athanasios Kallipolitis**[1], **Melina Tziomaka**[1], **Argyrios Zafeiriou**[1], **Stamatios Orfanos**[1], **Andreas Menychtas**[1], **Ilias Maglogiannis**[1], **Stamatia Michalopoulou**[2], **George Tsoulouhas**[2], **Athina Sioupi**[2], **Voula Giouli**[2]
[1]University of Pireus, [2]Aristotle University of Thessaloniki, Greece
{nkoulouris, nasskall, tziomakamel, zafeiriou, sorfanos, amenychtas, imaglo}@unipi.gr
george.tsoulouhas@athenarc.gr, {smichalo, sioupi, pgiouli}@del.auth.gr

Abstract

Foreign language learning is a cognitively and affectively demanding process, in which fluctuations in attention and motivation can negatively impact learner engagement. Emotions play a central role in this process, yet they are rarely modelled in a systematic, data-driven manner in authentic learning environments. This paper presents a prototype affective computing architecture that incorporates various modalities (audio, video, biosignals) to facilitate real-time or near-real-time emotion recognition in an educational scenario; the architecture is integrated within an emotion-aware and adaptive Language Learning application that harnesses Large Language Models in view of providing appropriate educational scenarios to learners. The system comprises modules for acquiring data for each modality and a processing pipeline for synchronizing and analyzing heterogeneous affective signals. We demonstrate both the feasibility and applicability of the approach through a proof-of-concept implementation and discuss its relevance for studying learner affect and supporting affect-aware educational scenarios. The results highlight both the applicability of multimodal affective data in educational settings and the need for further research on their pedagogical interpretation and use.

Keywords: affective computing, emotion recognition, educational application, affect-aware language learning

1. Introduction

Foreign language learning is a demanding process for both learners and educators, requiring sustained attention and concentration. Empirical studies show that students often experience negative emotions such as stress, anxiety, fatigue, and short lapses during instruction, which lead to reduced engagement and distraction (Hlas et al., 2019). At the same time, research has shown that positive emotions - such as enjoyment and satisfaction - play a significant role in shaping learners' motivation and sustained engagement (Dewaele and MacIntyre, 2014; Dewaele and Macintyre, 2016; Kantaridou and Psaltou-Joycey, 2023).

The paper presents a multimodal affective computing module integrated into EmoBot (Kallipolitis et al., 2026), an agentive language learning system that employs Large Language Models (LLMs) to deliver CEFR[1]-aligned practice and formative assessment activities within an interactive educational environment. While the system includes safeguards for detecting and pedagogically addressing toxic or inappropriate language, its primary contribution lies in the integration of a multimodal affective computing module. This module enables the investigation of learners' affective states during language learning interactions, positioning it not merely as an adaptive tutor but as a platform for computational affective science in education, capable of shedding light on affective engagement, emotional regulation, and learner behavior over time.

The rest of the paper is structured as follows: Section 2 outlines the aim and scope of the research; Section 3 presents the related work, focusing on multimodal emotion analysis and affective computing in foreign language learning, and identifies the literature gap that highlights the need for a biosignal-enabled multi-modal dual approach architecture for efficient emotion analysis in foreign language learning contexts. The system design, the architecture of the incorporated modules, emphasizing biosignals, and the technical implementation of the proposed system along with the adaptive affect-aware mechanisms, are detailed in Section 4. Section 5 demonstrates the system in practice, providing valuable insights into the proof-of-concept application and into a theory-driven qualitative evaluation. The challenges of the approach are discussed in Section 6 followed by conclusions, lessons learned, and directions for further research, in Section 7.

[1]Common European Framework of Reference for Languages (Council of Europe, 2001)

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 190–200
16 May 2026.

2. Aim and scope of the research

Recent advances in Natural Language Processing (NLP) have enabled increasingly accurate sentiment and emotion detection from textual and audio data. Neural architectures and LLMs have improved the modelling of affective cues in written and spoken language, supporting a range of applications in dialogue systems that enable human–computer interaction.

Parallel progress in affective computing has expanded both the methodological scope and the range of modalities through which emotional states can be inferred, especially in the field of health (Wang et al., 2024a). Computational emotion analysis increasingly relies on multimodal approaches, incorporating textual information, visual data, audio signals, and, more recently, physiological measurements recorded in real time or near-real time (NRT). The integration of heterogeneous yet complementary modalities has significantly enhanced the robustness of affective inference systems. Within this context, advanced human-computer interaction methods are aimed at identifying emotion through various sources. By utilizing intelligent pattern recognition techniques, like Neural Networks (NNs) and Support Vector Machines (SVMs), the latest emotion recognition trends include image analysis (Kundu and Saravanan, 2017) or textual analysis (Alslaity and Orji, 2024) with encouraging results.

At the same time, progress in mobile and wearable technologies has enabled the recognition of users' physiological signals through comodity devices (Dias and Cunha, 2018) or their incorporation into self-assessment procedures (Saganowski et al., 2023). Such wearable biosensor technologies are increasingly explored in learning applications, significantly impacting the field and providing insights regarding learners' performance (Hernández-Mustieles et al., 2024). More broadly, when multiple heterogeneous yet informative data sources are available, multimodal frameworks are widely adopted to support more comprehensive modelling of affective states (Lyons, 2015).

Within educational settings, various modalities have been used to monitor learner interaction and behavior. In most cases, systems, however, rely on a single interaction interface (e.g., text and audio capture during online instruction), typically performing modality-specific analysis followed by feature-level fusion (Miao, 2025). Nevertheless, the integration of multimodal affective inference into pedagogically grounded tutoring systems remains comparatively under-explored.

Addressing this gap, the EmoBot project (Kallipolitis et al., 2026) proposes a multimodal affective module integrated into an LLM-based, CEFR-aligned intelligent tutoring system. The system is currently tailored to learners of German as a foreign language. Rather than presenting the tutoring system as a whole, this work focuses specifically on the design and integration of the affective module within an agentive, NLP-driven architecture. The study therefore concentrates on the operationalization, synchronization, and fusion of multimodal affective signals, as well as on how these signals can inform adaptive interaction in foreign language learning scenarios.

A novelty of this work lies in the inclusion of various data sources through different user interaction points and the application of a dual data analysis architecture. The affective module is enhanced with the modalities of text, audio, and visual inputs from the learner's computer and biosignals, from the learner's already-owned smartwatch, offering a complete example for an end-to-end emotion-analysis-enabled language learning system (Koulouris et al., 2025). In this work, we highlight the importance of applying intelligent multi-resource and multimodal methods by implementing the proposed novel EmoBot architecture into an emotion recognition paradigm fed by textual, vocal, visual, and biosignal data. The proposed system uses the dual-architecture approach to collect, isolate, sync, and manipulate the data types of each modality and provide user feedback. This feedback is vital for learning, as evidenced by this proof-of-concept technical implementation.

One step further, a key contribution of the work is the transfer of intelligent mHealth-inspired methodologies to the domain of education, combining distributed systems, commodity wearable devices, AI, advanced data manipulation, and data fusion techniques to support affect-aware interaction in foreign language learning contexts.

3. State of the art

This section reviews the existing literature on emotion analysis in educational contexts. It focuses on multimodal approaches and their application to foreign language learning (FLL) and highlights current methodologies and technological trends. Finally, it identifies existing limitations and, thereby, establishes the research gap addressed by the present study.

3.1. Multimodal emotion analysis

Multimodal sentiment analysis is currently a central topic of research, enabling advanced human-machine interaction (Zhu et al., 2024) and providing a robust source of information (Ezzameli and Mahersia, 2023). The most common approaches include the use of visual analysis combined with NLP methods, either at neutral time (Hu and Flax-

man, 2018) or in real time. Real time examples include the utilization of speech (Kim et al., 2007), images or health-oriented data (Huang et al., 2016) such as electroencephalograms, respiratory belts, and electrodermal activity monitors (Nandi et al., 2021).

While these solutions provide classification for valence and arousal, they still require sophisticated, and particularly expensive, hardware and cannot run on commodity mobile devices. Other mHealth-oriented solutions include the utilization of feature extraction from textual and visual modalities, by utilizing Convolutional Neural networks (CNNs), and relying only on the commodity device hardware (Poria et al., 2016). Some approaches utilize text sentiment analysis, usually using SVM text classification models (Song, 2021).

Nevertheless, the integration of such multimodal architectures into pedagogically grounded educational systems remains comparatively limited.

3.2. Affective Computing in Language Learning

Intelligent Tutoring Systems (ITSs have been increasingly adopted in instructional settings, including language education, as tools for supporting interactive and personalized learning (Belda-Medina and Calvo-Ferrer, 2022; Barrot, 2023; Antoniou-Kritikou et al., 2024). Empirical research indicates that their integration into educational contexts can enhance learner performance, foster critical thinking skills, and provide targeted and individualized instructional support (Davar et al., 2025), while also contributing to positive learner experiences and increased satisfaction (Kuhail et al., 2022). Despite acknowledged limitations, several studies report that the pedagogical benefits of conversational agents generally outweigh their drawbacks for both learners and educators (Huang et al., 2022).

Recent advances in LLMs have further enhanced the capabilities of ITSs by enabling adaptive interaction, flexible guidance, automated educational content generation - especially in low-resource settings (Samuel et al., 2024), and instruction tailored to individual learner needs. In the context of FLL, conversational agents have been shown to offer particular advantages for learners who are reluctant to engage in classroom interaction due to stress, anxiety, or lack of confidence, suppressing negative emotions and providing a low-pressure environment for language practice and experimentation (Alwazzan, 2024).

In the field of education, various emotion analysis methods have been proposed. Emotion recognition was used for sentiment analysis in learning, focusing on identifying emotions and having learning as the centre of interest (Barrón Estrada et al., 2019). This method is also implemented in distant learning environments, where it monitors learners' perception, learning processes, and communication methods and highlights the importance of knowledge acquisition and decision-making (Duraes et al., 2021). Within the FLL context, there are systems that monitor the emotions experienced by learners (MacIntyre and Gregersen, 2012), the dynamics of which affect the learning outcome and proficiency (Wang et al., 2024b). Anxiety is a metric that is also measured in foreign language e-learning systems (Ismail and Hastings, 2019).

The utilization of the biosignal modality for emotion recognition is applied in web-based e-learning scenarios and provides the tutor with insights tailored to providing appropriate learning assistance or modifying guidance (Chen and Lee, 2011). Finally, camera-based systems with capabilities to recognize facial expressions using CNNs are studied in classrooms for assessing students' engagement, providing promising results (Pabba and Kumar, 2021).

Although the literature reports multimodal approaches in educational settings, there is limited work addressing languages other than English—such as German—particularly in systems that integrate biosignal-enabled emotion and sentiment analysis within affective computing frameworks. Existing approaches tend to rely either on single-modality inference or on multimodal fusion techniques that require sophisticated and often costly hardware. Our work addresses this gap by proposing a multimodal emotion analysis system that incorporates intelligent mHealth architectures and methods, enabling real-time or near-real-time estimation of learners' emotional states using commodity devices.

4. Methodology

The methodological approach adopted in this work aims to use affective signals in an authentic FLL setting in view of monitoring how learners' affective states emerge and evolve during interaction with an agentive educational system. Rather than treating emotion recognition as an end in itself, affective inference is used as a means for studying learner affective states and regulation within CEFR-aligned learning activities that support written production, reading comprehension, and use of German, oral, and listening comprehension. To this end, a multimodal affective analysis framework is embedded within EmoBot, enabling the continuous observation of affective signals during learning sessions.

The proposed system methodology is structured into three subsections: system design specifications, architectural details, and the implementation of the prototype.

4.1. System Design

The system is designed to facilitate efficient emotion analysis in the educational scenario of language learning through simultaneous utilization of biosignal, visual, and audio modalities. The pilot implementation is tailored to the needs of learners of German as a foreign language, as a proof-of-concept. The data inputs are recorded through handheld commodity devices, such as smartwatches and personal computers. A Cloud Processing Platform (CPP) retrieves the input signals, performs data processing and per-modality emotion classification, before fusing the data to extract an overall emotional status. In addition to the ML methods that are used, the CPP integrates an alternative LLM approach for analysing emotion, offering a dual architecture approach that recognizes emotion in real-time or NRT. The system design is shown in Figure 1.

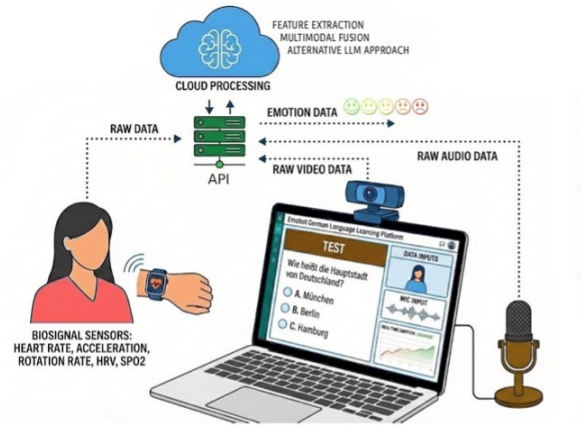

Figure 1: Overview of EmoBot. Data from different modalities (biosignals, audio, and visual) are transmitted to a cloud processing platform. Methods like feature extraction, data fusion, and LLMs are utilized to infer emotion in real-time or near-real-time.

4.2. System Architecture

The design of EmoBot follows a high-level interaction pipeline in which learner input is mediated through human–computer interaction, analyzed for affective cues, and used to inform adaptive pedagogical responses. Language understanding and generation are implemented through LLM-based NLP modules that support CEFR-aligned activity generation as well as feedback and explanation via retrieval-augmented generation. Within this pipeline, the emotion recognition module is aimed at providing affective signals that guide interaction flow and adaptation, enabling the system to adjust feedback, pacing, and task selection in a pedagogically appropriate manner.

The architecture of the proposed system fulfils the requirements of monitoring biosignals, along with recording the user expressions, importing the data in real time through a streaming bus service, synchronizing, extracting emotion from each modality and fusing. Hence, the system consists of a web e-Learning platform, which includes the learning interface and handles the visual and audio data input. Moreover, a smartwatch app is developed for biosignals recoding and a cloud environment incorporates a data input stream, an Affective Processing Unit (APU), an LLM unit and a Late Fusion Engine. An overview of the proposed architecture is shown in Figure 2.

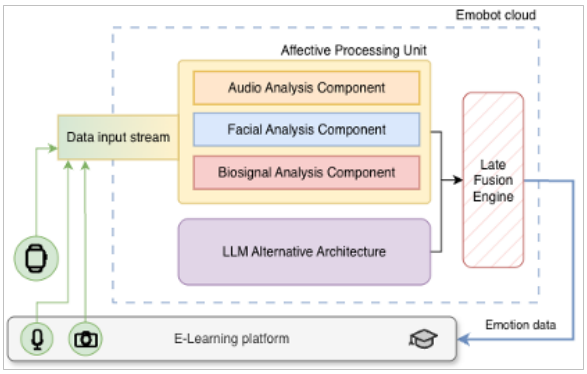

Figure 2: The EmoBot emotion analysis system architecture. Data relevant to each modality are retrieved from the educational platform (audio, visual data) and from a smartwatch application. These data are then fed into a unified input stream and transmitted to a cloud-based service for processing. Therein, the Affective Processing Unit (APU) handles the synchronization, feature extraction, and prediction generation. Data fusion is, then, performed through a Late Fusion Engine, enhanced and evaluated alternatively by LLMs, before returning a single emotional status to the educational platform.

The data input stream retrieves data from the three modalities in real time and transfers them to the APU. Therein, the Audio Analysis Component is responsible for extracting audio waveform data and performing emotion-based feature extraction. The Facial Analysis Component handles camera frames and the Biosignal Analysis Component analyses vital signs and IMU [2] data, performing also feature extraction. The results are synced and fused, incorporated alternatively by LLMs, contributing to a dual architecture approach.

[2]Intertrial Measurement Unit – A sensor which includes accelerometer and gyroscope and calculates accelerations and rotation rates.

4.3. Implementation

The system implementation is divided into three platforms: web, smartwatch and cloud.

4.3.1. Web

The web system initiates once a learning session begins and utilizes WebRTC to send video and audio to the cloud platform. It is implemented in the NodeJS environment. Moreover, through a data streaming service implemented using Firebase Database, the web platform awaits emotion data from the cloud in NRT.

4.3.2. Smartwatch

The smartwatch application is developed in watches that allow the custom application development, such as iOS WearOS and Android WatchOS. The recording session is initiated by the user by entering a 5-digit code generated over the web and retrieving the current session data. Recorded measurements include heart rate, SpO2, HRV calculated on device, accelerations, and rotation rates.

4.3.3. Cloud

The cloud service exposes the data input stream through an API and an event-based data streaming functionality. The input interface retrieves the data from the aforementioned modalities and initiates the appropriate analysis component. The Audio Analysis Component utilizes a CNN-LSTM architecture with MFCC features and bidirectional temporal modelling to extract emotion from speech. The Facial Analysis Component utilizes methods for real-time facial expression recognition, with planned extensions using MediaPipe for landmark detection and 3D-CNNs for temporal expression analysis. The Biosignal Analysis Component utilizes datasets like the WESAD dataset (Schmidt et al., 2018)r or SVM approaches (Hakim et al., 2018) to efficiently calculate emotions of stress, happiness, anger, or sadness.

Finally, an alternative LLM-based architecture is designed to extract fused affective information from audio and visual modalities. It is currently under development, and no fully operational prototype has been implemented to date. However, preliminary integration experiments have been conducted to assess its feasibility.

4.4. Adaptive Tutoring and human-computer interaction

The adaptive tutoring layer of EmoBot leverages affective inference by translating multimodal emotion predictions into pedagogically meaningful interaction strategies. Crucially, this affect-driven adaptation is tightly integrated with the system's ability to generate CEFR-aligned instructional content through an LLM-based educational module. Thus, adaptation occurs not only at the level of feedback style and tone but also at the level of task design, linguistic complexity, and progression across proficiency levels.

This CEFR-aware generation mechanism provides the structural backbone upon which affective adaptation operates. Rather than dynamically generating arbitrary content, the system selects or generates tasks within a constrained proficiency band and then modulates their complexity and scaffolding intensity in response to affective signals.

5. Evaluation

This section highlights the results of the proof-of-concept prototype, including the system in practice as a paradigm validation and a qualitative evaluation.

5.1. System in practice

For the proof-of-concept smartwatch prototype, a WearOS application is developed that can measure heart rate and SpO2 in real time, as well as HRV in NRT. Additionally, the data from the onboard IMU sensor are included in a batch that is sent to the cloud every 30 seconds. The authentication is performed by entering a 5-digit code bound to the current user session. Snapshots of the system in the authentication and measurement screens are shown in Figure 3.

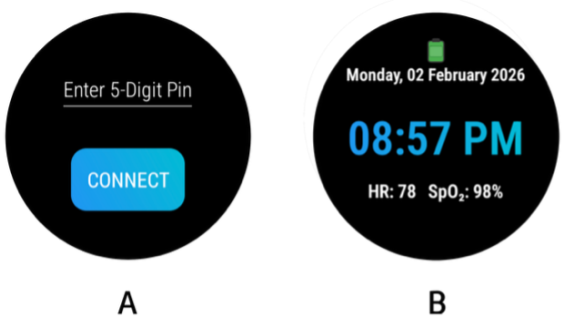

Figure 3: The mobile interface, including the authentication screen (A) and the live measurement screen (B), indicating the current heart rate and SpO2.

For the web interface, an HTML and JS-based iframe was created, including an interface for the camera and audio feed, and was included in the E-Learning platform. The platform is also responsible for showing the 5-digit code and rendering

the emotion result from the streaming output of the cloud service. A snapshot of the web interface is shown in Figure 4

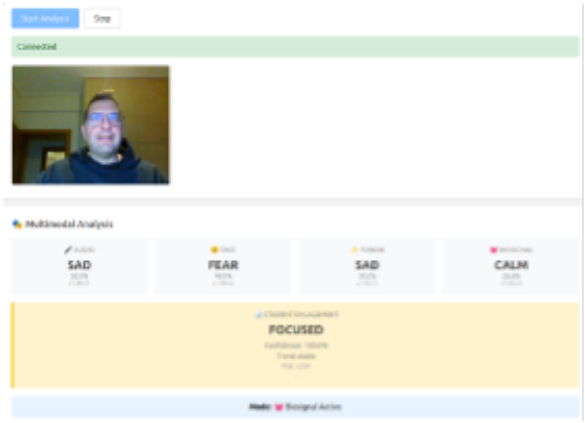

Figure 4: The web interface, including a camera preview utilizing WebRTC for audio and video input streams and multimodal analysis results in real time.

The components of the cloud service are built using a decoupled approach, in Python, dividing the data handling and the model inference. An LLM is utilized in the Late Fusion Engine to fuse the prediction of each modality into a general output, before notifying the E-Learning platform.

5.2. Paradigm evaluation

This proof-of-concept prototype acts as an example of the proposed architecture's implementation. The three modalities are validated using state-of-the-art experimental protocols. For the audio modality, a comprehensive evaluation across eight datasets (English and German) identified the CNN-LSTM architecture as the superior model, achieving up to 81.3% accuracy on German speech with high stability and cross-lingual generalization, significantly outperforming complex alternatives like EfficientNet.

For the visual modality, the DeepFace library (Serengil and Ozpinar, 2021) was employed for facial emotion recognition. DeepFace utilizes the Mini-Xception architecture (Arriaga et al., 2017), a lightweight CNN based on depthwise separable convolutions, trained on the FER-2013 dataset. The model achieves approximately 66% accuracy on the FER-2013 benchmark, which is comparable to human-level performance on the same dataset (65 ± 5%). With only around 60,000 parameters, Mini-Xception is highly efficient and well-suited for real-time or near-real-time deployment on commodity hardware, making it a practical choice for the visual emotion recognition component of the proposed system. These independent validations prove that the system uses the best possible methods and mitigates individual modality risks.

The alternative LLM-based architecture is under development and has not yet been fully deployed. It is designed to perform valence and arousal estimation from both audio and visual modalities, to produce a multimodal fused output. Preliminary integration experiments suggest that BERT-based models, particularly DistilBERT (Sanh et al., 2019), show promise in meeting the defined requirements. However, its evaluation falls outside the scope of this study, which focuses on biosignal-based affective computing in distributed systems using commodity devices.

5.3. Qualitative evaluation

In addition to component-level quantitative validation, and in the absence of a user study, a structured qualitative evaluation was conducted to assess the theoretical coherence, pedagogical interpretability, and instructional consistency of the affect-aware tutoring system within a CEFR-aligned FLL context.

Theoretical coherence: representation of affect and FLL theory. Our system models affect along the valence-arousal dimensions following Russell (1980) and Bradley and Lang (1994). This dimensional representation is consistent with the data captured via the various modalities including biosignals, while it also aligns with SLA research distinguishing positive (enjoyment, satisfaction) and negative (anxiety, fear of evaluation) affect (Dewaele and MacIntyre, 2014) in language learning contexts. Moreover, observed affect trajectories during task escalation (e.g., increasing arousal during time-constrained production tasks) correspond to theoretical expectations regarding foreign language anxiety (FLCA). Similarly, stabilized valence during scaffolded support aligns with positive psychology perspectives in SLA (MacIntyre and Gregersen, 2012). In this regard, the dimensional model produces affect patterns that are theoretically consistent with established emotion research in FLL contexts.

Pedagogical interpretability. The recognition and modelling of learners' emotional states provide an insight into variations in emotional states throughout the learning process. SLA literature identifies anxiety, fear of negative evaluation, and reduced motivation as key adult learner constraints (Schumann, 1975; Gao and Liu, 2022). System-level affect trajectories plausibly correspond to elevated arousal in difficult tasks, stabilized valence following immediate formative feedback, and reduced activation during highly scaffolded comprehension tasks. These patterns align with both anxiety research and positive psychology perspectives in SLA (MacIntyre and Gregersen, 2012).

Affect-aware adaptation and feedback. Systematic and timely feedback has been found to enhance performance and sustain learner engagement and system retention (Katinskaia et al., 2018; de Haas et al., 2020), while encouragement contributes to progression across CEFR proficiency levels (Council of Europe, 2001). Formative feedback is widely regarded as a major determinant of learning outcomes (Hattie and Timperley, 2007) and is particularly critical for the development of written communicative language competence. Feedback should be immediate, transparent, and explicitly linked to CEFR can-do descriptors, guiding learners toward greater alignment with targeted proficiency levels and enabling iterative performance regulation in relation to descriptor-based performance criteria (Hattie and Timperley, 2007).

In the proposed architecture, affect-aware adaptations are embedded implicitly within the pedagogical interaction. The system does not explicitly label learners' emotional states (e.g., "You seem anxious"); instead, it modulates tone, pacing, task difficulty, and linguistic complexity. This approach implements affect-informed scaffolding by adjusting instructional support according to inferred valence–arousal patterns and gradually reducing assistance as learners stabilize, consistent with theoretical principles (Vygotsky, 1978).

These adaptations remain aligned with CEFR descriptors to ensure structured proficiency progression. This approach is grounded in SLA literature, which highlights the importance of fostering psychological safety and learner autonomy within a non-threatening interactional environment (Pachler et al., 2009), thereby reducing anxiety and promoting active participation in communicative language activities (Council of Europe, 2001).

Moreover, emotional variables, particularly learner self-confidence in communicative tasks, influence task completion and sustained engagement (Council of Europe, 2001), especially in evaluative contexts marked by perceived inadequacy or anxiety. Instructional design in our system supports learner agency and strategic competence through interactive scenarios and systematic positive reinforcement. Such affective engagement underpins persistence and successful progression across learning trajectories.

One step further, beyond real-time inference, the system maintains session-level and cross-session affective profiles indexed to task categories and CEFR descriptors. This longitudinal modelling enables detection of recurring task-affect couplings, such as elevated negative valence during reading comprehension at a specific proficiency level or reduced activation in written production; identified patterns guide medium-term adaptation through targeted reinforcement and controlled task re-sequencing. This approach is consistent with SLA evidence linking learning efficiency to sustained attention, structured repetition, and strategic competence deployment (Council of Europe, 2001).

Overall, this systematic account of emotions is essential for identifying their contributing factors and for guiding research methodologies designed to mitigate their effects. From a computational perspective, the system can be interpreted as a multimodal affect-adaptive control architecture in which valence and arousal estimates function as state variables governing pedagogical adaptation.

6. Discussion

In this section, we discuss challenges that arise from this work. Although many affective computing applications are grounded in basic emotion typologies (Ekman, 1992; Plutchik, 1980), the affective constructs that are pedagogically salient in educational settings only partially overlap with these predefined emotion categories. Basic emotion frameworks were primarily developed to describe evolutionarily grounded, universally recognizable affective expressions (e.g., anger, fear, happiness), often associated with discrete and prototypical facial configurations. In contrast, educational contexts frequently involve more subtle, cognitively intertwined affective states such as confusion, cognitive overload, sustained engagement, frustration, boredom, or curiosity. These states cannot always be mapped onto single discrete categories and may reflect blends of arousal, valence, and task-related appraisal.

This conceptual gap is also attested in the structure of widely used benchmark datasets, which predominantly annotate prototypical facial expressions corresponding to basic emotions under controlled conditions. In the present work, emotion categories were harmonized into a unified eight-class scheme, namely, Anger, Disgust, Fear, Happiness, Neutral, Sadness, Contempt, and Surprise (Ekman and Friesen, 1986), to reconcile discrepancies across source datasets (see Section 5.2). While this harmonization improves cross-dataset consistency for model training and evaluation, it does not fully address the broader issue that these categories only partially capture educationally meaningful affective states.

Two methodological challenges arise from this discrepancy. First, a label mismatch problem emerges: models trained on basic or general emotion datasets must rely on indirect or heuristic mappings when applied to pedagogical scenarios. Second, benchmark datasets often contain expressions recorded in controlled environments, whereas classroom or self study interactions involve low intensity, context dependent affective fluc-

tuations embedded within complex cognitive processes.

These limitations highlight the need for domain-specific datasets that go beyond discrete basic emotion labels and incorporate pedagogically grounded taxonomies. Such datasets should ideally include temporally aligned signals, task metadata, and dimensional affect annotations (e.g., valence–arousal trajectories), enabling the study of affect as a dynamic and context-sensitive process. Developing and validating such resources remains an important direction for advancing affect-aware educational systems.

7. Conclusion

To conclude, we have presented a multimodal affective computing architecture integrated into an intelligent language learning system delivering CEFR-aligned activities within an interactive environment. The focus is on the design and system-level validation of a distributed framework that combines audio, visual, and physiological signals acquired through commodity devices to support real-time or near-real-time affect inference.

The proposed architecture and the empirical results demonstrated by the working prototype underscore opportunities for further research. As a future direction, different methods can be used for modality-based predictions. Moreover, additional hardware such as IMU sensors or cameras can be tested. A larger-scale deployment of the system is underway, and, hence, results from applying the system to real students will be carried out. The state-of-the-art affective emotion recognition architecture, along with the developed paradigm, validates its applicability within the context of computational emotion recognition. The contribution directs towards continued research in the field, particularly in relation to its pedagogical and cognitive implications for Foreign Language Learning, with specific relevance to the acquisition of languages other than English.

Finally, future work is also underway to empirically validate the system with end-users and to develop a pedagogically grounded multimodal dataset designed to support reproducible research in affect-aware educational applications. All data will be made publicly available for reproducibility purposes, subject to ethical and privacy constraints.

8. Limitations and ethical considerations

At the time of submission, the proposed framework has undergone system-level validation but has not yet been deployed in a large-scale learner study. No learner self-reports, expert annotations, or controlled comparative experiments were conducted at this stage. Consequently, affect inference accuracy and measurable learning gains remain to be empirically validated. A controlled evaluation with learners of German as a foreign language is currently in progress.

Furthermore, biosignals are probabilistic indicators of affective states and may be sensitive to individual variability, contextual noise, and model bias. Although temporal smoothing and bounded adaptation reduce instability, misclassification of affect may lead to suboptimal instructional modulation.

From an ethical perspective, affective computing in educational settings raises concerns related to privacy, transparency, and autonomy. The system processes multimodal signals that may be considered sensitive data; therefore, data minimization, secure storage, and explicit informed consent are required.

9. Acknowledgements

This research was supported by the National Recovery and Resilience Plan (NRRP) “Greece 2.0” under the “Clusters of Research Excellence” (CREs) program, SUB1.1, with project code $\mathrm{O\Pi\Sigma\ TA}$ 5180519 and title “Interactive Agent with Emotional Intelligence for Second/Foreign Language Learning”, Acronym: “EmoBot”.

10. Bibliographical References

References

Alaa Alslaity and Rita Orji. 2024. Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. *Behaviour & Information Technology*, 43(1):139–164.

Mona Saleh Alwazzan. 2024. Investigating the Effectiveness of Artificial Intelligence Chatbots in Enhancing Digital Dialogue Skills for Students. *Başlık*, 13(2):573–584.

Ioanna Antoniou-Kritikou, Voula Giouli, George Tsoulouhas, and Constandina Economou. 2024. Using LLMs in a language teaching and learning application. *ECRIM*, 136. Special Theme: Large Language Models; Guest Editors: Diego Collarana Vargas and Nassos Katsamanis.

Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

Jessie S. Barrot. 2023. Using Chatgpt for second language writing: Pitfalls and potentials. *Assessing Writing*, 57:100745.

Maria Barrón Estrada, Ramón Zatarain Cabada, and Raul Oramas. 2019. Emotion Recognition for Education using Sentiment Analysis. *Research in Computing Science*, 148(5):71–80.

Jose Belda-Medina and José Ramón Calvo-Ferrer. 2022. Using Chatbots as AI Conversational Partners in Language Learning. *Applied Sciences*, 12(17):8427.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Chih-Ming Chen and T.-H Lee. 2011. Emotion recognition and communication for reducing second-language speaking anxiety in a web-based one-to-one synchronous learning environment. *British Journal of Educational Technology*, 42:417–440.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, Strasbourg.

Narius Farhad Davar, M. Ali Akber Dewan, and Xiaokun Zhang. 2025. AI Chatbots in Education: Challenges and Opportunities. *Information*, 16(3).

Mirjam de Haas, Paul Vogt, and Emiel Krahmer. 2020. The effects of feedback on children's engagement and learning outcomes in robot-assisted second language learning. *Frontiers in Robotics and AI*, 7:101.

Jean-Marc Dewaele and Peter Macintyre. 2016. *Foreign Language Enjoyment and Foreign Language Classroom Anxiety. The right and left feet of FL learning?*, pages 215–236. Bristol: Multilingual Matters.

Jean-Marc Dewaele and Peter D. MacIntyre. 2014. The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, 4(2):237–274.

Duarte Dias and João Paulo Silva Cunha. 2018. Wearable Health Devices—Vital Sign Monitoring, Systems and Technologies. *Sensors (Basel, Switzerland)*, 18.

Dalila Duraes, Ramon Toala, and Paulo Novais. 2021. *Emotion Analysis in Distance Learning*, pages 629–639. Springer International Publishing.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Paul Ekman and Wallace V. Friesen. 1986. A new pan-cultural facial expression of emotion. *Motivation and Emotion*, 10(2):159–168.

Kaouther Ezzameli and Hela Mahersia. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847.

Lixiang Gao and Honggang Liu. 2022. Revisiting students' foreign language learning demotivation: From concepts to themes. *Frontiers in Psychology*, 13.

Lutfi Hakim, Adhi Dharma Wibawa, Evi Septiana Pane, and Mauridhi Hery Purnomo. 2018. Emotion Recognition in Elderly Based on Spo2 and Pulse Rate Signals Using Support Vector Machine. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 474–479.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

María A. Hernández-Mustieles, Yoshua E. Lima-Carmona, Maxine A. Pacheco-Ramírez, Axel A. Mendoza-Armenta, José Esteban Romero-Gómez, César F. Cruz-Gómez, Diana C. Rodríguez-Alvarado, Alejandro Arceo, Jesús G. Cruz-Garza, Mauricio A. Ramírez-Moreno, and Jorge de J. Lozoya-Santos. 2024. Wearable biosensor technology in education: A systematic review. *Sensors*, 24(8).

Anne Cummings Hlas, Krista Neyers, and Sarah Molitor. 2019. Measuring student attention in the second language classroom. *Language Teaching Research*, 23:107 – 125.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 350–358, New York, NY, USA. Association for Computing Machinery.

Weijiao Huang, Khe Foon Hew, and Luke K. Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Xiaohua Huang, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen, and Matti Pietikäinen. 2016. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Comput. Vis. Image Underst.*, 147(C):114–124.

Daneih Ismail and Peter Hastings. 2019. Identifying foreign language anxiety when using an e-learning system. pages 131–140.

Athanasios Kallipolitis, Dionysios Koulouris, Melina Tziokama, Kosmas Pinitas, Argyrios Zafeiriou, Andreas Menychtas, Ilias Maglogiannis, Voula Giouli, Athina Sioupi, Stamatia Michalopoulou, George Tsoulouhas, Michail Katras, Panagiotis Charalampopoulos, and Aristotelis Stamopoulos. 2026. Conversational agent with emotional intelligence for foreign language learning. In *Proceedings of the 14th International Conference on Information and Education Technology (IEEE-ICIET 2026), Koriyama, Japan*. IEEE.

Zoe Kantaridou and Angeliki Psaltou-Joycey. 2023. Positive emotions and self-regulation strategies in EFL classroom situations. In *Selected papers on theoretical and applied linguistics*, pages 401–419.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of its and call. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*. European Language Resources Association (ELRA.

Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth S. Narayanan. 2007. Real-time emotion detection system using speech: Multimodal fusion of different timescale features. *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 48–51.

Dionysis Koulouris, Melina Tziomaka, Argyrios Zafeiriou, Kosmas Pinitas, Athanasios Kallipolitis, Andreas Menychtas, and Ilias G. Maglogiannis. 2025. Design of a multimodal affective module for emotion recognition in adaptive language learning systems. *2025 10th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–6.

Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2022. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018.

Tuhin Kundu and Chandran Saravanan. 2017. Advancements and recent trends in emotion recognition using facial image analysis and machine learning models. *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–6.

Agnieszka Lyons. 2015. *Multimodality*, chapter 18. John Wiley & Sons, Ltd.

Peter MacIntyre and Tammy Gregersen. 2012. Emotions that facilitate language learning: The positive-broadening power of the imagination. *Studies in Second Language Learning and Teaching*, 2(2):193–213.

Guixian Miao. 2025. A text-audio multimodal weighted network for emotion recognition to enhance interactivity in online education. *Journal of Computational Methods in Sciences and Engineering*, 25:5764 – 5777.

Arijit Nandi, Fatos Xhafa, Laia Subirats, and Santi Fort. 2021. Real-time multimodal emotion classification system in e-learning context. In *Proceedings of the 22nd Engineering Applications of Neural Networks Conference*, pages 423–435, Cham. Springer International Publishing.

Chakradhar Pabba and Praveen Kumar. 2021. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, 39.

Norbert Pachler, Harvey Mellar, Caroline Daly, Yishay Mor, and Dylan Wiliam. 2009. *Scoping a vision for formative e-assessment: a project report for JISC*. WLE Centre and JISC, London.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1:3–33.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Stanisław Saganowski, Bartosz Perz, Adam G. Polak, and Przemysław Kazienko. 2023. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing*, 14(3):1876–1897.

Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA. Association for Computing Machinery.

John H. Schumann. 1975. Affective factors and the problem of age in second language acquisition. *Language Learning*, 25(2):209–235.

Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyperextended lightface: A facial attribute analysis framework. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4.

Gang Song. 2021. Sentiment analysis of japanese text and vocabulary learning based on natural language processing and svm. *Journal of Ambient Intelligence and Humanized Computing*.

Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Ke Wang, Lin Wei, Yuanyuan Liu, Jingying Chen, Yibing Zhan, Hua Jin, Chongchong Qi, and Zhe Chen. 2024a. Affective computing for healthcare: Recent trends, applications, challenges, and beyond. In *CSIG Conference on Emotional Intelligence*, pages 3–19. Springer.

Peng Wang, Lesya Y. Ganushchak, Camille Welie, and Roel van Steensel. 2024b. The dynamic nature of emotions in language learning context: Theory, method, and analysis. *Educational Psychology Review*, 36.

Xianxun Zhu, Chaopeng Guo, Heyang Feng, Yao Huang, Yichen Feng, Xiangyang Wang, and Rui Wang. 2024. A Review of Key Technologies for Emotion Analysis Using Multimodal Information. *Cognitive Computation*, 16:1504 – 1530.