







Article

A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset

Cristina Luna-Jiménez ^{1,*} , Ricardo Kleinlein ¹ , David Griol ² , Zoraida Callejas ² , Juan M. Montero ¹ 
and Fernando Fernández-Martínez ¹ 

¹ Grupode Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense 30, 28040 Madrid, Spain; ricardo.kleinlein@upm.es (R.K.); juanmanuel.montero@upm.es (J.M.M.); fernando.fernandezm@upm.es (F.F.-M.)

² Department of Software Engineering, CITIC-UGR, University of Granada, Periodista Daniel Saucedo Aranda S/N, 18071 Granada, Spain; dgriol@ugr.es (D.G.); zoraida@ugr.es (Z.C.)

* Correspondence: cristina.lunaj@upm.es

Abstract: Emotion recognition is attracting the attention of the research community due to its multiple applications in different fields, such as medicine or autonomous driving. In this paper, we proposed an automatic emotion recognizer system that consisted of a speech emotion recognizer (SER) and a facial emotion recognizer (FER). For the SER, we evaluated a pre-trained xlsr-Wav2Vec2.0 transformer using two transfer-learning techniques: embedding extraction and fine-tuning. The best accuracy results were achieved when we fine-tuned the whole model by appending a multilayer perceptron on top of it, confirming that the training was more robust when it did not start from scratch and the previous knowledge of the network was similar to the task to adapt. Regarding the facial emotion recognizer, we extracted the Action Units of the videos and compared the performance between employing static models against sequential models. Results showed that sequential models beat static models by a narrow difference. Error analysis reported that the visual systems could improve with a detector of high-emotional load frames, which opened a new line of research to discover new ways to learn from videos. Finally, combining these two modalities with a late fusion strategy, we achieved 86.70% accuracy on the RAVDESS dataset on a subject-wise 5-CV evaluation, classifying eight emotions. Results demonstrated that these modalities carried relevant information to detect users' emotional state and their combination allowed to improve the final system performance.

Keywords: audio-visual emotion recognition; human–computer interaction; computational paralinguistics; xlsr-Wav2Vec2.0 transformer; transformer; transfer learning; Action Units; RAVDESS; speech emotion recognition; facial emotion recognition



Citation: Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.M.; Fernández-Martínez, F. A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. *Appl. Sci.* **2022**, *12*, 327. <https://doi.org/10.3390/app12010327>

Academic Editors: Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira

Received: 22 November 2021

Accepted: 27 December 2021

Published: 30 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Emotions play a crucial role in our life decisions. Comprehending them awakens interest due to their potential applications since knowing how others feel allows us to interact and transmit information more effectively. With the help of an emotion recognizer, other systems could detect loss of trust or changes in emotions by monitoring people's conduct. This capability will help specific systems such as Embodied Conversational Agents (ECAs) [1,2] to react to these events and adapt their decisions to improve conversations by adjusting their tone or facial expressions to create a better socio-affective user experience [3].

Automobile safety is another important application of facial expression recognition. Detecting stress, rage, or exhaustion may be decisive in preventing traffic accidents [4] on intelligent vehicles by allowing cars to make decisions based on the driver's current psychological state. Another use for these systems is human–machine interaction in an assisted living experience for the elderly [5]. An emotion recognizer could monitor the emotional state of a person to detect anomalies in their behavior. When an anomaly arises, it could mean that the person requires attention. Additionally, the emotion recognizer could

be practical in the diagnosis of certain diseases (depressive disorders [6,7], Parkinson [8], and so on) by the detection of deficits in the expression of certain emotions, accelerating the diagnosis as well as the patient's treatment. Emotion recognizers will also be necessary for the future 'next revolution' [9], which will require the creation of social robots. These robots should know how to recognize people's emotions and convey and produce their own emotional state to display closer personal relationships with humans.

Historically, state-of-the-art speech emotion recognition systems (SER) have had low accuracy and extensive processing costs [10,11]. Currently, some models can work in real-time and demonstrate high performance in these settings, as exhibited in the work of Anvarjon et al. [12]. They suggested a lightweight CNN model with plain rectangular kernels and modified pooling layers, attaining state-of-the-art performance on IEMOCAP and EMO-DB datasets.

As a continuation of previous proposals, we developed a solution to categorize emotions from two sources of information: speech and facial features. These two modalities are combined to detect users' emotional states through independent models connected by a late fusion strategy.

As a summary, the main contributions of this study were:

- We implemented a speech emotion recognizer using a pre-trained xlsr-Wav2Vec 2.0 model on an English speech-to-text task. We analyzed the performance reached using two transfer-learning techniques: feature extraction and fine-tuning.
- This work also incorporated visual information, which is a rarely used modality on RAVDESS ('The Ryerson Audio-Visual Database of Emotional Speech and Song') due to the difficulties associated with working with videos. However, our results showed it is a valuable source of information that should be explored to improve current emotion recognizers. We designed a facial emotion recognizer using Action Units as features and evaluated them on two models: static and sequential.
- To our knowledge, our study is the first that assembles the posteriors of a fine-tuned audio transformer with the posteriors extracted from the visual information of the models trained with the Action Units on the RAVDESS dataset.
- We also leveraged our code to allow the replication of our results and the set-up of our experiments. In this way, we expect to create a common framework to compare contributions and models' performance on the RAVDESS dataset. We decided to continue with the formulation of our previous paper of Luna-Jiménez et al. [13] that consisted of a subject-wise 5-CV technique based on the eight emotions captured in the RAVDESS dataset.

The rest of the paper is organized as follows. Section 2 summarizes the related works on RAVDESS and other datasets. Section 3 describes the RAVDESS dataset and the methodology. Section 4 describes the experiments in detail, then, Section 5 gathers the main results obtained. Finally, in Section 6, we discuss the most relevant conclusions of this study and future investigation lines.

2. Related Work

Emotions represent psychological conditions. Due to their interest and multiple applications, several disciplines analyze them from different viewpoints such as medicine, psychology, and computer science [14–16].

Among all the suggested psychological hypotheses, the literature indicates that two primary theories appear as models for annotating most of the current emotion recognition datasets [17]: the discrete emotion theory and the core affect/constructionist theory [18].

On the one hand, the discrete emotion theory argues that there are discrete (or basic) emotions that build the human emotional experience [18]. Ekman's theory [19] is in this group; he proposes classifying emotions into six big families: anger, fear, sadness, enjoyment, disgust, and surprise. Even though each of these emotions has other sub-emotions detailed, most of the datasets only are labelled in terms of the first level of six basic emotions, mainly because of its simplicity since it maps emotions into a discrete

space. Some video-based corpus such as RAVDESS [20] or Emo-DB [21] are annotated with discrete emotions.

On the other hand, the core affect/constructionist theory frames emotions on an n-dimensional scale. Posner and Russell proposed: ‘the circumplex model of affect’. In this model, they suggest that emotions arise as a product of two separate neural systems, one carrying information about valence (i.e., how pleasant or unpleasant emotions are) and the other representing arousal (i.e., how intense or soft an emotion is) [22]. Databases like RECOLA [23], AffectNet [24], or IEMOCAP [25] contain this type of annotation.

2.1. Speech Emotion Recognition

When people engage in spontaneous conversational exchanges, their speech reveals their emotional state and personality traits in addition to the meaning of the words and their conveyance [10,26]. Paralanguage refers to the characteristics of the voice signal that can be utilized to change the meaning of phrases and transmit emotion, either intentionally or subconsciously. Schuller and Batliner [11] offered a comprehensive overview of computational paralinguistics, addressing the primary methods, tools, and techniques for recognizing affect, emotion, and personality in human speech. Berkeham and Oguz recently published an extensive review of speech emotion detection models, databases, features, preprocessing approaches, supporting modalities, and classifiers [27].

According to the reviews of Wani et al. [28] and Berkeham and Oguz [27], we can distinguish two main ways to perform speech emotion recognition: by using traditional classifiers or deep-learning classifiers.

For many years, the tendency was to apply feature engineering from low-level descriptors and feed with them a traditional classifier, such as SVM, logistic regression, or decision tree. These descriptors contained relevant information to categorize emotions. An example framed within this line of investigation is the work of Ancilin and Milton [29]. They studied a method to obtain Mel frequency cepstral coefficients calculating the magnitude spectrum, instead of the energy spectrum, without doing a discrete cosine transform, which generated their Mel Frequency Magnitude Coefficient.

However, feature engineering required large amounts of time to create and decide which descriptors were the most suitable for solving a specific task. As an attempt to reduce the researchers’ efforts, some frameworks emerged to obtain these hand-crafted features automatically, such as OpenSmile [30] or Praat [31]. Bhavan et al.’s [32] work was an example of this traditional line of investigation that employed descriptors. They passed MFCCs and spectral centroids to a bagged ensemble of support vector machines, giving an overall accuracy of 72.91% for RAVDESS.

Nowadays, most publications employ deep-learning models. These classifiers are usually neural networks capable of processing these descriptors or the complete audio records. For example, Singh et al. [33] suggested the use of prosody, spectral-information, and voice quality, to train a hierarchical DNN classifier, reaching an accuracy of 81.2% on RAVDESS. Pepino et al. [34] combined eGeMAPS features with the embeddings extracted from an xlsr-Wav2Vec2.0 to train a CNN model. They achieved an accuracy of 77.5% by applying a global normalization on this dataset. Issa et al. [35] also proposed a new method for feature extraction calculating Mel-frequency cepstral coefficients, chromagram, Mel-scale spectrogram, Tonnetz representation, and spectral contrast features from speech records. These features are the inputs of a one-dimensional CNN. Their proposal reached an accuracy of 71.61% for RAVDESS. Other works such as those proposed in [36–38] also employed CNNs, MLPs, or LSTMs to solve emotion recognition on RAVDESS using spectrograms or pre-processed features, obtaining accuracies of 80.00%, 96.18%, and 81%, respectively.

Although RAVDESS is appearing in a growing number of publications, there are not standard evaluation criteria yet, making it complex to quantify and compare contributions. For example, in [37] they achieved 96.18% accuracy using a 10-CV evaluation. Nonetheless, they did not specify how they distributed users in each fold, making it unclear whether

the same user participated in the training and test sets or not. Deciding whether the distribution of users is subject-wise or not is a relevant fact to consider when implementing an evaluation setup. Non-subject-wise scenarios will always result in higher performance rates because the training and the test sets have samples of the same user.

Continuing with the publications that exploit deep-learning models, some studies focus on the benefits of transfer learning to extract embeddings or fine-tune pre-trained models rather than extracting hand-crafted features [39–41]. DeepSpectrum [42], PANNs [43], and Hugging Face [44] are libraries that contain pre-trained models on audio, images, and/or text. Hugging Face is the most extensive library, with a repository dedicated to working with transformers to solve problems based on aural, visual, or textual modalities. For this reason, we employed it in this work.

2.2. Facial Emotion Recognition

Although voice is a crucial indicator of a subject's emotion, other modalities could enhance the SERs' performance, as demonstrated by Singh et al. in [33], which incorporated textual features to supplement the speech emotion recognizer. In our scenario, we included the visual information of the facial expressions.

Several libraries allow for the detection of facial variations and facial morphology. The dlib library [45], which estimates position of the landmarks on a face, is an example of these tools. According to Nguyen et al. [46] and Poulou et al. [47], landmarks encapsulate meaningful information about a person's facial expression that helps to solve automatic emotion recognition.

As an evolution of the facial landmarks, we have the facial action coding system (FACS) [48]. This system consists of a set of descriptions that refer to the movement of the facial muscles and are coded in what Ekman et al. called Action Units (AUs). Due to their simplicity and power, they have been investigated and utilized in several publications, such as in the work of Sanchez-Mendoza et al. [49], where they employed 12 AUs to classify emotions in the Cohn–Kanade (CK) database using a decision-tree; obtaining a recognition rate of 90%.

In the work of Li et al. [50], they used a first classifier to predict the presence of 14 AUs. Then, they fed the output of this classifier into an SVM to recognize emotions, reaching an average recognition rate of 94.07% for females and 90.77% for males on the CK database. In [51], Senechal et al. compared two systems: the first one consisted of training a classifier for detecting the presence of AUs, followed by an emotion recognizer; and the second one involved using only an emotion recognizer classifier. Their results confirmed that the system based on AUs reached higher scores than the version without including the AUs on several datasets, when they both receive facial images coded by local gaborbinary pattern (LGBP) histogram differences. These results reinforced the reliability of AUs for performing emotion recognition.

In the contribution of Bagheri et al. [52], they extracted 15 pairs of AUs' descriptors with the OpenFace library [53]. These descriptors indicated the presence of the AU and its associated intensity. After generating these features, they trained a stacked auto-encoder to compact the information and create an intermediate and more compact representation from the mentioned AUs. Then, they appended a fully connected layer to classify seven emotions. They reported a 96.5% classification rate at frame level on the RAVDESS dataset. However, they did not specify certain hyper-parameters and training settings necessary to replicate their experiments and compare proposals, like the learning rate or the cost function, among others.

Aside from using hand-crafted or automatically generated features for emotion recognition, several deep models that work directly with facial images can also be found in the literature. One of these models is EmotionalDAN [54], a CNN-based model that addresses emotion, valence, and landmark recognition all at once. Deep-emotion model [55] is another example of this. Deep-emotion model uses an STN architecture [56] with an attention

mechanism to address emotion recognition. Finally, the work of H. Kim et al. [57] employs an Xception model to perform facial emotion recognition.

Since our target was to estimate emotion in videos rather than in frames (or images), we evaluated two strategies to give a final prediction on the whole video:

The first proposal collapsed the sequence of AUs generated in each timestep into a single vector that was the average of all the temporal steps. These new features fed three distinct static models: an SVM, a k-NN, and an MLP.

The second proposal employed a sequential model, i.e., a bi-LSTM with an attention mechanism to extract the video verdict from the sequence of AUs retrieved from each frame of the video.

2.3. Multimodal Emotion Recognition

According to the review of Huang et al. [58], there are three basic ways for merging modalities: early fusion, joint fusion, and late fusion.

Early fusion consists of combining features or modalities extracted from various pre-trained models. Before training a final model, these attributes are grouped into a single vector.

Huang et al. [58] define joint fusion as “the process of joining learned feature representations from intermediate layers of neural networks with features from other modalities as input to a final model. The key difference, compared to early fusion, is that the loss is propagated back to the feature extracting neural networks during training, thus creating better feature representations for each training iteration”.

Late fusion, on the other hand, consists of two stages: a first stage in which as many models as modalities are trained and a second stage in which a final model receives the joined posteriors derived in the first stage to perform the definitive classification. The line between these procedures can be blurry at times since fusion tactics can occur at any time during the training [59].

Early fusion has the advantage of detecting feature correlations to eliminate redundant information and learn the interaction between distinct modalities. However, because of the varying sampling rates, it may have synchronization issues when aligning data from many modalities, as well as difficulties when the combined embeddings are high dimensional [60,61]. This method includes some works, such as the one proposed by Deng et al. [62]. They collected representative features, from the T5 transformer textual model and VGG, YAMNET, and TRILL aural models; then, these embeddings were concatenated and introduced into a co-attention transformer model, which enhanced the most relevant slots of each embedding to produce a fused representation, which was then used to train a final classifier. The fusion of these two modalities boosted the accuracy of the emotion recognizer in two datasets, IEMOCAP and SAVEEE.

As an alternative to an early fusion strategy, there exists a fusion at the decision level or late fusion. Sun et al. [60] employed features taken from pre-trained models on previous tasks to train a bi-LSTM model with an attention layer for each of their three used modalities (audio, video, and text) to recognize arousal and valence. Then, the posteriors of the bi-LSTM models were integrated by employing a late fusion technique to learn a final LSTM model.

Due to the simplifications and adequate performance of the late fusion strategy on similar tasks [60,63], we decided to apply a combination of the posteriors of each trained model per modality (aural or visual). Later, we fed a multinomial logistic regression with the generated outputs. This process could also be understood as an ensemble approach: we assembled the posteriors learned by each model on their own, and then we trained a multinomial logistic regression model for solving a single task, emotion recognition.

3. Methodology

Our framework is composed of two systems: the speech emotion recognizer and the facial emotion recognizer. We combined the results of these two systems with a late fusion strategy, as we can see in the diagram of Figure 1.

In this section, we will present the used dataset and go into the details of each system and the strategies applied.

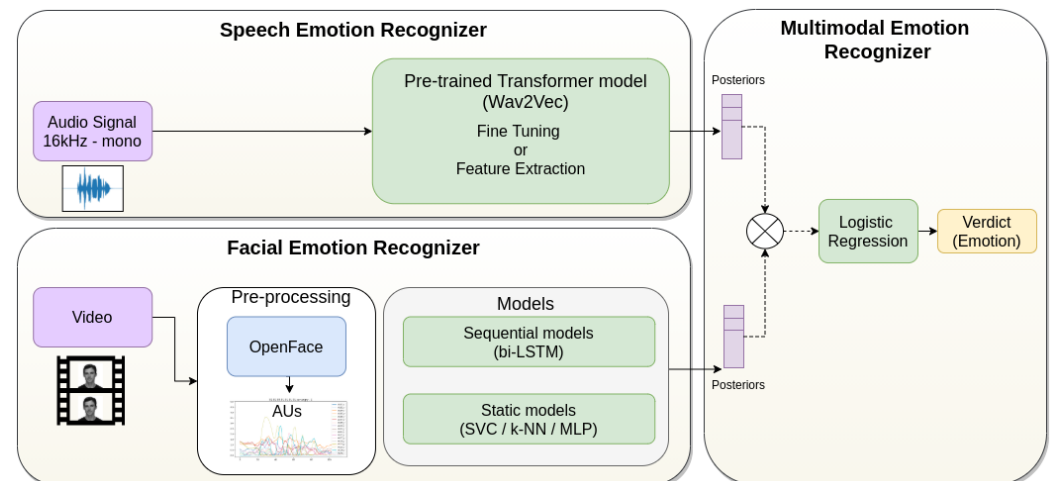


Figure 1. Block diagram of the implemented systems. The figure represents the analyzed models from the existent family of sequential models, static models, and transformers.

3.1. The Dataset and Evaluation

In our analysis, we have used RAVDESS [20]. This dataset includes 7356 recordings with acted-emotional content. The archives are distributed equally into three types of content (full AV, video-only, and audio-only) and two vocal channels (speech and song).

Except for the neutral emotion, which includes only regular intensity, the rest of the expressions are produced at two levels of emotional arousal: regular and strong. Each file contains a single actor representing one of the eight following emotions: calm, neutral, happy, sad, angry, fearful, surprised, and disgusted.

We only used the full AV material and the speech channel for our experiments because we were interested in audio-visual emotion recognition on speech rather than songs. This selection limited the number of files to 1440 videos with a maximum and minimum duration of 5.31 and 2.99 s, respectively. The corpus has 24 actors distributed in a gender-balanced way, who speak lexically-matched statements in a neutral North American accent. This setup is suitable to study the para-linguistics associated with emotions, isolating the lexical and reducing the bias in emotional expressions that culture may induce. Among its advantages, it also has a proportional number of files per emotion, which avoids problems derived from training algorithms with non-balanced data. Additionally, RAVDESS is a reference dataset in the research community, employed in several works [33,64,65].

Despite its simplifications, this dataset poses significant hurdles for emotion identification, even for humans. The human accuracy rate achieved utilizing only speech stimuli was of 67%, whereas when using visual information, this rate just increased until the 75%.

To evaluate and compare our results, we used a subject-wise 5-fold cross-validation strategy. The folds were randomly and stratified divided per classes and users, i.e., each fold had a similar number of samples per class randomly selected, but we always kept each actor in either the train or validation sets, never in both.

The distribution per actor for the validation folds was as follows:

- Fold 0: (2, 5, 14, 15, 16);
- Fold 1: (3, 6, 7, 13, 18);
- Fold 2: (10, 11, 12, 19, 20);
- Fold 3: (8, 17, 21, 23, 24);
- Fold 4: (1, 4, 9, 22).

We proposed this setup following the work of Issa et al. [35], who applied a similar subject-wise cross-validation methodology using the eight classes of the dataset. This evaluation procedure allowed us to compare our contribution to this previous work and with our prior solutions in [13].

Regarding the metrics, we compared our implementations with the average accuracy achieved by the cross-validation strategy at the video level. We also included a confidence interval to compare scenarios and evaluate the significance of our methods. Additionally, we calculated precision and recall per emotion for the best model.

3.2. Speech Emotion Recognizer

Training a deep neural network from scratch for emotion recognition requires a large amount of data to learn how to decide between several classes. Transfer learning techniques can alleviate this load by customizing pre-trained models. For this reason, we compared two different transfer-learning solutions: feature extraction and fine-tuning.

In this section, we will describe the implementation of these two techniques in the context of speech emotion recognition.

3.2.1. Feature Extraction

For the SER model, we used a pre-trained xlsr-Wav2Vec2.0 [66] model. This model had the original architecture of a Wav2Vec2.0 transformer [67]. Unlike Wav2Vec2.0, the xlsr version was trained in 53 distinct languages, reaching state-of-the-art performance in speech-to-text. Further, xlsr-Wav2Vec2.0 was a transformer trained in a self-supervised way from millions of raw audio data. After its pre-training on unlabelled data, the model was fine-tuned on labeled data to adapt it to downstream speech recognition tasks of different nature.

As Baevski et al. described in [67], this model consists of three distinct parts (which also appear in Figure 2): the feature encoder, the transformer, and the quantization module.

Firstly, the feature encoder contains several convolutional layers that receive the raw audio X and outputs a latent speech representation Z of each timestep of the recording.

Secondly, the transformer module receives the latent speech representations Z and creates the context representations C . This context representation is generated after passing through the 24 transformer blocks and 16 attention heads that conform to the transformer module.

Finally, the quantization module maps the output of the feature encoder Z into a discrete space using a product quantization that generates the compact vector Q .

For our experiments, we employed the LARGE version of xlsr-Wav2Vec2.0, fine-tuned on the English set of the Common Voice dataset [68] and available in the following Hugging Face repository: <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english> (accessed on 26 December 2021).

For the feature extraction strategy, we re-used the pre-trained network to obtain a latent speech representation of each recording. First, we subsampled the audios to 16 kHz and converted them into mono-channels using the FFmpeg tool [69]. Internally, the framework divided the audio into windows of 25 ms with an overlap of 15 ms and a stride of 20 ms. When the transformer generated the sequences of 512-dimensional embeddings from the convolutional feature encoder, we calculated the average of these embeddings along its temporal dimension. With the averaged 512-dimensional representation of each recording, we trained several static speech emotion recognizers employing the sklearn library [70]. Among the compared models, we used a support vector machine (SVM) with an 'RBF'

kernel, a k-Nearest Neighbours (kNN) with a majority voting to select the class, and a multilayer perceptron (MLP) with one or two layers of 80 neurons each and an output eight-neurons layer.

In this way, we re-used the original features of the speech-to-text task to solve speech emotion recognition, transferring the learned knowledge contained on the embeddings to the new models.

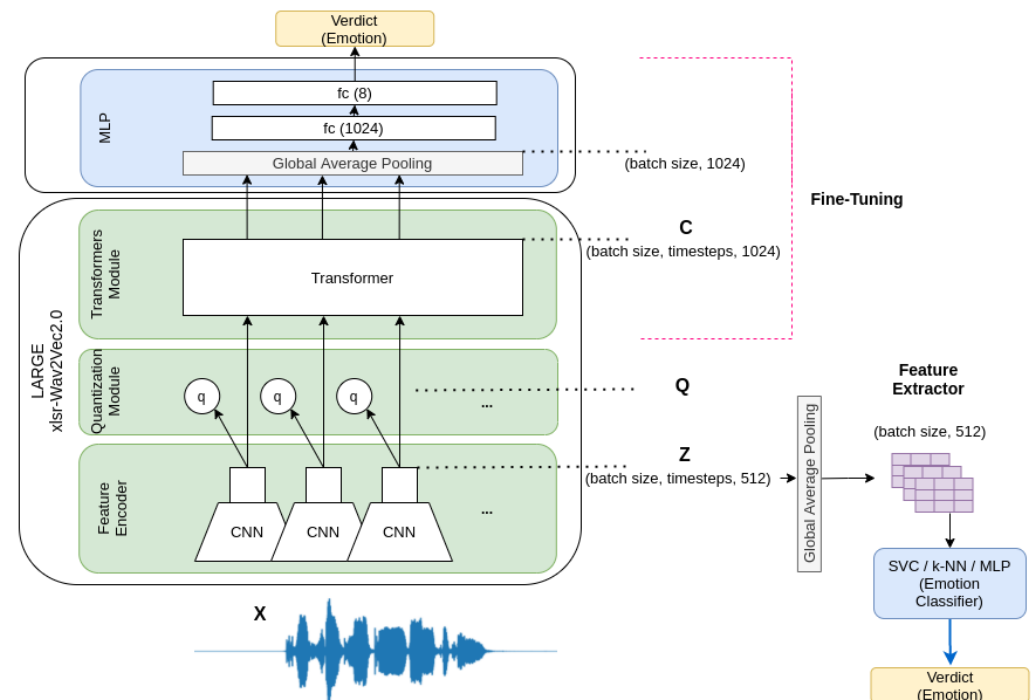


Figure 2. Proposed pipelines for speech emotion recognition.

3.2.2. Fine-Tuning

As an alternative to the embeddings extraction and to re-use the previous networks' expertise, we also fine-tuned the pre-trained xlsr-Wav2Vec2.0. By fine-tuning a base pre-trained model, we unfroze a few of its top layers and jointly trained both the newly-added classifier layers and the last layers of the base model. This technique allowed us to 'fine-tune' the higher-order feature representations in the base model, to adapt them to the new specific task, maintaining the knowledge acquired from the training on millions of data samples.

In our case, the new task to solve was speech emotion recognition. To adapt the xlsr-Wav2Vec2.0 architecture, we introduced a global average pooling on top of the output of the transformer module. This layer collapsed all the timesteps of the context representation C into a single 1024-dimensional vector. These averaged embeddings were passed to an MLP of two layers with 1024 and eight neurons, respectively, stacked on top of the pooling layer.

During the fine-tuning process, all the layers were adjusted except for the convolutional layers of the feature encoder. The layers of this module stayed frozen because they contained embedded knowledge from a large amount of data and were robust enough for being used without adaptation.

In Figure 2, we show a diagram that clarifies the feature extractor and the fine-tuning strategies. Under the square of 'LARGE xlsr-Wav2Vec2.0', we can see in green the default layers that the model had. From the output of the feature encoder stage, we extracted the embeddings that we fed to the static models (SVM, k-NN, and MLP). Regarding the fine-tuning version, the pink lines indicate the layers we re-trained with RAVDESS. Inside the blue box, we can also identify the added layers on top of the transformer for performing emotion classification.

3.3. Facial Emotion Recognizer

To address facial emotion recognition, we used Action Units (AUs) as inputs. Action Units are the basic units of the facial action coding system (FACS), which taxonomizes human facial movements by their appearance on the face. Hence, each Action Unit encodes a specific facial movement, usually associated with a modification of a facial expression due to the changes in people's psychological state.

To solve video emotion recognition from the Action Units, we compared the performance of static against sequential models. In this section, we will illustrate the characteristics of each method.

3.3.1. AUs Extraction

Among all the available resources to extract Action Units from the frames of the videos, we used the OpenFace toolkit [53]. This API delivers the presence and intensity of 18 different Action Units (see Appendix A). The presence of an AU is binary-encoded as 1 if the AU is displayed, or 0 otherwise. In contrast, the intensity is a continuous variable ranging from 0 to 5. Both predictions are derived from two independent networks that followed the same pre-processing stage; firstly, faces were aligned to compute geometrical and appearance-based features; second, the features fed two disconnected SVM models that return existence of the AUs and its arousal in each frame, respectively [71]. By default, the features at the video level were normalized under the assumption that many frames represented a neutral emotion. In our work, we also applied this default configuration.

In Figure 3, we present two instances of the intensity-AUs generated by the tool for two different videos. The first actor displayed an angry emotion, and the second showed a happy expression. In the angry sample, the AU4 ('Brow Lowerer' FACS' name) and AU7 ('Lid Tightener' FACS' name) reach high-intensity values during most of the video because these AUs are associated with the eye movements that reflect anger. For the happy sample, the AU12 ('Lip Corner Puller' FACS' name) shows the maximum values for many frames since it is associated with a 'smile mouth movement' that correlates with an expression of happiness. Unlike the sample of anger, the AU4 and AU7 report lower intensities in the plot of happiness.

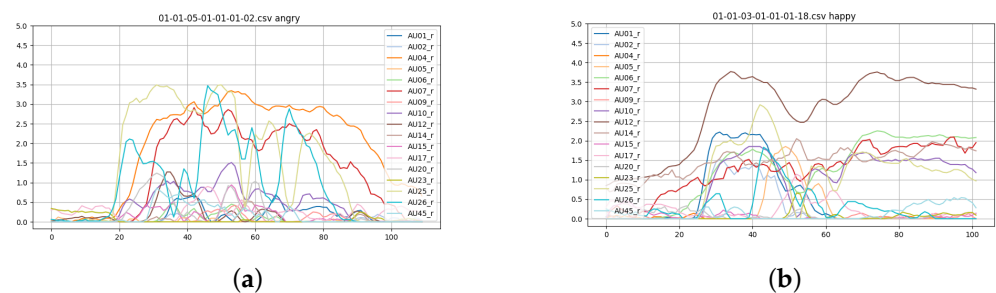


Figure 3. Plots of the 18 AUs generated by the OpenFace library [71] for the intensity version ranging from 0 to 5 of two videos of the RAVDESS dataset. (a) AUs' intensities for an angry sample; (b) AUs' intensities for a happy sample.

3.3.2. Static vs. Sequential Models

Once we obtained the AUs of each video frame, we needed to accomplish a pre-processing step to adapt the input format to the specific model, whether static or sequential.

To evaluate the static models, we computed the average vector of the sequence of AUs extracted from each video, collapsing all the temporal steps into a single softened vector.

In this stage, we tested two strategies, applying a normalization of the AUs in the range of 0–1 for each column, or not employing any normalization.

After adapting the AUs to the static problem, we introduced the samples into different models: an SVC, a k-NN classifier, and an MLP.

This approach, adopted as our baseline, had two main benefits: the first one was its simplicity, and the second one was ‘the average effect’. To illustrate this, suppose that a video has a deviation on several frames from a prototype emotion because the person on it has closed their eyes. Thanks to the average pooling, these frames would not severely influence the final recognition, as long as all the other frames encapsulated the correct emotion.

Nevertheless, this method has also an evident drawback: sequential data may exhibit a natural temporal ordering. However, this heuristic aggregation of individual frame-level features ignores temporal order, which in turn may result in a sub-optimal discriminative capability.

As an alternative to the static models, we have adopted a sequential model, which is usually effective, especially for sequential data, assuming that there is relevant information in the order of the frames.

As a sequential model, we employed an RNN. This RNN used a long short-term memory (LSTM) network. This model can process its inputs sequentially, performing the same operation, $h_t = f_W(\overrightarrow{AU}_t, h_{t-1})$, on each of the distinct timesteps that conform to our input series, where h_t is the hidden state, t the time step, and W the weights of the network.

As input to this model, we introduced the sequence of AUs produced per frame of the video ($\overrightarrow{AU}_1, \overrightarrow{AU}_2, \dots, \overrightarrow{AU}_N$), to exploit the temporal patterns enclosed in the AUs and make a final prediction at the video level.

Regarding the architecture of the sequential model, it consisted of several bidirectional-LSTM layers with a deep self-attention mechanism, similar to the proposed in [72]. In Figure 4, we show a picture of the structure of the employed Bi-LSTM.

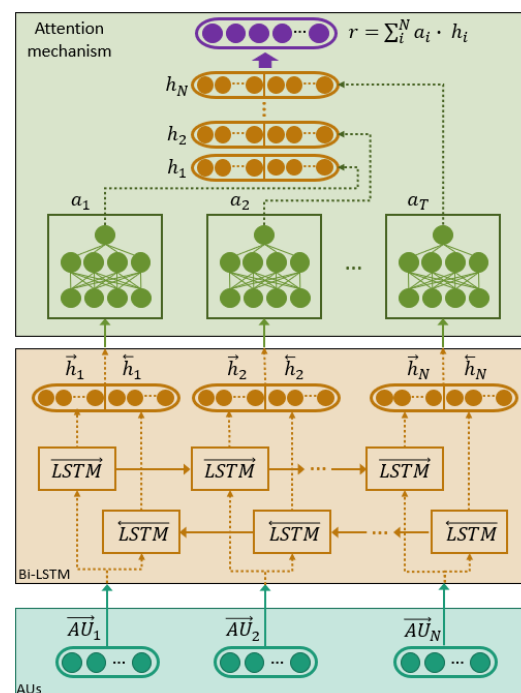


Figure 4. Bidirectional-LSTM with attention mechanism for emotion recognition using sequences of Action Units at the input. Modified version from source [73].

The Bi-LSTM layer works in a bidirectional way, which allowed us to collect the sequential information in both directions of the hidden states h_1, h_2, \dots, h_N of the LSTMs.

In particular, our Bi-LSTM consisted of two LSTMs: *forward* \overrightarrow{LSTM} , which permitted the analysis of the frames from \overrightarrow{AU}_1 to \overrightarrow{AU}_N , and an *inverse or backward* \overleftarrow{LSTM} , which permitted a similar analysis to be carried out but in the opposite direction, from \overrightarrow{AU}_N to \overrightarrow{AU}_1 .

To obtain the emotional tag from the Bi-LSTM layers, we concatenated the embeddings of the outputs of each specific direction (see Equation (1) in which $||$ corresponds to the concatenation operator and L to the size of each LSTM).

$$h_i = \overrightarrow{h_i} || \overleftarrow{h_i}, \text{ where } h_i \in R^{2L} \quad (1)$$

The main task of the attention layer was to distinguish the most relevant AUs associated with a specific frame when determining the emotion portrayed in the whole video. The actual contribution of each embedding was estimated through a multilayer perceptron (MLP) with a non-linear activation function (tanh), as the proposed in [74].

The attention function g was a probability distribution applied on the hidden states h_i that allowed us to obtain the attention weights a_i that each embedding (or frame of the video) receives. At the output of the attention layer, the model calculated the linear combination of the LSTM outputs h_i with the weights a_i .

Finally, r was used as a feature vector that fed a final task-specific layer for emotion recognition. In particular, we used a fully connected layer of eight neurons, followed by a softmax activation, which returned the probability distribution over the classes.

3.4. Multimodal Recognizer

We applied a late fusion strategy to combine the learned knowledge from both modalities. To begin with, we extracted the posteriors from the last fully-connected layer of the models, which in our case consisted of an eight-dimensional vector per model. Next, we concatenated these embeddings from each modality. At the end, we trained a multinomial logistic regression with the sklearn library's default parameters.

As we did in the static models, we also applied a normalization in the range of 0–1 on the joined posteriors. This normalization gave better results compared to the version without it. Likely this is because the normalization equalizes the contribution of each model.

4. Experiments

In this section, we will tackle the training parameters (batch size, epochs, optimizers, learning rate...). Furthermore, we will explain the main differences of the tested architectures on each modality.

We will follow the same structure that we established in the rest of the paper. First, we will examine the settings for the two TL strategies applied on the speech emotion recognizer: feature extraction and fine-tuning. Then, we will move to the facial emotion recognizer and describe the specifications of the static and sequential models.

4.1. Speech Emotion Recognizer Setup

As we commented in Section 3.2, for the feature extraction, we extracted 512-dimensional embeddings from the last layer of the feature encoder of the xlsr-Wav2Vec2.0 network. After that, we calculated the average of these embeddings and compared the performance of three algorithms: SVC with 'RBF' kernel, k-NN, and MLP. For the SVM, we varied the regularization parameter between 1, 10, and 100. For the k-NN, we modified the number of neighbors to 10, 20, 30, and 40. Finally, for the MLP we tested architectures with one or two layers, always with 80 neurons each, except for the last output layer that always had eight neurons. The rest of the parameters took the default values set in the sklearn library.

Concerning the training configuration and hyper-parameters for the Fine-Tuning experiments, we selected a batch size of 100 samples, adequate to the capacity of our GPU; and a maximum number of training epochs of 10, since higher values showed an increment in the overfitting of the model and did not improve the validation metrics.

As we were solving a classification task, we utilized the cross-entropy loss implemented in the Hugging Face library [44]. To optimize this objective function, we employed the default optimizer of the library, AdamW, with a learning rate of 0.001, betas with values

0.9 and 0.999, and epsilon of 1×10^{-6} , fine-tuning the previous weights learned from the English set of CommonVoice dataset.

Notice that the xlsr-Wav2Vec2.0 model was adapted to include an MLP on top of the transformer module and an average pooling layer. The average pooling compacted the timesteps at the output for having a single vector of 1024 dimensions per recording. This vector passed to an MLP with a hidden layer of 1024 neurons, activated with a tanh function and an output layer of eight neurons that returned the final probabilities of each class thanks to its softmax activation.

4.1.1. Facial Emotion Recognizer Setup

Continuing with the visual experiments, we compared the performance reached using the AUs extracted from OpenFace on the static and sequential models.

Regarding the static models, we also tested an SVC, k-NN, and MLP from the sklearn library, changing their hyper-parameters, similar to the ones used for the feature extraction version on the speech emotion recognizer.

For the sequential model, we created a bi-LSTM of two layers with 50 neurons. The bi-LSTM also contained an attention mechanism with two layers. The whole model was implemented in Pytorch [75]. We trained the model for 300 epochs, as maximum, in batches of 64 samples. To avoid overfitting, we also implemented an early-stopping strategy to finish the training when the F1 score of the validation set did not improve in 30 epochs.

4.1.2. Multimodal Emotion Recognizer Setup

For the late fusion models, we employed a multinomial logistic regression algorithm implemented using the sklearn library [70]. We set the regularization parameter (C) to 1. The rest of the parameters maintained their default values.

5. Results

Through this section, we will present the main outcomes obtained from the experiments on speech and facial modalities. We will also include an error analysis and a comparison of our systems with previous publications on the same dataset.

5.1. Speech Emotion Recognition Results

The results of Table 1 summarize the performance of the speech emotion recognition models tested. For the feature extraction strategy using the 512-dimensional representation obtained from the feature encoder of the xlsr-Wav2Vec2.0, we reached the maximum accuracy when we passed them to the MLP of one layer. This model outperformed all the k-NNs tested in a statistically significant manner. Additionally, the MLP reached a higher average accuracy than the SVCs.

Although the embeddings extracted from the feature encoder improved the accuracy of the ZeroR, demonstrating there was relevant information embedded in these representations, the Fine-Tuning strategy beat all the feature extraction-based methods. This fact demonstrated the plasticity of the xlsr-Wav2Vec2.0 transformer to adapt its weights with new information. The fine-tuning version surpassed the top MLP in 25.29% points. This rise also suggests that our dataset has enough size to let the networks learn effectively from the recordings. Furthermore, the pre-trained weights contained consistent and compatible knowledge to solve the speech emotion recognition task.

To understand the errors of the top solution, we extracted the confusion matrix from the predictions of the fine-tuned xlsr-Wav2Vec2.0 that reached an accuracy of 81.82%. The confusion matrix displayed in Figure 5 is the rounded average value of the errors and the predictions obtained in the 5-CV. For this reason, this matrix has 288 samples in total (1440/5).

Table 1. Quantitative evaluation of the different strategies on speech emotion recognition. In bold, the best models per TL strategy.

TL Strategy	Inputs	Models	Hyper-Parameters	Accuracy \pm 95% CI
-	-	Human perception	-	67.00
-	-	ZeroR	-	13.33 \pm 1.76
Feature Extraction (Static)	Average xlsr-Wav2Vec2.0 embs. from feature encoder	SVC	C = 1.0	50.13 \pm 2.58
			C = 10.0	53.12 \pm 2.58
			C = 100.0	53.10 \pm 2.58
		kNN	k = 10	36.07 \pm 2.48
			k = 20	37.90 \pm 2.51
			k = 30	38.65 \pm 2.51
			k = 40	38.47 \pm 2.51
		MLP	1 layer (80)	56.53 \pm 2.56
			2 layers (80,80)	55.82 \pm 2.56
Fine Tuning (Sequential)	Raw audio	xlsr-Wav2Vec2.0 + MLP	MLP 2 layers (1024,8)	81.82 \pm 1.99

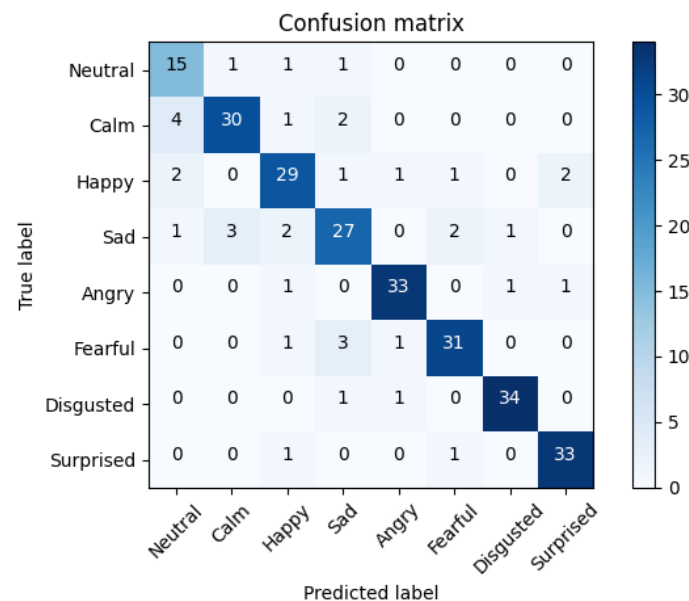
**Figure 5.** Average confusion matrix of the fine-tuned xlsr-Wav2Vec2.0 experiment with an accuracy of 81.82%.

Figure 5 reveals that the model exhibited good performance for most samples except for some cases. For example, the ‘Neutral’ and ‘Calm’ classes were sometimes confused due to the similarities between these two emotions. Likewise, the ‘Sad’ class was occasionally mistaken with ‘Fearful’, ‘Calm’, or ‘Neutral’. The errors between ‘Sad’ and ‘Fearful’ were understandable because of the low arousal present in both emotions. On the other hand, the mistakes between the predictions of ‘Sad’, ‘Neutral’, or ‘Calm’ could come from the neutral voice chunks at the beginning of some recordings that may have created the false perception that they belonged to ‘Calm’ or ‘Neutral’ classes.

As a conclusion from the results of the speech modality, we could confirm that fine-tuning the pre-trained models on similar tasks helped to reach higher scores since the dataset had enough samples to train the model. Additionally, using deeper and pre-trained models that work with sequences rather than functionals of features made a significant difference in the accuracy that they achieved.

5.2. Facial Emotion Recognition Results

In Table 2, we can see the results for the facial emotion recognizer. The static models (SVC, k-NN, and MLP), trained with the average of the AUs, obtained comparable performances to the sequential model despite the lower computer time consumption and the reduced complexity of these models. Among the static models, the MLP with a single layer of 80 neurons achieved the best accuracy, 58.93%. This metric improved when we normalized the input features between 0 and 1, to 60.22%. These results seem to indicate that the normalization emphasized the differences in the representation of the AUs for each emotion and gave more relevance to AU's binary attributes of presence against the intensity attributes.

As an alternative to the static models, we also considered using the AUs extracted from each frame of the videos as individual instances, inheriting the label of their parent video. However, training the models with all the samples and applying a max. voting algorithm to predict the video level class, reported inferior values. For this reason, they are not present in Table 2, but the reader can consult these results in the Appendix B.

Table 2. Quantitative evaluation of the different strategies applied for the facial emotion recognizer. In bold, the best models per input type.

Inputs	Models	Hyper-Parameters	Norm.	Accuracy \pm 95% CI
-	Human perception	-	-	75.00
-	ZeroR	-	-	13.33 \pm 1.76
Average Action Units	SVC	C = 0.1	Yes	53.25 \pm 2.58
			No	48.07 \pm 2.58
		C = 1.0	Yes	59.93 \pm 2.53
			No	54.88 \pm 2.57
		C = 10.0	Yes	55.93 \pm 2.56
			No	51.65 \pm 2.58
	kNN	k = 10	Yes	53.10 \pm 2.58
			No	46.80 \pm 2.58
		k = 20	Yes	54.30 \pm 2.57
			No	49.07 \pm 2.58
		k = 30	Yes	55.18 \pm 2.57
			No	48.82 \pm 2.58
		k = 40	Yes	55.40 \pm 2.57
			No	50.20 \pm 2.58
	MLP	1 layer (80)	Yes	60.22 \pm 2.53
			No	58.93 \pm 2.54
		2 layers (80,80)	Yes	57.77 \pm 2.55
			No	55.82 \pm 2.56
Sequence of Action Units	bi-LSTM	2 bi-LSTM layers (50,50) + 2 attention layers	No	62.13 \pm 2.51

Concerning the sequential model, we noticed an increment of 3.2 points regarding the non-normalized top MLP model. This result suggests that there was also relevant information in the temporal structure of the data. Even though this rate was higher than the obtained with the static models, it also revealed that there was still room for improvement on this modality since it still did not surpass human perception accuracy. Moreover, it opened an attractive research line to understand which frames are the most relevant to decide when a video belongs to one or another emotion. This understanding could support generating more precise algorithms for performing emotion recognition on videos.

To analyze the causes of the errors, we plotted the confusion matrix of the bi-LSTM experiment that reached an accuracy of 62.13%. We display this matrix in Figure 6.

The matrix illustrates that most errors happened between ‘Sad’ and ‘Fearful’, or between ‘Disgusted’ and ‘Sad’. These results may indicate that the AUs used as features were not enough to separate these classes due to the similar facial expressions that these emotions have in common. Another possible explanation is that these expressions do not follow a temporal pattern that our bi-LSTM can model. Since emotions vary over time, some frames could be more informative than others, and our sequential model did not attend them correctly, so far.

One possible solution to this problem could be the creation of a detector of ‘qualified’ frames. For instance, this system could rely on a threshold over the posteriors of an emotion recognition model trained only with images. This model could act as a filter of the most relevant samples on a video. This pre-processing stage ideally would detect the most informative frames that, later, would be used to fine-tune the video-based model. An alternative may have been to train the model with the whole clip, in an end-to-end way, using features of a different nature together with the AUs.

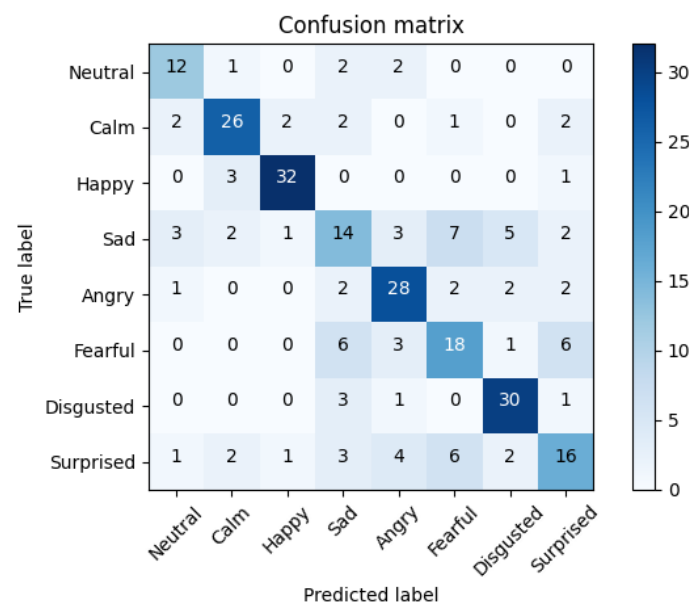


Figure 6. Average confusion matrix of the bi-LSTM with 2 layers of 50 neurons and 2 attention layers trained with the AUs. Accuracy of 62.13%. See Table 2.

5.3. Multimodal Fusion Results

Even though the results achieved by the visual modality were inferior to those reached by the speech-based system, the late fusion of the posteriors of these models improved both aural and visual modalities. Combining the aural and visual emotion recognizers allowed one to achieve an overall accuracy of 86.70%, against the 62.13% of the visual modality and the 81.82% of the top speech-based strategy.

This accuracy was obtained by a multinomial logistic regression when we combined and normalized the posteriors of the FT-xlsr-Wav2Vec2.0 model for SER (81.82% of ac-

curacy), the bi-LSTM with attention mechanism for FER (62.13% of accuracy), and the Static-MLP of 80 neurons for FER fed with the averaged and normalized AUs (60.22% of accuracy).

In Figure 7, we compared the static models (SVC, k-NN, and MLP) with the sequential ones (Wav2Vec and bi-LSTM). We have plotted the performance of the top recognizers for each modality (aural and visual), besides the results of combining the static and sequential models. The overall fusion was represented in cyan using the top three models, which had an accuracy of 86.70%. In Table 3, we added the average precision, recall and accuracy of the 5-CV to compare the performance of this top algorithm for predicting each type of emotion.

To study when this fusion strategy failed, we pictured the confusion matrix in Figure 8. The diagonal of the matrix shows a higher amount of correctly predicted samples regarding the aural modality. More specifically, the ‘Happy’ class improved its accuracy, which seems reasonable since the visual modality distinguishes this emotion with high certainty. For other categories such as ‘Angry’, ‘Fearful’, ‘Disgusted’, or ‘Surprise’, the accuracy was similar to the one reached by the aural model. ‘Calm’ and ‘Sad’ also reflected a slight improvement after the fusion, while ‘Neutral’ maintained similar values as in the xlsr-Wav2Vec2.0 model after fine-tuning it. In summary, we can conclude that the fusion led to better results in almost all the emotions categorized in our study.

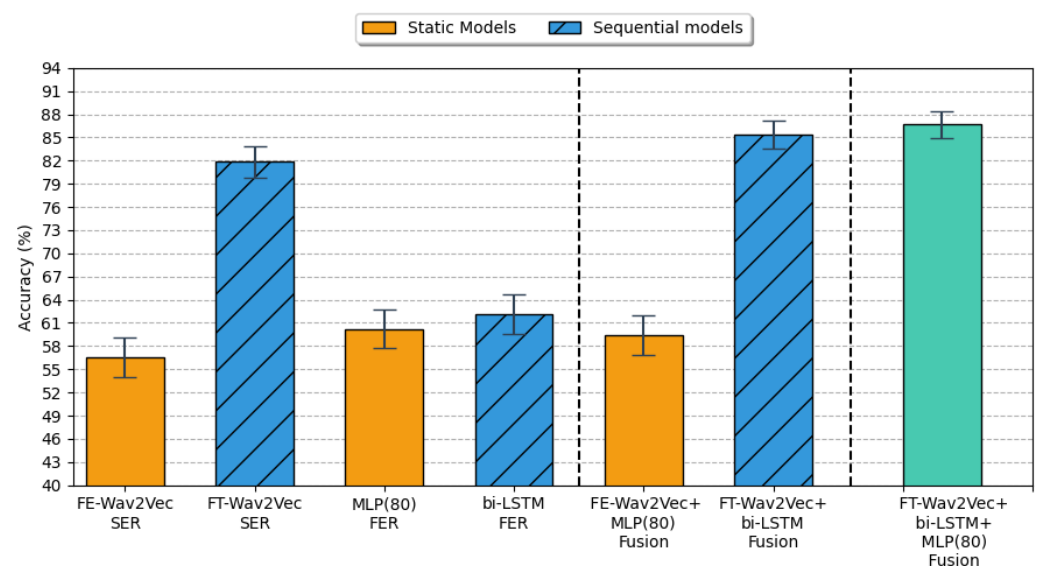


Figure 7. The top average accuracy of the 5-CV obtained for speech and visual modalities with a 95% confidence interval. In orange, the experiments of the top static models; in blue, the top models with sequential inputs; and in cyan, the top fusion model.

Table 3. Precision, recall, and accuracy metrics per emotion for the top model that achieves an accuracy of 86.70%. All the metrics are calculated as an average of the 5-CV strategy.

	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgusted	Surprised
Precision	88.31	91.45	89.79	71.22	91.78	88.26	93.99	89.50
Recall	82.25	85.63	89.25	81.88	92.37	83.62	88.50	87.87
Accuracy	98.05	96.88	97.05	92.73	97.83	96.20	97.70	96.95

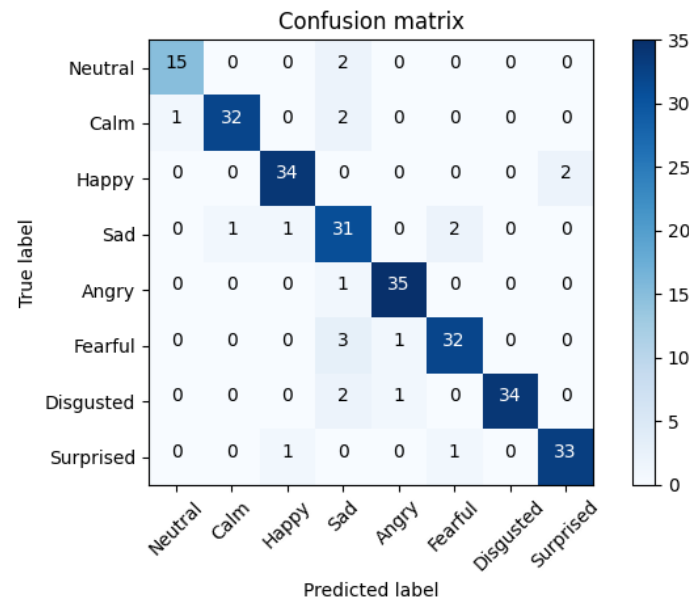


Figure 8. Average confusion matrix of the top late fusion model that combines the posteriors of the FT-xlsr-Wav2Vec2.0, the bi-LSTM, and the MLP of 80 neurons. Accuracy of 86.70%. See Figure 7.

5.4. Comparative Results with Previous Works

One of the main drawbacks of comparing the results on RAVDESS was the multiple set-ups used in the literature. The most reliable and closed evaluation methodology for comparing our experiments was the publication of Issa et al. [35]. Here, the researchers developed a speech emotion recognizer and applied a subject-wise 5-CV, obtaining an accuracy of 71.61%. Other works such as [76] used the last two participants in the validation and test sets, respectively, reaching an accuracy of 56.71% on the speech modality. With a variation of this set-up, we also found the work of Pepino et al. [34], which used as the test set only the last two participants and combined the ‘Calm’ and ‘Neutral’ emotions, passing from a problem with eight emotions to a problem with seven different classes. On these conditions, the top accuracy reached by their model is 77.5%, applying a global normalization.

With our audio-based model, we achieved an accuracy of 81.82%, improving by 10.21% the audio-based solution of Issa et al. [35], and by 15.09% if we compare this result to our best multimodal solution.

Regarding our previous publications, we can see that both methods, the feature extraction and the fine-tuning of the xlsr-Wav2Vec2.0, surpassed our previous proposals for the SER using CNNs in [13]. More specifically, for the feature extraction, we have obtained an improvement of 10.73 points, and for the fine-tuning, an increment of 5.24 with the current transformer-based approach. Regarding the visual modality, the AUs got a slight increment in comparison with the embeddings extracted from the STN on [13]. In our previous work, we reported an accuracy of 57.08%, and now we achieved 62.13%. As both modalities improved, the late fusion was also 6.62 points higher than in our previous study.

6. Conclusions

Automatic emotion classification is a difficult task. Although similar patterns seem to exist, there are still many differences between individuals, even when they are actors of the same nationality and speak the same language variety, as it happens in the RAVDESS corpus.

In this paper, we proposed a multimodal system for emotion recognition based on speech and facial data.

Concerning the speech-based models, we have demonstrated that the fine-tuned model using a pre-trained transformer outperformed the feature-extraction strategy by 25.29 points. When compared to human perception, our speech model achieved a 14.82 percent point increase, demonstrating the robustness of the proposed procedure for this modality. Furthermore, our proposal outperformed previously proposed solutions in [35] by 10.21 percent.

For the visual modality, the results showed that the sequential model achieved the highest accuracy. The results using the static and sequential models still fall short of the scores obtained with the SER and human capability. However, from this study, we have found some issues that will be researched further in the future in order to model the dynamic nature of emotions. An example of said issues was that we discovered that some frames in the video appeared to contain more important information than others, and neither implemented temporal nor static models were capable of capturing this knowledge from the Action Units.

Despite the lower performance of the visual modality regarding the speech modality, the fusion of both sources achieved an accuracy of 86.70% in automatic emotion classification, improving both single modalities.

In the future, we intend to improve the visual models by modifying the tested architectures or applying other transformer models. In addition, we will investigate how to extract the most relevant frames that contain a higher emotional load. If we succeed in this study, we expect to achieve closer performance of our models to human perception. Finally, we will test these strategies in real-world scenarios too.

Author Contributions: Conceptualization, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M.; Data curation, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M.; Formal analysis, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M.; Funding acquisition, D.G., Z.C., J.M.M. and F.F.-M.; Investigation, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M.; Methodology, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M.; Project administration, D.G., Z.C., J.M.M. and F.F.-M.; Resources, D.G., Z.C., J.M.M. and F.F.-M.; Software, C.L.-J. and R.K.; Supervision, D.G., Z.C., J.M.M. and F.F.-M.; Validation, C.L.-J., R.K. and F.F.-M.; Visualization, C.L.-J. and F.F.-M.; Writing—original draft, C.L.-J., R.K. and F.F.-M.; Writing—review editing, C.L.-J., R.K., D.G., Z.C., J.M.M. and F.F.-M. All authors have read and agreed to the published version of the manuscript.

Funding: The work leading to these results was supported by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), CAVIAR (TEC2017-84593-C2-1-R, funded by MCIN/AEI/10.13039/501100011033/FEDER “Una manera de hacer Europa”), and AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by “the European Union “NextGenerationEU/PRTR”). This research also received funding from the European Union’s Horizon2020 research and innovation program under grant agreement N° 823907 (<http://menhir-project.eu>, accessed on 17 November 2021). Furthermore, R.K.’s research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

Institutional Review Board Statement: Ethical review and approval were waived for this study because the datasets used are accessible under request for research purposes and the authors of this work adhered to the terms of the license agreement of the first dataset.

Informed Consent Statement: Subject consent was waived because the datasets used are accessible under request for research purposes and the authors of this study were adhered to the terms of the license agreement of the datasets.

Data Availability Statement: The RAVDESS database used in this paper is available under request from https://zenodo.org/record/1188976#.YTscC_wzY5k, accessed on 26 December 2021. All the code implemented for this publication will be available at <https://github.com/cristinalunaj/MMEmotionRecognition>, accessed on 26 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FER	facial emotion recognition
SER	speech emotion recognition
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song
Bi-LSTM	Bi-directional long short-term memory networks
GAN	generative adversarial networks
embs	embeddings
fc	fully-connected
SVC	support vector machines/classification
k-NN	k-Nearest Neighbours
MLP	multilayer perceptron
AU	Action Unit
FACS	facial action coding system
TL	transfer-learning
FT	fine-tuning
FE	feature extraction
CI	confidence interval

Appendix A. Generated AUs by the OpenFace Library

In Table A1, we show the 18 Action Units codes that can be extracted using the OpenFace toolkit [53], associated with the facial movement that they represent, following the FACS of Ekman in [48].

Table A1. Action Units extracted using the OpenFace library with its associated facial muscle movements of the facial action coding system.

AU Number	FACs Name
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
6	Check raiser
7	Lid tightener
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
20	Lip stretcher
23	Lip tightener
25	Lip part
26	Jaw drop
28	Lip suck
45	Blink

Appendix B. Training with as Many Samples as Frames + Max. Voting

Here, we report the results obtained from using the AUs extracted from each frame of the 1.440 videos that conform to the dataset. In total, we had 159,222 samples distributed per fold following the subject-wise 5-CV strategy that we described in Section 4. Each of the frames or samples inherited the label of the video they belong to. All results in Table A2 are given at video level, after applying a maximum voting algorithm on the posteriors of each generated output of the trained model on the test sets of each fold.

Table A2. Quantitative evaluation of the strategies applied for the FER using all the frames as training. Results reported at video level after applying max. voting.

Inuts	Models	Hyper-Parameters	Norm	Accuracy
Average Action Units	SVC	C = 0.1	Yes	52.23 ± 2.58
			No	54.18 ± 2.57
		C = 1	Yes	54.15 ± 2.57
			No	53.02 ± 2.58
		C = 10	Yes	53.20 ± 2.58
			No	52.48 ± 2.58
	k-NN	k = 10	Yes	49.78 ± 2.58
			No	49.20 ± 2.58
		k = 20	Yes	51.12 ± 2.58
			No	50.90 ± 2.58
		k = 30	Yes	52.57 ± 2.58
			No	51.67 ± 2.58
		k = 40	Yes	52.35 ± 2.58
			No	51.60 ± 2.58
	MLP	1 layer (80)	Yes	51.08 ± 2.58
			No	50.07 ± 2.58
		2 layers (80,80)	Yes	48.75 ± 2.58
			No	49.78 ± 2.58

References

- Kraus, M.; Wagner, N.; Callejas, Z.; Minker, W. The Role of Trust in Proactive Conversational Assistants. *IEEE Access* **2021**, *9*, 112821–112836. [\[CrossRef\]](#)
- Cassell, J.; Sullivan, J.; Prevost, S.; Churchill, E.F. *Embodied Conversational Agents*; The MIT Press: Cambridge, MA, USA, 2000.
- de Visser, E.J.; Pak, R.; Shaw, T.H. From ‘automation’ to ‘autonomy’: The importance of trust repair in human–machine interaction. *Ergonomics* **2018**, *61*, 1409–1427. [\[CrossRef\]](#)
- Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R.W. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Comput. Surv.* **2020**, *53*, 1–30. [\[CrossRef\]](#)
- Thakur, N.; Han, C.Y. An Ambient Intelligence-Based Human Behavior Monitoring Framework for Ubiquitous Environments. *Information* **2021**, *12*, 81. [\[CrossRef\]](#)
- Nyquist, A.C.; Luebke, A.M. An Emotion Recognition–Awareness Vulnerability Hypothesis for Depression in Adolescence: A Systematic Review. *Clin. Child Fam. Psychol. Rev.* **2019**, *23*, 27–53. [\[CrossRef\]](#) [\[PubMed\]](#)
- Greco, C.; Matarazzo, O.; Cordasco, G.; Vinciarelli, A.; Callejas, Z.; Esposito, A. Discriminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Systematic Review. *IEEE Access* **2021**, *9*, 112850–112870. [\[CrossRef\]](#)
- Argaud, S.; Vérin, M.; Sauleau, P.; Grandjean, D. Facial emotion recognition in Parkinson’s disease: A review and new hypotheses. *Mov. Disord.* **2018**, *33*, 554–567. [\[CrossRef\]](#)
- Franzoni, V.; Milani, A.; Nardi, D.; Vallverdú, J. Emotional machines: The next revolution. *Web Intell.* **2019**, *17*, 1–7. [\[CrossRef\]](#)
- McTear, M.; Callejas, Z.; Griol, D. *The Conversational Interface: Talking to Smart Devices*; Springer: Cham, Switzerland, 2016. [\[CrossRef\]](#)
- Schuller, B.; Batliner, A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st ed.; Wiley Publishing: Hoboken, NJ, USA, 2013.
- Anvarjon, T.; Mustaqeem; Kwon, S. Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. *Sensors* **2020**, *20*, 5212. [\[CrossRef\]](#)
- Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; Montero, J.M.; Fernández-Martínez, F. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors* **2021**, *21*, 7665. [\[CrossRef\]](#)
- Shah Fahad, M.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digital Signal Process.* **2021**, *110*, 102951. [\[CrossRef\]](#)
- Naga, P.; Marri, S.D.; Borreo, R. Facial emotion recognition methods, datasets and technologies: A literature survey. *Mater. Today Proc.* **2021**. [\[CrossRef\]](#)

16. Clavel, C.; Callejas, Z. Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. *IEEE Trans. Affect. Comput.* **2016**, *7*, 74–93. [\[CrossRef\]](#)
17. Ashraf, A.; Gunawan, T.; Rahman, F.; Kartiwi, M. A Summarization of Image and Video Databases for Emotion Recognition. In *Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering*; Springer: Singapore, 2022; Volume 730, pp. 669–680. [\[CrossRef\]](#)
18. Thanapattheerakul, T.; Mao, K.; Amoranto, J.; Chan, J. Emotion in a Century: A Review of Emotion Recognition. In Proceedings of the 10th International Conference on Advances in Information Technology (IAIT 2018), Bangkok, Thailand, 10–13 December 2018; Association for Computing Machinery: New York, NY, USA, 2018; Volume 17, pp. 1–8. [\[CrossRef\]](#)
19. Ekman, P. Basic Emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1999; Chapter 3; pp. 45–60. [\[CrossRef\]](#)
20. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391.
21. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech 2005, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520. [\[CrossRef\]](#)
22. Posner, J.; Russell, J.A.; Peterson, B.S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715–734. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalande, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8. [\[CrossRef\]](#)
24. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [\[CrossRef\]](#)
25. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower Provost, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
26. Prasanth, S.; Roshni Thanka, M.; Bijolin Edwin, E.; Nagaraj, V. Speech emotion recognition based on machine learning tactics and algorithms. *Mater. Today Proc.* **2021**. [\[CrossRef\]](#)
27. Akçay, M.B.; Oguz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [\[CrossRef\]](#)
28. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [\[CrossRef\]](#)
29. Ancilin, J.; Milton, A. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl. Acoust.* **2021**, *179*, 108046. [\[CrossRef\]](#)
30. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462. [\[CrossRef\]](#)
31. Boersma, P.; Weenink, D. PRAAT, a system for doing phonetics by computer. *Glott Int.* **2001**, *5*, 341–345.
32. Bhavan, A.; Chauhan, P.; Hitkul; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [\[CrossRef\]](#)
33. Singh, P.; Srivastava, R.; Rana, K.; Kumar, V. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowl.-Based Syst.* **2021**, *229*, 107316. [\[CrossRef\]](#)
34. Pepino, L.; Riera, P.; Ferrer, L. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 3400–3404. [\[CrossRef\]](#)
35. Issa, D.; Fatih Demirci, M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [\[CrossRef\]](#)
36. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* **2021**, *102*, 107101. [\[CrossRef\]](#)
37. Atila, O.; Şengür, A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoust.* **2021**, *182*, 108260. [\[CrossRef\]](#)
38. Wijayasingha, L.; Stankovic, J.A. Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health* **2021**, *19*, 100165. [\[CrossRef\]](#)
39. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37. [\[CrossRef\]](#)
40. Akhand, M.A.H.; Roy, S.; Siddique, N.; Kamal, M.A.S.; Shimamura, T. Facial Emotion Recognition Using Transfer Learning in the Deep CNN. *Electronics* **2021**, *10*, 1036. [\[CrossRef\]](#)
41. Ahmad, Z.; Jindal, R.; Ekbal, A.; Bhattacharyya, P. Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding. *Expert Syst. Appl.* **2020**, *139*, 112851. [\[CrossRef\]](#)
42. Amiriparian, S.; Gerczuk, M.; Ottl, S.; Cummins, N.; Freitag, M.; Pugachevskiy, S.; Baird, A.; Schuller, B. Snore Sound Classification Using Image-Based Deep Spectrum Features. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3512–3516.
43. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [\[CrossRef\]](#)

44. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Association for Computational Linguistics (EMNLP 2020), Virtual Conference, 16–20 November 2020; pp. 38–45.
45. King, D.E. Dlib-ML: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
46. Nguyen, B.T.; Trinh, M.H.; Phan, T.V.; Nguyen, H.D. An efficient real-time emotion detection using camera and facial landmarks. In Proceedings of the 2017 Seventh International Conference on Information Science and Technology (ICIST), Da Nang, Vietnam, 16–19 April 2017; pp. 251–255. [\[CrossRef\]](#)
47. Poulou, A.; Kim, J.H.; Han, D.S. Feature Vector Extraction Technique for Facial Emotion Recognition Using Facial Landmarks. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 20–22 October 2021; pp. 1072–1076. [\[CrossRef\]](#)
48. Ekman, P.; Friesen, W.V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978. [\[CrossRef\]](#)
49. Sanchez-Mendoza, D.; Masip, D.; Lapedriza, A. Emotions Classification using Facial Action Units Recognition. In *Artificial Intelligence Research and Development: Recent Advances and Applications*; Frontiers in Artificial Intelligence and Applications; Museros, L., Pujol, O., Agell, N., Eds.; Catalan Assoc Artificial Intelligence: Barcelona, Spain, 2014; Volume 269, pp. 55–64. [\[CrossRef\]](#)
50. Yao, L.; Wan, Y.; Ni, H.; Xu, B. Action unit classification for facial expression recognition using active learning and SVM. *Multimed. Tools Appl.* **2021**, *80*, 24287–24301. [\[CrossRef\]](#)
51. Senechal, T.; Bailly, K.; Prevost, L. Impact of Action Unit Detection in Automatic Emotion Recognition. *Pattern Anal. Appl.* **2014**, *17*, 51–67. [\[CrossRef\]](#)
52. Bagheri, E.; Esteban, P.G.; Cao, H.L.; De Beir, A.; Lefebvre, D.; Vanderborght, B. An Autonomous Cognitive Empathy Model Responsive to Users' Facial Emotion Expressions. *ACM Trans. Interact. Intell. Syst.* **2020**, *10*, 20. [\[CrossRef\]](#)
53. Baltrušaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66. [\[CrossRef\]](#)
54. Tautkute, I.; Trzcinski, T. Classifying and Visualizing Emotions with Emotional DAN. *Fundam. Inform.* **2019**, *168*, 269–285. [\[CrossRef\]](#)
55. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* **2021**, *21*, 3046. [\[CrossRef\]](#)
56. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
57. Kim, J.H.; Poulou, A.; Han, D.S. The Extensive Usage of the Facial Image Thresholding Machine for Facial Emotion Recognition Performance. *Sensors* **2021**, *21*, 2026. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Huang, S.C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **2020**, *3*, 136. [\[CrossRef\]](#)
59. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Sun, L.; Xu, M.; Lian, Z.; Liu, B.; Tao, J.; Wang, M.; Cheng, Y. Multimodal Emotion Recognition and Sentiment Analysis via Attention Enhanced Recurrent Model. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, Virtual Event China, 24 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 15–20. [\[CrossRef\]](#)
61. Sun, L.; Lian, Z.; Tao, J.; Liu, B.; Niu, M. Multi-Modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop, Seattle, WA, USA, 16 October 2020; pp. 27–34. [\[CrossRef\]](#)
62. Deng, J.J.; Leung, C.H.C. Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition. In *Brain Informatics*; Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N., Eds.; Springer: Cham, Switzerland, 2021; pp. 179–188.
63. Pandeya, Y.R.; Lee, J. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimed. Tools Appl.* **2021**, *80*, 2887–2905. [\[CrossRef\]](#)
64. Abdulmohsin, H.A.; Abdul wahab, H.B.; Abdul hossen, A.M.J. A new proposed statistical feature extraction method in speech emotion recognition. *Comput. Electr. Eng.* **2021**, *93*, 107172. [\[CrossRef\]](#)
65. García-Ordás, M.T.; Alaiz-Moreton, H.; Benítez-Andrades, J.A.; García-Rodríguez, I.; García-Olalla, O.; Benavides, C. Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. *Biomed. Signal Process. Control.* **2021**, *69*, 102946. [\[CrossRef\]](#)
66. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 2426–2430. [\[CrossRef\]](#)

67. Baeovski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
68. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, 20–25 June 2020; pp. 4211–4215.
69. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, *2006*, 10.
70. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
71. Baltrusaitis, T.; Mahmoud, M.; Robinson, P. Cross-Dataset Learning and Person-Specific Normalisation for Automatic Action Unit Detection. In *Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, Slovenia, 4–8 May 2015; pp. 1–6. [\[CrossRef\]](#)
72. Baziotis, C.; Nikolaos, A.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; Potamianos, A. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, Orleans, LA, USA, 5–6 June 2018. [\[CrossRef\]](#)
73. Romero, S.E.; Kleinlein, R.; Jiménez, C.L.; Montero, J.M.; Martínez, F.F. GTH-UPM at DETOXIS-IberLEF 2021: Automatic Detection of Toxic Comments in Social Networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, Co-Located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, 21 September 2021; Volume 2943, pp. 533–546.
74. Pavlopoulos, J.; Malakasiotis, P.; Androutsopoulos, I. Deep Learning for User Comment Moderation. In *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada, 4 August 2017. [\[CrossRef\]](#)
75. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 8024–8035.
76. Dissanayake, V.; Zhang, H.; Billinghamurst, M.; Nanayakkara, S. Speech Emotion Recognition ‘in the Wild’ Using an Autoencoder. In *Proceedings of the Interspeech 2020*, Shanghai, China, 25–29 October 2020; pp. 526–530. [\[CrossRef\]](#)