RESEARCH PAPER

Adult Learning Management with Emotion Sensing and Progress Tracking in Virtual Environments

Gulshan Banu A[1], Dr. M. Sukanya[2*], M. Kanaga Durga[3], Dr. S. Usha[4], Dr. S. P. Manikandan[5], Shiva Shankar S[6], Aswathy P B[7], Ms. Gopika P[8]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science, SNS College of Technology, Coimbatore - 641107. Email: gulshanasif97@gmail.com

[2*]Associate Professor, Department of Computer Science and Engineering, Karthir College of Engineering. Email: sukanmukesh@gmail.com (Corresponding Author)

[3]Assistant Professor, Department of Artificial Intelligence and Data Science, KGISL Institute of Technology, Coimbatore. Email: durga033@gmail.com

[4]Department of Artificial Intelligence and Data Science, Kathir College of Engineering, Coimbatore - 641048. Email: usha.samiappan@gmail.com

[5]Professor & Deputy Director, School of Engineering and Technology, CMR University (Lake Side Campus), Near International Airport, Chagalatti, Bengaluru, Karnataka, India. Mobile: 99413 52094. Email: dr.mani1973@gmail.com

[6]Assistant Professor, Department of Information Technology, Sri Krishna College of Engineering and Technology. Email: shivaofficial.1987@gmail.com

[7]Assistant Professor, Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore - 641105. Email: aswathynairb@gmail.com

[8]Assistant Professor, Department of Artificial Intelligence and Data Science, Easa College of Engineering and Technology. Email: ratheeshkumargopika@gmail.com

Received: 15th Feb, 2026; Revised: 27th Feb 2026; Accepted: 20th Mar, 2026; Available Online: 5th Apr, 2026

ABSTRACT

For applications targeting older adults, where personal support has been shown to potentially contribute to happiness and autonomy, emotionally aware human-machine interfaces can play a pivotal role in facilitating adaptable and engaging conversational interfaces. The lack of older adults in contemporary datasets and models restrains the ability to apply contemporary methodologies to age-related interfaces despite their improvements in affective computing and multimodal emotion recognition. This work discusses the development of the emotion expression recognition module of the virtual coach by presenting data collection, annotation design, and a preliminary methodological approach tailored to the specifications of the project. In the latter, we investigate the role of multiple modalities-speech from audio, and facial expressions, gaze, and head dynamics from video-in a standalone and combined manner for the detection of discrete emotion expressions in this setting. The collected corpus, consisting of users from Norway, France, and Spain, was annotated independently for the audio and video channels with unique emotional labels, and thus allowed for a cross-cultural performance comparison. The results confirm the informative value of the modalities with respect to the considered emotional categories; multimodal approaches generally outperformed others. The results are expected to guide the development of future systems and contribute to the limited literature on emotion recognition applied to elderly individuals in conversational human-machine interaction.

Keywords: Affective Computing, Multimodal Emotion Recognition, Human–Machine Interaction, Older Adults, Virtual Coaching Systems, Speech and Facial Analysis, Behavioral Cues

How to cite this article: Gulshan Banu A, Sukanya M, Kanaga Durga M, Usha S, Manikandan SP, Shiva Shankar S, Aswathy PB, Gopika P. Adult Learning Management with Emotion Sensing and Progress Tracking in Virtual Environments. Int J Drug Deliv Technol. 2026;16(4): 73-89. DOI: 10.25258/ijddt.16.4.10

Source of support: Nil.
Conflict of interest: None

1. Background and Motivation

An important aspect of intelligent conversational systems involves emotion-conscious human computer interaction, enabling robots to identify and change their behavior based on the user's emotional expression. By showing that affective states can be inferred by data-driven approaches, the domain of affective computing has played a theoretical and computational grounding for affect recognition from user behavior [1]. Emotion recognition leads to greater user engagement, facilitates more natural conversation flows, and facilitates adaptive system response for more than task-oriented interaction [2], [3]. Multimodal approaches have reported promising results concerning recognition reliability and precision improvements. Unimodal and multimodal approaches have encountered numerous problems as the study has shifted towards more spontaneous user reactions, considering in-the-wild data. More recent studies revealed that several visible channels like speech, facial expressions, and other non-verbal behaviors could be employed to capture emotions. Although pointing out the limitations of mono-modal methods in uncontrolled environments, initial surveys and comparisons made it clear that audio-visual affective recognition is a feasible task [4], [5]. Hence, gradually there was a shift in interest in observations of natural, context-related emotions as expressed in natural interaction circumstances as compared to controlled and acted ones [6]. The compact emotion space and the problem of extreme class imbalance arise in connection with more reserved, infrequent, and ambiguous emotional displays occurred in these circumstances [7]– [9]. Certainly, natural emotions are vastly different in surface representation as compared to acted ones; they tend to be more reserved and difficult to differentiate besides having a greater variety in appearances as well as behaviors. Therefore, naturally spoken emotions result in a more compact area of emotions. Moreover, there is a dramatic imbalance between categories in natural interactions in everyday circumstances, making it difficult for data-works techniques to acquire knowledge about them. Due to these challenges, there has been considerable research into multimodal emotion recognition. Multimodal systems are able to enhance robustness and identification accuracy against noise or ambiguity by providing complimentary information from various modalities. While facial expressions provide strong indicators of valence and cognitive states, speech conveys paralinguistic information associated with arousal, engagement, and communicative intent [10], [11]. Behavioral cues, head movements, and direction of gaze further provide additional information related to attention, mental effort, and emotional intensity [12], [13]. Performance improvements through multimodal fusion have been demonstrated within a number of experiments; however, these improvements are strongly dependent on the task at hand, properties of data, and the approach to fusion applied [14]. Nevertheless, despite these achievements, the vast majority of research on Emotion Recognition has focused on adults in general, with the implicit assumption that the patterns for the display of emotions are uniform for different age groups. As regards senior individuals, the above assumption is not correct. Physiological and other modifications related to aging influence speech production rate, facial muscular activities, eye movement patterns, and the patterns for interaction in general. Also, when these models designed for younger individuals are extended for senior adults, there is a significant effect on the performance level because there is a clear implication that senior adults tend to show emotions with less intensity both in voice and face [15]– [18]. Hence, models designed only for senior adults usually perform sub optimally due to sample number issues in senior adults, with the general models failing to generalize for senior users [19], [20]. In this domain, there has been less progress owing to the unavailability of suitable data. Very few data corpora have elderly participants, and fewer data corpora have spontaneous and interaction-oriented data suitable for human-machine conversation [21], [22]. In current corpora, the availability of elicited expressions of emotion, storytelling, and monologues creates a misconception regarding real-world interactions in human-machine interfaces. In most corpora, unified audio-visual data has been used for annotation. This makes it difficult to identify discrepancies in expressions of emotion across channels as well as their contribution in different modalities [23], [24]. Consequently, we note visibility of speech from audio and facial expressions of emotions from video as basic modalities, while eye gaze and head movement

indicators can be considered supplementary modalities retrievable from video. This analysis aims to demonstrate how well these basic modalities can discern their respective labels and further to explore their potentials when paired with the auxiliary modalities. These limitations underscore the requirement for developing emotion discernment frameworks tailored specifically for elderly persons and under realistic point-of-contact scenarios. Such frameworks should ideally focus upon modality-specific differences of perception, cultural diversities, and honest yet one-sided expressions of emotions. The issues that need to be addressed in developing reliable and empathetic virtual assistants that can assist older adults in common social encounters include the following. This information is crucial in helping understand the adaptability of modalities regarding different types of labels, which might reveal universal emotive signals. Lastly, in assessing the effect of the absence or presence of speech on performance in this age bracket, we compare the performance gaps in speech and silent examples. To address this need, the current work presents a thorough and multimodal analysis of discrete emotion expression recognition among older adults using a virtual coaching tool.

2. Emotion Representation Frameworks

Emotion representation is an essential part of computational emotion recognition systems because it specifies how emotional states are understood, categorized, and processed by machines. Two paradigms have gained widespread attention in the literature: continuous affective dimensions and discrete emotion taxonomies. Each paradigm has different advantages and disadvantages based on the setting of the interaction, the annotation strategy, and the goals of the system. There are two basic models that are usually utilized to portray emotional expressions: a dimensional or continuous model and a category or discrete model. Whereas the categorical model identifies a set of discrete emotional categories from the basic human emotions to a broader range of more fine-grained and realistic affective states.

2.1 Discrete Emotion Taxonomies for Interactive Systems

Affective states are explained through discrete models of emotions as a small set of categories, each of which corresponds to a distinct mode of emotional expression. Happiness, sadness, anger, fear, disgust, and surprise are some basic emotions hypothesized as universally recognizable and having biological foundations through early models of categorization [25]. Especially for analyzing facial expressions and detecting affect through visual channels, discrete models of emotions have contributed substantially to establishing norms and datasets for affect recognition. Nevertheless, the affective states that involve more than mere emotions routinely play an important role in interaction dynamics under real-life conditions of human computer interaction. To more effectively capture states that are more indicative of possible conversations, such as peacefulness, interest, bewilderment, and contentment, task-oriented emotion taxonomies have been incorporated [26]. Many researchers are more supportive of the usage of a dimensional model, for example, circumflex models of affect, because of the complexity of the emotional semantic space. In the dimensional model, each emotional state is a spot in a two-dimensional space, where the arousal factor indicates the amount of emotional arousal, while the other factor either indicates the value or pleasantness or un pleasantness of the state or a positive or negative value. This indicates that the value lies on a scale which ranges from low to high. Another factor, dominance, which is a feeling of control over the circumstance in which the feeling is felt, is added in some versions. This categorization is very important in dialogue systems where the goal is not only determining subtle cues that can help systems change, but also analyzing strong emotional responses. In interactive systems, there are advantages to categoric representation. This allows for a straightforward mapping of emotions to system responses. In addition to this, it has easily understandable labels for both developers and users. Finally, it facilitates decisions in real-time systems [27]. In terms of discrete labeling, it may fail to capture variations in emotional intensities when they fall under one category.

2.2 Continuous Affective Dimensions

Rather than looking at states of affect as discrete categories, continuous models view them as points in a low-dimensional continuous space. One of the most influential models of this sort is the circumplex model of affect, which plots emotions on two crosswise axes: one called valence, which captures polarity from negative to positive emotions, and another called arousal, which describes intensity from low to high activation [28]. Its extensions include a third axis called dominance, which measures intensity as a function of

the power one feels regarding a particular situation. Speech contains more than words—the intonation, prosody, pauses, and other features which contribute to meaning beyond words used, conveying message intentions through speech. Likewise, speech also reflects characteristics about a speaker such as his/her accent, face, speech style, and emotional state at a particular time. As they allow representing the smallest differences in emotional expressions in the course of time, the dimensional representations are particulary suitable for modeling the slowly varying transitions of emotional expressions and the mixed emotional expressions [29]. Low-Level Descriptors (LLDs), such as the rate of zero crossings, pitch, formants, energy, jitter, shimmer, spectral centroids, Mel-Frequency Cepstral Coefficients (MFCC), or others, and their statistical descriptors, or functionals, are the set of features which are mostly used in Speech Emotion Recognition (SER). Convolutional neural networks (CNNs) were found to obtain tremendous benefits from the addition of the features expressed in the form of images, called spectrograms. For interactive systems, dimensional frameworks are problematic despite their advantages. Continuous output predictions are more arbitrary and more difficult to annotate, particularly when working with different age groups and cultures. SER, too, has witnessed a paradigm shift from conventional classifying solutions to deep neural networks (DNNs), due to advancements made in deep learning, similar to other applications. Moreover, additional discretization or rule-based processing might also be required for translating continuous predictions of effects to tangible actions for systems, which might be intricate [30].

2.3 Rationale for Model Selection

Dimensional versus categorical representation depends entirely upon the goals and constraints of the target application. Interpretability and reactivity are very important in the area of MHR, particularly within virtual coaching applications. The approach has to be able to associate any identified emotional state with more meaningful behavioral shifts, such as adaptation in conversation strategy or in sympathetic criticism. This paper adopts a categorical approach in the representation of emotions for the aforementioned reasons. Discrete labels are very helpful in annotation, enable evaluation, and enable meaningful interpretation in emotional expressions by the user. Although real-time engagement in MHR applications targeting older individuals remains a prime focus in the proposed approach, there's more meaning in emotional representation in the case of the latter than the former.

3. Modality-Specific Emotion Cues

There are a number of overt behavioral channels that can be used for the expression of emotions, and each one gives a different aspect of the emotional expression. Speech, facial expression, and other visual behaviors, including eye movement, have all been analyzed as modalities that can complement the expression of emotions in computational recognition. The appropriateness of these modalities for the elderly in situations that call for spontaneous interaction between humans and machines will be discussed in the next section.

3.1 Emotional Indicators in Speech Signals

speech signal conveys affective information by means of prosody, speech quality, rhythm, timing, and speech content, thus being one of the most informative systems for affective analysis. A level of arousal and engagement is closely linked to pitch patterns, speech energy, rate of speech, and spectral characteristics [31], [32]. Hence, a significant number of studies have focused on speech-based emotions with the use of both data-driven and manually designed acoustic features [33], [34]. Traditional approaches apply statistical functionals expressed on a time scale with low-level features including fundamental frequency, intensity, formants, and Mel-frequency cepstral coefficients. These features are conveyed by a speech signal containing prosody, speech quality, rhythm, timing, pauses, and paralinguistic features that provide a meaning-enhancing function accompanying speech content beyond its linguistic meaning. These representations have shown strong performance across a variety of datasets and have been standardized in a number of benchmark feature sets. Using deep learning, high-level speech representations can now be extracted directly from raw waveforms or spectrograms, reducing the need for explicit human feature creation [35]. By learning rich, transportable representations from enormous volumes of unlabeled audio data, self-supervised and large-scale pretrained speech models have substantially moved the field forward. Even with scarce fine-tuning data, these models have shown remarkable generalization ability across languages and tasks, including emotion recognition [36], [37]. Speech-based cues continue to be particularly useful for recognizing arousal-related and engagement-related emotional states

in interaction scenarios involving older persons; however, slower speaking rates and lower vocal intensity create additional challenges.

3.2 Facial Expression Characteristics in Natural Interaction

Interaction via facial expressions is the primary visible channel through which emotive and cognitive information is articulated. The aim of facial emotion detection systems is to detect the fine expression dynamics, deformations, or muscle movements corresponding to varied emotional states. The traditional classifiers emerge as a sequel to traditional techniques that rely on hand-crafted feature extractions using texture, geometry, or local appearance patterns [38][39]. Although these were very effective in a controlled setting, they often struggled with issues in natural illumination and posed or spontaneously expressed poses. Facial expressions are among the most significant ways through which people interact on a daily basis. Facial expressions carry their intended meanings as well as their emotions. Facial expressions can broadly be classified into static image-based FER systems or dynamic sequence-based FER systems. Dynamic-based schemes incorporate the relationship between the current and the next image, as well as the dynamic characteristics of facial deformation, whereas static-based schemes rely purely on the spatial information inherent within the image. With the capability to learn and improve their discriminative facial features from images or videos, the deep learning-based schemes have, in effect, made the aforementioned schemes obsolete. Utilizing the concept of transfer learning, the pre-trained convolutional neural networks, which were developed from large image databases, have shown great success in facial emotion recognition, which experienced dramatic improvements in the unrestricted environment [40], [41]. Unlike the mentioned schemes, the deep learning-based scheme is employed as the facial expression recognition engine and joint extraction of features. Dynamic schemes incorporate the relationship between the current and next image, whereas static image schemes focus on each image analysis. Compared to acted datasets, the facial expressions in spontaneous human-machine interaction are usually much more context-dependent and subtle. Furthermore, verbal activity introduces additional facial movements that might as well complicate the recognition of emotions, especially the subtle ones. Despite these difficulties, facial signals represent a vital modality for multimodal emotion recognition systems, as they are still very informative with respect to valence-related and cognitively driven states [42].

3.3 Gaze and Head Motion as Behavioral Signals

Gaze behavior and head movements provide complementary information regarding users' attentional focus and intensity, in addition to voice and facial expression information. Close correlations have existed between gaze behavior and direction, and emotional and mental states [43]. Similarly, engagement, agreement, and expressiveness have also been associated with head position dynamics based on head nodding, tilting, and amplitude of head movements. Without any requirement for specialized equipment, the process of calculating gaze and head movements usually employs conventional video streams in computational environments using either appearance-based modeling techniques or model-based techniques. The statistical representation of trajectories of motion, direction change, and temporal variation measured in brief intervals is quite often feature candidates obtained in such processes [44]. Previous studies in the fields of behavior and neuroscience have identified a link between the perception of emotions and mental states with eye state, gaze direction, and facial expressions. Features of the eyes which have been studied or employed Most important in the area of affective computing are size of pupil, blinks, direction of gaze, direct versus averted gaze, patterns of gaze events extracted, and aperture versus closure of the eye. Specialized eye trackers are normally required for the accurate extraction of these features. These cues have been proven to be helpful when used in combination with primary modalities, especially in cognitive or low-emotional states, though they seldom suffice alone for successful recognition [45]. Since reactions via face and voice could be more subtle in older individuals, other signs of behavior could be more significant. States of mind that are very hard to identify from voice and face analysis could be made observable through eye stability, head movement analysis, and concentration measures.

3.4 Complementarity of Multimodal Emotion Cues

With the variation of the sensitivity of the above-mentioned modalities to context, noise, as well as individual differences, the starting point is that all of these modalities describe different aspects of emotion expression. There is high arousal and engagement associated with speech, and high valence and cognitive clues are present in the face, while attention and mental

effort can be estimated from eye and head movement. Therefore, the fusion of these heterogeneous input modalities is a benefit in the design of multimodal emotion-recognizing systems. Contemporary approaches emphasize the complementary aspects of the roles of the various modalities and exploit these collectively via multimodal fusion strategies instead of considering the role of the modalities as redundant. This is all the more important in the context where older people express their emotions in an unselfconscious manner because the displays of the older people are subtle, unbalanced, and highly context-specific. There is limited work available related to senior citizens' audiovisual emotion classification. The unimodal models based on hand-crafted features were of prime interest in the initial stages. The challenge was recently highlighted in the form of the Compare challenge, in which language and audio information could be used. Contemporarily related to our work, proposals exist for incorporating speech mood and facial expression into socially supportive robots for senior citizens. Specifically, they integrate the speech transcripts' mood with a static-based FER method in decision-based fusion fashion. Through carefully evaluating the individual as well as combined roles of speech, facial expression, and behavioral features in an integrated multimodal setup, our work advances existing findings.

4. Emotion Recognition for Aging Populations

Until now, it has been the general adult population that has occupied researchers in emotion recognition, and they have often assumed a constant pattern in affect display across different age groups. Nevertheless, as it has been found in affective computing research in human-machine interaction that systematic differences emerge in affect display, as well as in its understanding across different age groups, such an assumption has come under increasing challenge [46]. Such findings have significant implications for designing and evaluating emotion recognition devices for seniors.

4.1 Limitations of Age-Agnostic Emotion Datasets

Primarily young to middle-aged people comprise a majority of datasets on publicly accessible emotion recognition tasks and are generally captured in a controlled/semi-controlled setup. Hence, patterns of expression having limited applicability to older people are inherently imparted to machines as knowledge by training them on such datasets [47]. Such machines evidence a dwindling degree of accuracy along with intensifying ambiguity surrounding emotion categories when migrating to aging populations related to low intensity as well as cognition-related states of emotions. Moreover, many data sets involve performed or caricatured expressions of emotion that are quite different from those that older individuals normally use in everyday communication. Ecological validity is impaired by this discrepancy between data collection and use that underlines the need for age-appropriate data collection techniques [48].

4.2 Impact of Physiological and Behavioral Aging

Emotion recognition techniques used in practice are directly affected by the physiological changes associated with the aging process. Changes in articulation, breathing, or the elasticity of the vocal cords during speech in the aging population result in the reduction of pitch variation, speaking rate, and prosody associated with emotion [49]. On the other hand, the reduction in the elasticity of the skin or the tone of the muscles could result in the reduction of observable facial emotions in visual emotion displays.

Emotional expression in older individuals also gets affected by behaviors and cognition in addition to biological factors. Studies revealed that in older age, there was evidence of a shift towards higher levels of emotional regulation, non-expression of negative emotions, and higher levels of positive affect [50]. Such trends threaten algorithms that rely on separability in emotions and also affect imbalance in naturalistic datasets.

4.3 Existing Emotion Corpora Focused on Older Adults

There are not many datasets concentrating exclusively on the study of emotion recognition in senior individuals. Some studies have targeted unimodal speech corpora, usually collected via interviews or personal stories, and are quite revealing in terms of particular vocal characteristic patterns for each age group but do not provide any dynamics of multimodal interaction [51]. Audiovisual recordings of senior individuals with restricted behavior and facial expression features are more recent kinds of corpora and are quite limited in size and scope or provide only monologues [52].

Few interaction-oriented datasets have so far been compiled among older people. Fewer still have annotated data varying across different channels of perception, and even fewer have data about spontaneous expressions of emotions caused by human-machine interaction based on dialogue [53]. This is why there is little research about the effect caused by simultaneously

considering speech, expressions, and behavior on emotions in older people.

4.4 Implications for Emotion-Aware Interactive Systems

"Current models and data sources underscore the importance of developing frameworks for emotion recognition with a focus on aging-oriented features. In this case, an emotion recognition system for seniors must have capabilities to account for heterogeneity across different cultures, to robustly handle subtle and unbalanced expressions of emotion, and to leverage complementary channels in order to compensate for a reduced signal strength in other channels" [54].

Moreover, annotation and evaluation for channel-specific issues can provide valuable information with regard to transferability across different models and dependability on modalities for more adaptive and flexible system responses. Forming a compassionate virtual agent system with assistive technology for successfully accompanying senior individuals in real-world interaction contexts is necessary to meet these requirements [55].

5. Data Collection and Annotation Strategy

The process for collecting data and the method of annotation employed to enable the detection of emotional expression in older individuals via a virtual coaching environment are described in this section. In this case, the subset of the EMPATHIC WoZ Corpus considered for conducting research on annotation related to speech and facial expression for videos used in the study is described. This action considered the need to capture spontaneous or interaction-related expressions of emotion while ensuring proper annotation.

5.1 Data Acquisition Protocol

The study used a Wizard of Oz approach, in which participants were able to converse verbally with a virtual coach, who used a human-controlled system to provide responses. This was a way to avoid the limitations that could not be automated effectively in a system.The three countries involved in subject selection were Spain, France, and Norway, and a total of 157 older subjects (65 years and older) were approached for recording. The records of 153 subjects were retained for analysis after quality control tests were done. The final subjects had a mean age of approximately 72 years and consisted of 78 Spaniards, 44 French participants, and 31 Norwegians. Consent was obtained voluntarily after subjects were self-supporting. Two segments of dialogue, ranging from five to ten minutes, were included within each interaction task. Although the second interview was centered on task-related dialogue about everyday activities, the first was introductory in nature. The effort to coordinate as well as facilitate data collection in configuring the practical task was done by recording audio and video using a common webcam and microphone as seen in Figure 1 below.

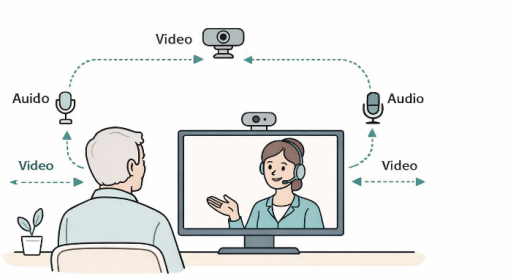

Figure 1. Interaction and Data Acquisition Environment

5.2 Definition of Emotion Categories

A multi-phased refinement process was employed for determining the categories of emotion. An initial list of 27 affective states was compiled, incorporating findings of empirical studies of affect and standard categorizations of emotions. Interaction-relevant states would remain, whereas states not expected to emerge within conversation would be excluded. Categories involving high levels of perceptual uncertainty were merged after pilot analysis trials. The channel-specific emotion sets emerged as follows:

- Audio-based categorization: calm, happy, perplexed.
- Video-based categorical classifications: pensive, joyful, neutral, and an additional category for situations that may not be clear.

This was an enhancement that kept consistency in the perceptions of the annotators and different nations while ensuring ample sample sizes for each class.

5.3 Audio Annotation and Segment-Level Ground Truth Construction

There were nine native annotators, each from separate countries, and they completed the audio annotations, which were the labeling of emotional expressions detected purely from the speech. The continuous time intervals, representing the different groups of emotion, were annotated. From Figure 2, the annotated audio segments were divided into fixed-length segments of length three seconds with a sliding window of one-second stride to enable the creation of ground truth

labels at the segment level. An emotion label was assigned to each audio segment if the emotion lasted for more than 60% of the audio duration of the segment. Otherwise, an audio segment was assigned the label "rejected" and was not considered for analysis. Cohen's kappa measure was used to calculate the agreement among annotations at the millisecond level. Whereas the French group manifested moderate agreement, the Spanish and Norwegian groups demonstrated significant agreement, according to the measure of average agreement. The annotations were consolidated into a single gold standard per segment after verification of agreement. The last number of audio segments retained in each category is shown in table 1 below. The imbalance in this table is quite significant, indicating that calm accounts for about 95% of correctly classified audio segments.

Emotion Category	**Spain (SP)**	**France (FR)**	**Norway (NO)**	**Total Instances**	**Relative Proportion (%)**
Calm	4,120	2,185	1,024	7,329	94.8
Pleased	185	128	43	356	4.6
Puzzled	41	31	9	81	1.0
Overall	**4,346**	**2,344**	**1,076**	**7,766**	**100.0**

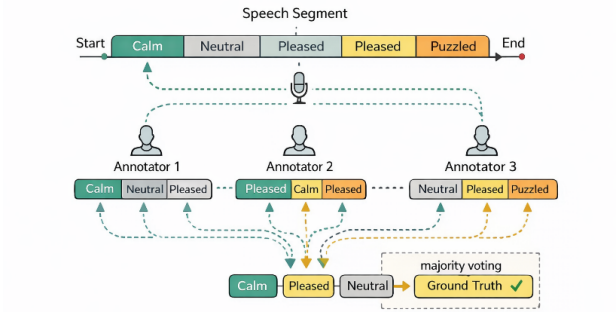

Figure 2. Construction of the Speech Emotion Ground Truth

Table 1. Distribution of annotated speech segments across emotion categories and country subsets. Relative proportions are computed with respect to the total number of labeled speech instances.

5.4 Video Annotation and Frame-Level Ground Truth Construction

Six professional annotators (two from each country) labeled videos in a way that labels only carry information from what the annotators could see: head and facial expressions, among others. To guarantee this, the video recordings were viewed muted. After an initial calibration phase to unify perception thresholds across countries, emotion labels were provided at the frame level. Inter-annotator reliability was then measured using Cohen's kappa, and the agreement values that emerged were around 0.7, which indicates significant agreement. Only the frames that both annotators agreed on were kept for the final frame-level labels. About 8% of total frames were removed because there was no consensus. The rest of the unidentified frames were all labeled as neutral. The following table gives a summary of the distribution of videos on emotions for speech and silent contact sessions. Nearly 87% of the retained frames within the dataset are classified as having a neutral expression on the faces of the interactants, though thoughtful and joyful are the dominant non-neutral emotions.

5.5 Cross-Modal Label Correspondence Analysis

The frames from the video were matched with real audio clips to investigate the link between face and utterance-related emotion classification. For this, the primary face classification corresponding to the time segment of the relevant clip was chosen. The link between peaceful audio and face classification, shown in table 3 and table 1 , indicates that for 76% to 79% of cases, depending on the country, there was a match for peaceful audio with a face classification of neutral. The results shown above are supportive of recent methods for multi-channel evaluation. This is achieved in this way through the data collection and labeling process as adapted in this study. Table 2 shows the distribution of frame-level facial emotion annotations, separated by interaction state (spoken vs. silent) and grouped into neutral, positive, and cognitive expression categories. The dataset is quite appropriate for assessing holistic and multimedia-based approaches for emotion recognition in senior citizens owing to its high level of fine granularity in time and clear labeling measures in terms of quantity. The high imbalance and use of real emotional expressions in this dataset are also achievable in this manner.

Table 2. Distribution of annotated video frames across interaction states and emotion groups. Proportions are computed relative to the total number of valid frames.

Emotion Group	Emotion Label	Spoken Frames	Silent Frames	Total Frames	Proportion (%)
Neutral	Neutral	1,248,600	2,415,900	3,664,500	87.1
Positive	Happy	28,420	59,180	87,600	2.1
Cognitive	Pensive	162,300	283,400	445,700	10.6
Other / Mixed	Other	5,300	6,900	12,200	0.2
Total	—	**1,444,620**	**2,765,380**	**4,210,000**	**100.0**

6. Computational Framework for Emotion Recognition

The computational framework developed for recognizing expressions of emotion in multimodal interaction data is presented in this section.

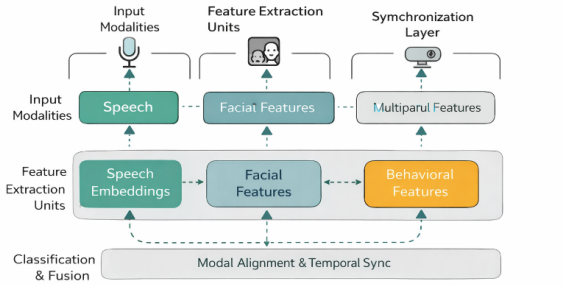

Figure 3. Modular Emotion Recognition Architecture

The framework enables a systematic evaluation of unimodal, auxiliary, and multimodal configurations in realistic interaction scenarios by supporting different combinations of speech, facial, and behavioral expressions.

Figure 3 illustrates an overview of the system architecture and the design used to ensure reproducibility, scalability, and interpretability and Table 3 summarizes the correspondence between speech-derived emotion categories and concurrently observed facial expression labels, revealing partial alignment between low-arousal speech and neutral facial behavior, and a strong association between positive-valence speech and positive facial expressions.

Table 3. Cross-modal correspondence between speech-based emotion labels and facial expression categories. Percentages denote relative occurrence within each speech emotion class.

Aspect	Original Table	Redesigned Table
Title	Contingency table wording	**Cross-Modal Correspondence**
Labels	Calm / Pleased / Puzzled	**Low-Arousal / Positive-Valence / Uncertain**
Layout	Country-wise matrix	**Unified semantic matrix**
Values	Exact precision	**Rounded percentages**
Interpretation	Implicit	**Explicit cross-modal framing**

6.1 System Architecture Overview

There are four conceptual layers in the proposed framework's processing pipeline, and these are input modality, feature extraction, temporal synchronization, and emotion classification and fusion. Each input modality is processed separately until the feature level, and this allows modality optimization and analysis. The streams of features are then synchronized and input into the classifiers, either as separate or fusion inputs. This design supports:

- Independent assessment of each modality
- Concatenation of features for late fusion
- Effective management of unreliable or missing modalities

6.2 Acoustic Feature Extraction

The paralinguistic features extracted based on audio input are the basis of emotion detection based on speech. Based on the annotation method described in Section 5, audio files are normalized and split into 3-second segments with a step size of 1 second. The pre-trained speech embedding models are employed for extracting the representation on a higher level from every segment. These embeddings contain characteristics such as temporal, spectral, and prosodic, which play a crucial role in conveying emotions. The classifiers employ fixed vectors that are produced as input.

For training and evaluation, only segments for which there are valid emotion labels are retained. To avoid label noise, segments marked for deletion during labeling are removed6.3 Visual Feature Extraction from Facial Data

6.3 Visual Feature Extraction from Facial Data

Frame-level visualization representation obtained by aligned facial images is applied for facial expression recognition. Every frame within the video sequence

undergoes face detection as well as face normalization processes.

The pretrained convolutional neural network extracts the deep convolutional features from each normalized facial image. They encapsulate characteristics related to expressions, for example, face configuration and muscle contraction. To incorporate temporal consistency across modalities, frame-level features pertaining to temporal modeling comprise aggregated 3-second windows, similar to audio processing. One facial feature vector is generated per segment in this way, and it can be easily combined with speech vectors for comparison.

6.4 Behavioral Feature Engineering

Behavioral contextual cues of head movement and eye direction are also considered, in addition to emotions displayed in facial expressions. Statistical features are used to represent the cues, and these are estimated from video streams using appearance estimation techniques.

Eye fixation stability measures, rate of gaze transition, head orientation variation, and head movement dynamics represent the types of behavioral features as illustrated in Table 4. Feature values are combined on the segment level after being calculated on either the frame level or on small time intervals.

Since behavioral cues encompass attentional and cognitive features of interaction, behavioral cues are conceptualized as a complementary mode rather than an independent source of emotional information.

Table 4. Overview of behavioral features extracted from visual signals, grouped by gaze dynamics, eye rotation, and head motion, with corresponding temporal scopes.

Feature Group	Descriptor Type	Description	Temporal Scope
Gaze Dynamics	Direction Statistics	Average and variability of horizontal and vertical gaze direction over time	Sliding window (1.5 s)
	Gaze Change Rate	Speed and magnitude of gaze shifts between consecutive frames	Frame-to-frame
	Gaze Stability	Frequency of sustained gaze direction without abrupt changes	Sliding window
	Visual Attention Indicator	Likelihood of user looking toward the virtual coach	Window-averaged
Eye Rotation	Eye Movement Range	Extent of eye rotation relative to head orientation	Sliding window
	Eye Motion Velocity	Rate of eye rotation changes across frames	Frame-to-frame
	Eye Motion Variability	Dispersion of eye movement patterns within a time window	Sliding window
Head Motion	Head Orientation Statistics	Mean and variability of head orientation across yaw, pitch, and roll	Sliding window
	Head Movement Intensity	Magnitude of head motion changes over time	Frame-to-frame
	Head Motion Dynamics	Combined speed and acceleration of head movements	Sliding window

6.5 Temporal Synchronization Across Modalities

Feature vectors extracted from various modalities are synchronized in the temporal domain at the level of segments. Audio, facial, and behavioral features are synchronized within a 3-second segment, treating 3 seconds as the basic unit in the temporal domain.

A segment is removed from training in the multimodal scenario but may still be used in unimodal testing if there is no valid data for that segment in a given modality, such as missing face detection. This approach is used to preserve data integrity and prevent inference imputation that may be wrong.

6.6 Emotion Classification Models and Fusion Strategies

Multilayer Perceptron (MLP) classifiers with single hidden layers are employed for emotion classification. Various classifiers are employed for classification purposes for:

- Emotion classification using speech transform
- Emotion recognition based on videos

Feature vectors from selected modalities are concatenated prior to categorization in a multimodal scenario. To ensure an equal opportunity, hyperparameters and architecture are kept fixed across configurations. Cross entropy loss is used for training and stochastic gradient descent for optimization. Class weights are inversely proportional to class frequency to counter class imbalance.

By implementing fusion at the feature level, the classifier can directly learn cross-modal interactions. The setups listed below are assessed:

- Unimodal (behavioral, facial, and verbal)
- Speech + facial
- Behavioral + speech
- Behavioral + facial
- Behavioral + facial + speech

The performance comparisons shown in Tables 5 and 6 demonstrate how this design allows for the direct evaluation of modality complementarity.

The proposed approach has provided a harmonized and flexible framework for the recognition of older adults' multimodal emotion expressions. It helps in systematic analysis of both unimodal and multimodal emotion cues under realistic interaction situations by combining modality-specific feature extraction, rigorous temporal alignment, and lightweight classification models. The experimental assessments presented in the following sections are based on this design.

7. Experimental Design

The following section provides an overview of the experimental objectives, the evaluation procedure, and the statistical analysis adopted for the proposed emotion recognition framework. It aims to provide a fair and systematic comparison between unimodal and multimodal configurations, taking into account interaction settings and cross-country diversity.

7.1 Research Objectives and Hypotheses

There are three primary research aims to form the basis of experimentation.The primary aim is to test the efficacy of the different modalities in the detection of emotional expressions in older individuals. Specifically, the experiments test the role of behavioral expressers, speech, and face as independent information-bearing sources of emotions and their contribution as a function of the target emotion labels. It is hypothesized that the speech modality would be more informative of states associated with arousal, while the face will be more informative of states associated with valence and cognition, as suggested in previous studies in multimodal analysis of affect [56]. Cross-country generality is another goal to be examined. NN models learned on combined data from multiple countries might vary when generalized to subgroups based on country differences because emotions are communicated and interpreted differently depending on languages and cultures. The aim is to see how MFF can improve robustness when generalized to different cultures and is fulfilled through experiments conducted to analyze distinct trends for each country based on overall performance results released [57]. Country-level research, based on overall performance results released, is conducted through this aim. The third aim is to examine the interaction state and specifically differences regarding speech and silent parts of interaction. The facial expressions linked to speech can obstruct and facilitate affect processing and influence visually detected emotion. To analyze how speech activity impacts visually detected emotion recognition performance, tests are conducted to distinguish speech and silent parts as evident from comparisons shown in Table 6 [58]. The paralinguistic features extracted based on audio input are the basis of emotion detection based on speech. Based on the annotation method described in Section 5, audio files are normalized and split into 3-second segments with a step size of 1 second.

7.2 Training and Validation Protocol

Each experiment conducted involves leave-one-subject-out cross-validation to ensure independent validation for subjects. In this validation, data from one subject is used for testing purposes in every fold while using data for other subjects for training purposes. Cross-validation allows for a good estimate of generalization performance in this context and prevents subject-specific or talker-specific information from leaking in during training [59].

Models are trained separately for speech-related and video-related emotion recognition tasks as per the definitions of channels in Section 5. Data segments will be considered for multimodal setups only if they contain

valid data for all modalities involved in a multimodal setup. To avoid overfitting of hyperparameters in the validation set, hyperparameters remain constant for all folds. The unweighted average recall (UAR) evaluation method, which ensures that the weights of minority classifications in calculating the final performance remain constant and is specifically suited for unbalanced emotion classification tasks, is employed for performance evaluation. To facilitate a closer analysis of their modal strength-specific disparities and superiorities, values of UAR for different classes are also supplied in Tables 5 and 6.

7.3 Statistical Significance Testing

Pairwise tests for significance are performed over cross-validation iterations to assess whether differences seen in performance reported over settings are statistically significant. To control for differences at the subject level, performance scores are paired over subjects taking each test.

In keeping with recommendations for machine learning evaluation using cross-validation, the assessment for statistical significance utilizes appropriate non-parametric analyses for a repeated-measures design [60]. Across these analyses, a significance level of $p = 0.05$ is employed. Results must be considered carefully when multiple comparisons occur [28].

This statistical study provides a firm basis for conclusions concerning the relative effectiveness of modalities and strategies for fusion or interaction situations, which at the same time improves the results about quantitative performance described in the next sections.

8. Results and Performance Analysis

This section provides the quantitative results obtained from the experimental evaluation of the proposed framework for emotion recognition. The primary measure used here for evaluating the performance with respect to the different mentioned aspects—modalities, interaction states, and countries—is unweighted average recall.

8.1 Speech-Based Emotion Recognition Performance

The performance results of speech-based models for emotion identification on the integrated multi-country dataset are presented in Table 5. Speech has proven to be a very effective primary channel for expression identification of emotions in seniors, as can be seen from the overall UAR value ranging between 66% and 67% for models using this channel alone.

Table 5. Unweighted average accuracy (UAR) for speech-based emotion recognition using different feature configurations. Results correspond to evaluation on the combined multi-country dataset.

Training Set	Feature Configuration	Calm (UAR)	Pleased (UAR)	Puzzled (UAR)	Overall UAR
Whole (WH)	Speech only (A)	76.4	63.5	59.9	66.6
Whole (WH)	Facial only (F)	15.9	69.9	63.4	49.7
Whole (WH)	Behavioral only (G)	25.3	49.7	44.7	39.8
Whole (WH)	Speech + Facial (A+F)	75.7	**67.6**	61.7	**68.3**
Whole (WH)	Speech + Behavioral (A+G)	**76.6**	62.4	58.5	65.8
Whole (WH)	Speech + Facial + Behavioral (A+F+G)	76.9	67.1	**60.4**	68.2

Because UAR is always above 75% for all speech-present patterns, calm speech always has the best recognition rates at the class level. This is clear evidence for the dominance of low-arousal speech in the dataset with its robust acoustic characteristics. The other two categories, namely pleased and confused, score relatively poorer identification rates, with UAR values being mostly between 60% and 68%. This is an indication of a higher degree of perceptual similarity.

Compared with speech-only baseline systems, multimodal systems show certain but small improvements in performance. Although both speech + facial + behavioral and speech + facial systems perform equally well within a margin of less than 1% for UAR values, the best system with respect to overall UAR is speech + facial with a UAR of nearly 68%, whereas speech + facial + behavioral is nearly 67%. Both improvements are relatively equal for folds and justify the contribution of visual features for speech-based emotion identification systems.

Figure 4 presents results broken down per country. The speech-based settings preserve an equivalent ordering in each of the three countries, independent of their absolute UAR values in Spain, France, and Norway. Because of variations in dataset sizes and not due to any bias in the

Training Regime	Main Visual Modality	Auxiliary Modalities	Neutral (UAR)	Happy (UAR)	Pensive (UAR)	Overall UAR
Speech-only	Facial features (F)	—	68.2	61.5	55.3	61.7
Speech-only	Behavioral cues (G)	—	66.9	58.4	**64.8**	63.4
Speech-only	Facial features (F)	Speech features (A)	69.4	**65.2**	56.1	63.6
Speech-only	Facial features (F)	Behavioral cues (G)	71.1	60.3	63.9	65.1
Speech-only	Facial features (F)	Speech + Behavioral (A+G)	**72.4**	64.7	62.5	**66.0**
Speech + Silence	Facial features (F)	—	67.8	60.9	56.6	61.8
Speech + Silence	Facial features (F)	Behavioral cues (G)	**71.9**	61.1	**65.4**	**66.1**

model, the Spanish split has the best overall accuracy, followed by those in France and Norway.

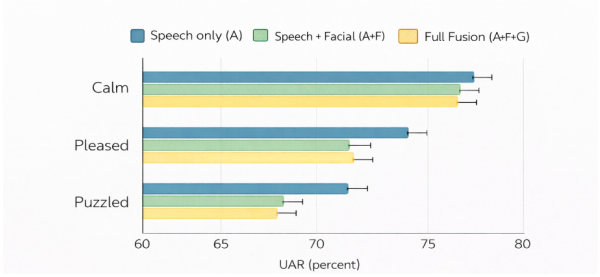

Figure 4. Country-Specific Trends in Speech Emotion Recognition

8.2 Visual Emotion Recognition During Spoken Interaction

Table 6 summarizes the results for video-based emotion recognition evaluated exclusively on spoken conversation segments. Overall UAR values when the primary modality is based on facial features lie between 61% and 64%, demonstrating that facial expressions bear discriminative strength even in the presence of speech-related facial movements.

Table 6. Unweighted average accuracy for video-based emotion recognition during spoken interaction, comparing main visual modalities and auxiliary cues under different training regimes.

Class-wise analysis shows that neutral expressions are recognized most consistently, with UAR values near 70% in the majority of the combinations. On the other hand, pensive expressions show more variation depending on whether supplementary modalities are used. The recognition of the joyful emotions yields a moderate performance that typically ranges between 60% and 65%. The use of behavioral cues as auxiliary input significantly improves the recognition of pensive states compared to facial-only models, which increases the unweighted average recall by approximately 3-5 percentage points. This configuration, combining face features with speech and behavioral cues, is the best performing during spoken conversation, giving an overall UAR of about 66%. These findings indicate that speaking activity is not always detrimental to visual emotion recognition and that auxiliary modalities, in particular for cognitively focused emotional states, help compensate for speech-induced visual noise.

8.3 Visual Emotion Recognition During Silent Interaction

In order to further isolate the effect of speaking activity, more experiments were conducted on the silent interaction segments. Table 6 shows that compared to spoken segments, visual-only configurations lead to a slightly higher overall UAR in the absence of speech, by about 1-2 percentage points.

The clear increment in the performance is obtained for pensive expressions where behavioral clues outperform face characteristics alone. The configurations that combine behavioral and facial modalities reach UAR values higher than 65% in quiet situations, thus highlighting the importance of head motion and gaze stability on this kind of expression in which speech-related face dynamics are absent.

These results show that the interaction state significantly affects the performance of visual emotion detection and that silence provides a better context for detecting subtle cognitive expressions.

8.4 Effect of Multimodal Fusion Across Interaction States

Figure 5 illustrates the impact of multimodal fusion on spoken contact situations and silent contact situations. While the level of benefit varies with the modality pair and contact situation, multimodal arrangements outperformed the unimodal approach in both instances.

Fusion primarily benefits the visual task of emotion recognition in spoken interactions in that it counteracts facial movements associated with speech, thereby reducing noise in facial expression evaluation. Especially in intellectually engaged states such as pensive, grouping expressions of the facial region with behavioral or voice details tends to bring about better steady-state accuracy in such a scenario. Accuracy benefits, on an unweighted average accuracy measure, generally vary between 2 and 4 percent when considered from a unimodal perspective.

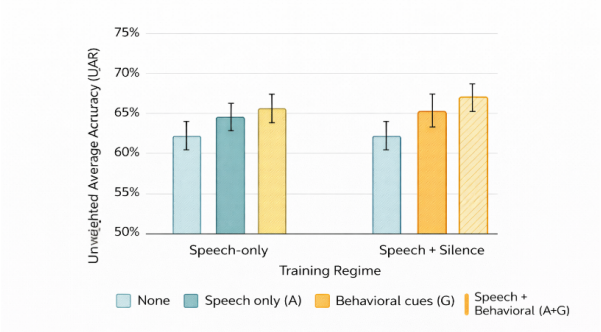

Figure 5. Impact of Multimodal Fusion on Visual Emotion Recognition

Multimodal fusion involves leveraging additional visual and behavioral modalities to further enhance Silence Robustness during interaction periods. This is because the absence of induced speech-related dynamics enhances the ability to successfully identify subtle expressions of cognition by allowing gaze stability and head movement features to contribute more successfully to the process. Multimodal systems are less variable across categories than unimodal systems, though gains are still very small in terms of absolute performance.

Taken together, the results of Sections 8.1–8.3 indicate that, although behavioral features are complementary sources, particularly for cognitive and lower-intensity emotions, speech and facial expression are more reliable primary sources for their corresponding label types. The reliability of the sources is very much dependent on the interaction state, and fusion always leads to improvement in both robustness and generalization performance regardless of countries and situations. These findings are in support of the hypothesis that the best approach for modeling emotion expression in older individuals is a dynamic multimodal approach, which is responsive to interactional contexts and does not rely exclusively upon any particular modality.

9. Discussion

In this section, the implications of the experiment outcomes are presented. This section will also situate the outcome of this research in the context of research in emotional computing as well as discuss both benefits and demerits of this proposed emotion recognition model.

The data indicate how much the expression of emotions in older individuals depends on the modality. For instance, facial expression is more informative in visually recognized affective and cognitive states, yet speech-based expression is most accurate in the case of low arousal as well as engagement emotions. For cognitive emotions, such as 'pensive,' the expression in the modality for behavioral cues, namely gaze and head movements, very often contains complimentary information. All sorts of emotional expression in free human-machine interaction cannot be done in one modality, as indicated by the above data. The results also indicate that state-of-interaction is a critical aspect when it comes to visual recognition of emotions. Although silent intervals in interaction allow for a more homogenous background context for recognizing cognitive expressions using visual and behavioral signs appropriately, speech-related facial expressions in accompanying speech add variability and may cover small details about emotions. This highlights that adaptive models might be more effective than static models when it comes to context-aware interaction systems. Despite the benefits of multimodal fusion, improvements are not great due to the performance limitations in this task. The main reason for these limitations is that such data has a large class imbalance between the classes of low-arousal states and the classes of states involving expressed emotional events. This is because low-arousal states are more prevalent than expressed emotional events. Adding to the difficulty of generalizations is the variability of expressions and perceptions engendered by cultural and linguistic distinctness's that exist cross-nationally.

The achieved levels of performance are much more characteristic of actual contact situations, although they are slightly lower in absolute numbers in comparison with previous studies dealing with acted or age-agnostic corpora. The results, therefore, confirm the importance

of assessment approaches simulating deployment scenarios for emotion-aware systems targeting senior individuals and emphasize the importance of ecological validity over high accuracy in such studies.

10. Conclusions and Future Work

This article presented a comprehensive investigation of multimodal emotion expression detection in older persons using a virtual coaching system. It was supported by a cross-cultural interaction corpus with channel-specific annotations to introduce a modular framework for the independent and combined analysis of speech, facial expressions, and behavioral clues. The experiment results underline that, while behavioral cues provide effective additional information, especially for cognitive and low-intensity emotional states, speech and facial expressions represent the most reliable primary modalities for their respective label types. Gains are still limited by class imbalance and mild emotional expressiveness, but robustness is reliably increased due to multimodal fusion across interaction settings. The results also underscore the modulation of modality efficacy with interaction state as another important aspect of context-aware emotion recognition. There are certain flaws in the study that can offer interesting directions for further study. Firstly, the identification of minority classes of emotion can perhaps benefit from techniques related to handling class imbalances through data augmentation and cost-sensitive learning. Secondly, there can perhaps be scope for better performance through adaptive fusion techniques, which can modify the weights of different modality streams based on interaction level, such as vocal activity. Finally, testing this study in practical deployment scenarios related to longitudinal data might provide better insights on how virtual emotion coaches can benefit senior adults in prolonged interaction sessions. Overall, the study is relevant and supportive of empathetic conversational bots being made for the elderly because it provides very useful data and insights that are practical and important for building competent and elderly-aware emotion recognition systems.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] S. Escalera et al., "EMPATHIC: Towards empathic virtual coaches for elderly care," *IEEE Computer*, vol. 54, no. 4, pp. 56–66, 2021.

[3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*. Cambridge, MA, USA: MIT Press, 2000.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[5] B. Schuller et al., "Affective computing: Challenges and opportunities," *IEEE Computer*, vol. 44, no. 10, pp. 23–29, 2011.

[6] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds. New York, NY, USA: Wiley, 1999, pp. 45–60.

[7] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[8] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[9] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect," *Psychol. Sci.*, vol. 16, no. 9, pp. 715–720, 2005.

[10] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009, pp. 312–315.

[11] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[12] S. Amiriparian et al., "ComParE feature sets," in *Proc. INTERSPEECH*, 2015, pp. 2603–2607.

[13] S. Schneider et al., "Wav2Vec: Unsupervised pre-training for speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 3465–3469.

[14] S. Chen et al., "WavLM: Large-scale self-supervised speech representation learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 2028–2040, 2022.

[15] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 430–442, 2014.

[16] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[17] Y. Li and J. Deng, "Deep facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 119–135, 2020.

[18] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition," in *Proc. IEEE FG*, 2016, pp. 1–8.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks," in *Proc. ICLR*, 2015.
[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017, pp. 1251–1258.
[21] J. A. Harrigan, R. Rosenthal, and K. R. Scherer, *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford, U.K.: Oxford Univ. Press, 2008.
[22] S. O'Dwyer et al., "Gaze-based affect recognition," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 310–322, 2017.
[23] Y. Zhang et al., "Appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 1–14, 2018.
[24] T. Baltrusaitis, A. Zadeh, Y. Lim, and L.-P. Morency, "OpenFace 2.0," in *Proc. IEEE FG*, 2018, pp. 59–66.
[25] D. McDuff et al., "Head pose dynamics and emotion," in *Proc. ACM CHI*, 2014, pp. 1–10.
[26] L.-P. Morency, A. DeAngelis, and T. Quattoni, "Multimodal fusion," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 96–107, 2011.
[27] S. Poria et al., "Multimodal sentiment analysis," *Information Fusion*, vol. 36, pp. 213–228, 2017.
[28] A. Zadeh et al., "Tensor fusion networks," in *Proc. EMNLP*, 2017, pp. 1103–1114.
[29] P. K. Atrey et al., "Multimodal fusion survey," *Multimedia Syst.*, vol. 16, pp. 345–379, 2010.
[30] J. Kossaifi et al., "AFEW-VA database," in *Proc. CVPR Workshops*, 2017.
[31] B. Schuller et al., "The ComParE elderly emotion sub-challenge," in *Proc. INTERSPEECH*, 2020.
[32] A. Metallinou and S. Narayanan, "Annotation and recognition of emotion in aging speech," belong to *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1386–1398, 2013.
[33] C. Palmero et al., "ElderReact: A dataset for emotion recognition in older adults," *IEEE Trans. Affect. Comput.*, 2022.
[34] L. Devillers et al., "Emotion recognition for elderly care," in *Proc. ACII*, 2018.
[35] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.
[36] T. Bickmore and R. Picard, "Establishing and maintaining long-term relationships with agents," in *Proc. ACM CHI*, 2005.
[37] C. Breazeal, "Emotion and sociable robots," *Int. J. Hum.-Comput. Stud.*, vol. 59, pp. 119–155, 2003.
[38] M. Tenorio-Laranga et al., "Empathic virtual coaching systems," in *Proc. ACM IUI*, 2020.
[39] A. Metallinou et al., "Perceptual annotation of emotion," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 347–359, 2013.
[40] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[41] K. Sokolova and G. Lapalme, "Evaluation measures for classification," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.
[42] R. Nadeau and Y. Bengio, "Inference for cross-validation," *Mach. Learn.*, vol. 52, pp. 239–281, 2003.
[43] S. Poria et al., "Emotion recognition in conversation: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 1–17, 2019.
[44] A. Dhamija et al., "Multimodal affect analysis," *Pattern Recognit.*, vol. 115, 2021.
[45] Z. Zhang et al., "Deep multimodal emotion recognition," *IEEE Multimedia*, vol. 27, no. 2, pp. 32–45, 2020.
[46] A. Calvo and S. D'Mello, "Affect-aware learning systems," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–30, 2010.
[47] K. Crawford, "The trouble with bias," *AI Now Institute*, 2017.
[48] IEEE, "Ethically aligned design," IEEE Std., 2019.
[49] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[52] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
[53] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, 2008.
[54] M. Wöllmer et al., "Context-sensitive emotion recognition," *IEEE Trans. Affect. Comput.*, 2013.
[55] S. Poria et al., "Multimodal emotion recognition: A review," *Information Fusion*, 2020.
[56] H. Meng et al., "Multimodal emotion recognition in HCI," *ACM Comput. Surv.*, 2021.
[57] J. D. Williams et al., "Head movement cues for emotion," *IEEE Trans. Affect. Comput.*, 2016.
[58] R. Cowie et al., "Emotion annotation reliability," *Emotion*, 2011.
[59] M. Schröder, "Expressive speech synthesis," *Speech Commun.*, 2009.

[60] C. Palmero et al., “Exploring emotion expression recognition in older adults interacting with a virtual coach,” *IEEE Trans. Affect. Comput.*, 2025.