applied sciences

Article

Analysis of Deep Learning-Based Decision-Making in an Emotional Spontaneous Speech Task

Mikel de Velasco , Raquel Justo *, Asier López Zorrilla and María Inés Torres

Department of Electricity and Electronics, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain; mikel.develasco@ehu.eus (M.d.V.); asier.lopezz@ehu.eus (A.L.Z.); manes.torres@ehu.eus (M.I.T.)

* Correspondence: raquel.justo@ehu.eus

Abstract: In this work, we present an approach to understand the computational methods and decision-making involved in the identification of emotions in spontaneous speech. The selected task consists of Spanish TV debates, which entail a high level of complexity as well as additional subjectivity in the human perception-based annotation procedure. A simple convolutional neural model is proposed, and its behaviour is analysed to explain its decision-making. The proposed model slightly outperforms commonly used CNN architectures such as VGG16, while being much lighter. Internal layer-by-layer transformations of the input spectrogram are visualised and analysed. Finally, a class model visualisation is proposed as a simple interpretation approach whose usefulness is assessed in the work.

Keywords: emotion detection; speech processing; explainable artificial intelligence; machine learning

check for updates

Citation: de Velasco, M.; Justo, R.; López Zorrilla, A.; Torres, M.I. Analysis of Deep Learning-Based Decision-Making in an Emotional Spontaneous Speech Task. *Appl. Sci.* **2023**, *13*, 980. https://doi.org/10.3390/app13020980

Academic Editor: Douglas O'Shaughnessy

Received: 30 November 2022
Revised: 3 January 2023
Accepted: 8 January 2023
Published: 11 January 2023

1. Introduction

Emotion theories agree that an emotional episode consists of several components, such as the stimulus, motivation for action, central and peripheral physiological responses, behaviour (e.g., facial and vocal expressions, among others) and subjective experiences or feelings [1]. In addition, the behavioural and physiological expression of emotions and the stimulus quality depend on the person and on the specific scenario [2].

Affective computing often uses a categorical model based on a set of predefined emotional labels that are roughly supported by the basic emotions defined by the affect program theory [3], which might cover the whole emotional space. Each basic emotion encompasses a wide subset of emotions that can be understood as blends or elaborations of the basic ones. An alternative theory [1] discriminates emotions on the basis of combinations of continuous variables aiming to characterise the contents of feelings [4]. Typical variables are Valence and Arousal, which define a 2D space for representations, even though Dominance has also been proposed, resulting in a 3D model usually called a VAD model.

Furthermore, emotional responses result in changes in gaze, facial and vocal expressions, speaking style, in the way the language is used as well as in changes in physiological signals, such as the electroencephalographic signals or galvanic skin responses, among others [5]. The information provided by each signal has distinctive features, which can be complementary. This might result in a variety of approaches and systems for emotion recognition with different goals and application tasks [6]. However, universal facial expressions [7], also considered as short-term stereotypical responses, are the more extensively analysed emotional expressions.

Speech signals encode speaking styles, paralinguistic features, the usage of language, message contents to be transmitted, environmental sounds, etc., which contain varied information about the speaker profile, intent, current emotional status, and even information about some mental diseases [8].

In contrast to this complexity, computational researchers of emotions need an exact ground truth to be used for supervised learning, decoding and evaluating computational models of emotions. Usually, human annotators establish their own perception of the emotional data as the ground truth and reference for the automatic identification of emotions. These perceptual experiments add subjectivity and complexity to the already complex and, to some extent, subjective emotional constructions, mainly in speech processing.

In the next step, researchers submit the data to black boxes, i.e., to complex architectures of neural networks, whose behaviour is not fully understood but that might perform well in terms of usual scores. Therefore, the key to successful or unsuccessful classification rates remains unknown. In other words, we are unaware of what the computational model identifies as emotional cues.

Over the last few years, some techniques have been proposed to explain the internal behaviour of complex computational models, resulting in what is called XAI, i.e., eXplainable Artificial Intelligence. Some of them propose simple models that can represent the aforementioned external behaviour. However, the most amount of effort has been put into the image analysis domain because the action of the network on the original images can be visually represented and, thus, it can be more easily understood. In contrast, XAI methodologies have been scarcely used in voice processing.

The aim of our work is to contribute to the understanding of computational methods and decision-making involved in the identification of emotions in spontaneous speech. To this end, we selected a task consisting of TV debates in which spontaneous emotions can be investigated. To this end, we follow the transformation of the input data, from layer to layer, until the classification is carried out in the output layer of the network. If the whole architecture becomes too deep, this process is hard and it becomes difficult to extract valuable conclusions. Thus, we propose a CNN-based deep architecture capable of providing good results but simple enough to be able to follow and interpret the decisions taken.

The main contributions of the work can be summarized as follows:

- We develop a multitask architecture to simultaneously classify discrete categories and VAD dimensions, in the aforementioned realistic task. This requires a previous annotation of the corpus in terms of both categorical and VAD models through human perception experiments, which define the ground truth. The proposed model is also compared to a more complex state-of-the-art image processing network such as VGG-16 [9], resulting in a better performance (even if slightly) for the target task.
- In an attempt to explain the decisions of our automatic system, we analyse the evolution of the categorical representations of our model layer-by-layer. Thus, we analyse the evolution of the data until they become predictions, i.e., from input spectrograms to the results.
- As a final contribution, we use the spectrogram to parameterise the voice signal to process it as an image. This allows us to obtain a visual class model [10] (deep dream) that can be used to visualise the patterns learnt by the proposed network. This technique is widely used when dealing with images, but as far as we know, it has never been applied to speech.

The paper is organised as follows: Section 2 reports some related works. Section 3 describes the methodology selected to develop an automatic recogniser of emotions from spontaneous speech. This section includes the description of the task and corpus, the neural network model proposed for the joint classification of categories and emotional dimensions, and also a comparison of classification results obtained with our network and a pretrained VGG-16 net. Then, Section 4 deals with the interpretation of the model behaviour. It first presents a joint analysis of the results in terms of both categories and dimensions. Then, the evolution of the model across the layers is visualised and examined. Finally, the proposed simple interpretation model, i.e., the class visualisation model is introduced and assessed. Finally, Section 5 reports the main conclusions of this work.

2. Related Work

Affective computing has become more relevant due to its impact on person–computer interactions [11,12]. This has translated into significant progress in all its modalities [13–15]: face [7,16], gestures [17], text [18–21], audio [22] and others. Some investigations do not only focus on a single modality but also multi-modal approaches [23].

With regard to the detection of emotion from facial cues, most studies deal with the categorical model for emotional state representation. Within this framework, the most employed set of emotions is the one proposed by Ekman [3], which is widely accepted under the name of "The Big Six" [24]. Ekman's proposal consists of six basic and universal emotions: surprise, disgust, sadness, anger, fear, and happiness. However, in other works dealing with speech, emotions are also represented using a dimensional model [25–27]. Dimensional theories postulate that the vast array of emotions cannot be simplified to a basic set but can be mapped to a continuous low-dimensional spatial representation. Most of the works in this context propose a two-dimensional model, comprised of valence (whether the emotional state is positive/pleasant or otherwise negative/aversive) and arousal (intensity or level of arousal) [28,29]. Dominance is a third dimension also included in some works that encodes the level of control (leading to feelings of power/dominance or weakness/submission) [30]. The aforementioned two models (categorical and dimensional) show a close relationship according to the Core Affect theory [4], where each categorical emotion is represented in a point/area of the dimensional model. An example of this is the illustration of Sherer's circumplex [31], which makes use of the arousal/valence two-dimensional model to represent categorical emotions. In this work, the two models were considered to represent the emotional status of the speakers because they can complement each other.

In order to build an emotion detection system from scratch, annotated data are needed, assuming the supervised machine learning paradigm. Finding corpora where real emotions appear is a really difficult task. Thus, most of the research in this field relies on data sets where emotions have been acted or forced [26], as occurs with the EMODB [32] or IEMOCAP [33] corpora. However, in recent years, there has been an attempt to put emphasis on creating corpora with spontaneous emotions such as AVEC2012 [34], EmoL6N [35] or DBATES [36]. However, this is a challenging task because, on the one hand, the perception of emotions is not as intense as in the corpora with acted emotions [35] and, on the other hand, the annotation procedure is very subjective, leading to low inter-annotator agreements [37–39]. This work deals with a task in which spontaneous emotions are involved. This entails an additional challenge for the system that has to deal with speech chunks with subjective and subtle emotional representations.

Another important issue to be addressed is how to identify the most suitable features, i.e., speech representations, for detecting emotions. In recent years, there have been several attempts to build a set of features suitable for the identification of emotions in the speech signal [40,41]. Several works are based on Low-Level Descriptors [42–46], whose characteristics are related to prosody (pitch, formants, energy, jitter and shimmer), the spectrum (centroid, flux, entropy) and their functionals (mean, std, quartiles 1–3, delta, etc.). In this context, Ref. [47] proposes the GeMAPS set of speech features that has been considered as a standard. However, and thanks to challenges such as INTERSPEECH [48], other sets have also been proposed (ComParE) to become a reference in this area. However, none of these sets has actually proven to be superior to the rest in a global environment. Several works [43,46,49] suggest that there are no universal acoustic features that extract the emotional content and work well in all contexts. In this direction, some works propose working with the spectrogram [45,46,50–54] since it contains almost all the information about a speech signal. More recently, self-learning based approaches [55,56] have shown to find good representations of the speech signal. Indeed, self-learning has been applied in a variety of applications of speech processing with successful results in solving different tasks [56–59]. In fact, our task has already been addressed using such speech representations [2]. However, pretrained (and not fine-tuned) models were needed to obtain good results, and thus the analysis of the decisions made by such models cannot be easily conducted.

The understanding of which patterns are detected by deep neural networks, or how they work, can help to design new architectures or new learning paradigms that can make a difference. The introduction of such advances has been decisive to achieve the current automatic systems' performance. One of the clearest examples appears when the field of computer vision introduces convolutional [60] and pooling [61] networks, which marked the beginning of a new age. In 1997, the recurrent networks based on LSTM cells were proposed [62], which are known for their ability to process long sequences that were first used in NLP or speech processing. However, attention networks [63,64] have been the ones that have made progress in the NLP field. In emotion recognition, although some work has been conducted [2,65,66], promising results have not yet been achieved. Different works based on CNNs [67], LSTMs [68,69] and attention mechanisms [67,69] have reached accuracy values (or F1 scores) of around 0.7 with the most commonly employed acted data sets.

Moreover, this kind of deep neural architecture is sometimes so complex that even experts are hardly able to interpret it [70]. As a consequence, understanding the behaviour of a model to make predictions is becoming as important as its accuracy. In fact, interpretability is nowadays a key to improving the performance of complex neural architectures. Several methodologies have recently been proposed to explain the importance of particular features for decision-making. These methodologies are sometimes integrated into the models, but they also very often consist of postprocessing analysis and models [71]. On the other hand, explanation and interpretation are context-, domain- and task-dependent concepts. In this way, XAI targets are the end-users who depend on the decisions taken by the automatic system [72]. Some recent works also argue that explanations must be related to the perceptual process from cognitive psychology [73]. In brief, XAI is still a domain to be explored by AI researchers in relation to the domain addressed. Due to the intuition of vision and the availability of data, much of the XAI research has focused on image prediction tasks [10]. On the contrary, few techniques have been developed for audio or speech prediction [6,74,75]. In this work, an architecture based on CNNs and inspired by computer vision was designed. Moreover, employing the spectrogram as the input allows us to represent the audio as an image and apply XAI image processing techniques.

In summary, we propose an emotion detection system capable of providing two emotion representation levels: a categorical one and another one based on three-dimensional VAD space. The proposed model is simple enough to allow a detailed analysis of the network behaviour layer by layer while providing accurate classification results. In order to increase the level of explainability of the decisions made by the network, the Visual Model Classification XAI technique was selected. To this end, the spectrogram was selected as the input of the system along with a CNN-based deep neural architecture.

3. Emotion Detection

In this section, we describe the selected neural model capable of detecting emotions in TV debates. This model is inspired by previous works [2,39], but it has been adapted for the joint classification in terms of both categories and emotional dimensions.

3.1. Task and Corpus

This task consists of human–human spontaneous conversations gathered from the La Sexta Noche Spanish TV program. This TV show addresses the hot news of the week in social and political debate panels led by two moderators. A very wide range of talk-show guests (politicians, journalists, etc.) analyse social topics from their perspectives. Given that the topics under discussion are usually controversial, it is expected to have emotionally rich interactions. However, the participants are used to speaking in public so they mostly do not lose control of the situation. Nevertheless, even if participants might overreact sometimes, it is a real scenario in which emotions are subtle.

In order to build a corpus, the programs of La Sexta Noche broadcasted during the electoral campaign of the Spanish general elections in December 2015 were selected. Then, speech signals were extracted from the videos of the TV shows. Then, they were split into shorter segments or chunks. The segments have to be short enough to avoid changes in emotional content but long enough to allow for their identification. Thus, the speech signal was divided into clauses. A clause was defined as "a sequence of words grouped together on semantic or functional basis" [76], and it can be hypothesised that the emotional state does not change inside a clause. An algorithm that considered silences and pauses, as well as text transcriptions, was designed to identify the utterances compatible with the clauses [39]. This produced a set of 5500 audio chunks in Spanish, ranging from two to five seconds long that was used as our data set. Regarding the speaker features, the resulting gender distribution in the processed data was 30% female and 70% male, with a total number of 238 different speakers and an age ranging from 35 to 65. These data just reflect the nature of the described TV shows.

The corpus was emotionally annotated in the framework of the AMIC "Affective multimedia analytics with inclusive and natural communication" project [77], as described in [2,39]. The annotation was carried out through perception experiments in which crowd annotators were asked to identify both emotional categories and Valence–Arousal–Dominance dimensions [78,79]. A crowdsourcing platform [38] was used to gather five annotations for each audio chunk. All the annotators filled out a questionnaire related to the perceived emotions in each audio chunk, and an agreement higher than 60% was required as a quality guarantee for the categorical annotation [35,39]. The questionnaire related to the dimensional model considered discrete labels to facilitate the crowd annotation process. However, instead of asking a 60% of agreement again, a consensus of the different annotations was achieved by converting the labels to real values, as shown below:

- Valence: Positive = 1, Neutral = 0.5, Negative = 0;
- Arousal: Excited = 1, Slightly excited = 0.5, Neutral = 0;
- Dominance: Rather Dominant = 1, Neutral = 0.5, Rather intimidated = 0.

Then, the average values attached to each audio chunk by different annotators were computed. In this way, each label of the VAD model corresponds to a real value. This scenario suggests a classification problem for the categorical model and a regression problem for the VAD. However, previous works [39] showed that the regression problem might be too ambitious for this task, and if it is addressed as a classification task, a better performance might be achieved. Specifically, the distribution of the annotated data for each of the VAD dimensions was analysed. According to these distributions, a discretisation of each VAD dimension was carried out in order to learn a categorical classifier to predict each of the discretised classes. This procedure led to a set of 4118 annotated chunks distributed, as shown in Table 1. Let us note that Table 1 shows a high imbalance between classes. The tendency to neutrality that is observed is related to the spontaneity conditions in which the corpus was acquired.

Table 1. Class distribution of the annotated data for categorical and VAD model.

Categorical Model (%)	Dimensional Model		
	Arousal (%)	**Valence (%)**	**Dominace (%)**
Angry: 30.2 Happy: 15.3 Calm: 54.5	Excited: 25.5 Neutral: 74.5	Positive: 29.0 Neutral: 54.4 Negative: 16.6	Dominant: 26.2 Neutral: 73.8

3.2. Convolutional Neural Model

Convolutional neural architectures have become a standard in image processing over the last few years. However, other types of tasks have also taken advantage of these architectures by adapting the problem and addressing it with computer vision techniques. For example, audio analysis can be performed with computer vision techniques if the audio is represented by a spectrogram. In addition, other areas such as speech and natural language processing have also taken advantage of the potential of convolutions for the analysis of temporal sequences [80].

In this work, we propose a simple and light convolutional network architecture (a network with 43K parameters) and compare it with the VGG16 convolutional network [9], a model widely used in the literature but that consists of 134M parameters, which makes it difficult to understand its behaviour.

Both network architectures are designed to obtain, from the speech spectrogram, the joint classification of emotional state in terms of both representations, the categorical model and VAD dimensions, as shown in Figure 1. On the one hand, after a number of convolutional and pooling layers are applied, three scalar values corresponding to each dimension of the VAD model are computed through three linear layers with a sigmoid activation function at point A. These values are then converted to discrete VAD predictions linearly, at point B. On the other hand, the categorical model is inferred in two ways. First, at point C, the categories are predicted based on the output of the CNNs. Second, at point D, the scalar predictions of the VAD model are used instead. Intuitively, this second prediction might perform better as it could explicitly take advantage of the multi-tasking capabilities of the networks.

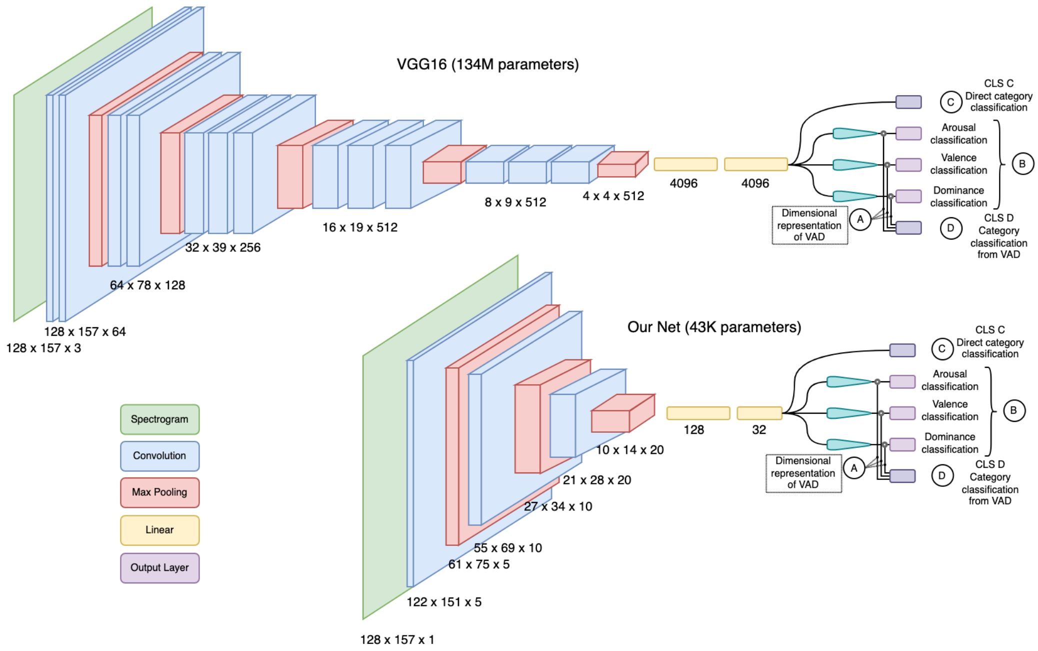

Figure 1. Representation of the structure of the VGG16 [9] network and our proposal. In both cases, the convolutions (blue boxes) and max-poolings (red boxes) help to extract a set of appropriate features, while the linear ones are in charge of performing the class discrimination. Point A represents a dimensional prediction of the VAD model, while point B provides the discretisation of each of the dimensions. On the other hand, the categorical model is inferred directly in point C, and from the scalar VAD model predictions, in point D.

As for the training procedure, a resampling strategy is used when training both networks to deal with the imbalance of the data, as observed in Table 1. The employed method consisted of selecting the samples inside a batch using a random function that can take into account the weight given to each sample. The weight W_x for each sample x was computed, as Equation (1) shows

$$W_x = min\left\{ \frac{|X|}{|X_c|}, \beta \cdot \min_{\forall c \in C} \left\{ \frac{|X|}{|X_c|} \right\} \right\} \tag{1}$$

where $|X|$ is the number of samples in the corpus, $|X_c|$ is the number of samples in the class which sample x belongs to, and β is the oversampling coefficient (in this work it was chosen a value $\beta = 2$). In this way, the samples of the minority classes appear proportionally more times, but never more than twice as much as a sample from the majority class.

Regarding the optimisation hyperparameters, the Adam optimiser was chosen with a learning rate of 10^{-4} and with a batch size of 16. The models were trained throughout 7K iterations. Note that our network is trained from scratch, whereas VGG16 is fine-tuned from the publicly available pretrained checkpoint. The cross-entropy loss function is used for each classification task. The five losses (two for the categorical model, and three for the VAD model) are then averaged (with different weights) to obtain the final loss. The weights for the categorical losses were half the weights for the VAD model, since this led to the best results, empirically.

3.3. Classification Results

The performance of the proposed classification system for both emotion categories and VAD dimensions are compared to the VGG16 model in Tables 2 and 3. All results were obtained after a 10-fold cross-validation procedure. For the categorical model, two different results are given, one corresponding to the direct categorical classification associated with output B in Figure 1 (CLS C) and another one making use of the predicted VAD floating point values associated with output D (CLS D). The average and standard deviation of five metrics commonly employed to evaluate emotion classification systems [81,82] are reported: F1 score, Unweighted Accuracy (UA, also known as balanced accuracy or unweighted average recall), average precision, Matthews Correlation Coefficient and Area Under the ROC Curve (AUC). Additionally, paired *t*-tests were computed to assess the statistical significance of the performance differences between our model and the VGG16 network.

Table 2. Classification performance of our proposal and VGG16 for the categorical model prediction task. Two ways of predicting the emotion categories were tested. Comparisons where a network significantly outperforms the other (i.e., *p*-value < 0.05) are marked with the symbol *. The best result for each comparison is highlighted in bold.

Our Net/VGG16	**CLS C**	**CLS D**
F1	**0.58** ± **0.04**/0.57 ± 0.06	**0.59** ± **0.05**/0.57 ± 0.05
UA	**0.57** ± **0.04**/0.54 ± 0.06	**0.58** ± **0.04**/0.54 ± 0.05
Average precision	0.60 ± 0.05/**0.63** ± **0.07**	0.60 ± 0.06/**0.63** ± **0.07**
Matthews corr. coef.	**0.39** ± **0.06**/0.38 ± 0.08	**0.41** ± **0.05**/0.38 ± 0.06
AUC	**0.80 *** ± **0.03**/0.75 ± 0.03	**0.81 *** ± **0.02**/0.74 ± 0.04

Table 3. Classification performance of our proposal and VGG16 for the VAD model prediction task. Comparisons where a network significantly outperforms the other (i.e., *p*-value < 0.05) are marked with the symbol *. The best result for each comparison is highlighted in bold.

Our Net/VGG16	**Arousal**	**Valence**	**Dominance**
F1	**0.67** ± **0.11**/0.67 ± 0.02	**0.42** ± **0.03**/0.41 ± 0.05	**0.57** ± **0.05**/0.56 ± 0.03
UA	**0.67** ± **0.09**/0.66 ± 0.03	**0.45 *** ± **0.04**/0.41 ± 0.03	**0.57** ± **0.03**/0.56 ± 0.03
Average precision	0.69 ± 0.13/**0.69** ± **0.02**	0.44 ± 0.05/**0.45** ± **0.09**	0.58 ± 0.07/**0.58** ± **0.04**
Matthews corr. coef.	0.35 ± 0.17/**0.35** ± **0.03**	**0.14** ± **0.04**/0.12 ± 0.04	**0.15** ± **0.06**/0.14 ± 0.06
AUC	**0.74** ± **0.11**/0.72 ± 0.02	**0.65 *** ± **0.03**/0.60 ± 0.02	**0.63** ± **0.07**/0.60 ± 0.03

First, and most importantly, our network performs slightly better than the fine-tuned VGG16 network for most classification tasks and metrics. This is already remarkable because our architecture uses around 3000 times fewer parameters than the VGG16 CNN. Furthermore, the differences in performance are statistically significant in four comparisons: when measuring the AUC for CLS C, CLS D and arousal, and also for the Unweighted Accuracy in the arousal prediction task. Importantly, these results support the use of our light network to apply XAI techniques.

If we further analyse the results, Table 2 shows a slight tendency towards a better performance of CLS D, i.e., the classifier that uses the predicted VAD values before classifying the categories, particularly in the case of our proposed network. Looking at the results for the VAD dimensional classification of Table 3, it can be concluded that the best results are achieved for Arousal. However, it should be noticed that, in this case, there are only two different categories (Excited and Neutral), whereas, in Valence, there are three different ones (Positive, Neutral and Negative).

For a better understanding of the VAD results, Figure 2 shows the VAD predicted values vs. the annotated, i.e., perceived, values. Straight lines in the figure show the borderlines learnt to discretise the problem, i.e., to transform the regression problem into a categorisation one. This figure shows the good performance of our proposal. In fact, opposite diagonals are very low-density regions. When it comes to Arousal, for instance, it seems to be easier to predict accurately higher values than lower ones that are more scattered in the lower part of Figure 2. In Valence, it can be concluded again that positive and negative categories are sometimes mixed with Neutral, but rarely among each other (see secondary diagonal in the figure). Finally, Dominance shows a lower correlation between the predicted and annotated values.

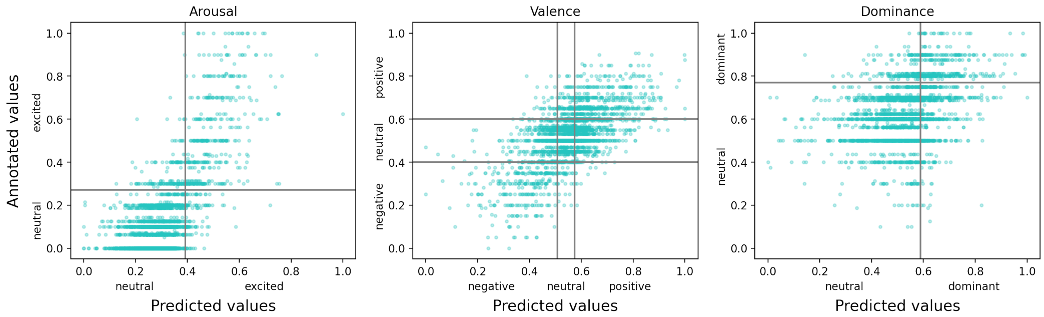

Figure 2. Comparison between the samples annotated by humans and predicted by the network in the dimensional model. Each of the plots depicts a VAD dimension. Each sample is placed at a point based on the actual value of the annotations (y-axis) and the actual value of the predictions of the network (x-axis). The lines show how the samples have been separated for the classification problem on each dimension (y-axis) and how the network has separated them (x-axis).

4. Interpreting the Model Behaviour

This section aims to explain the proposed model's decisions, providing a better understanding of the work of each layer, as well as an analysis of the input features learnt by the model.

4.1. Analysis of the Classification Results

Figure 3 shows the three projections of the VAD values in a 2D space. The colours of the points show the category they belong to. In the first row, both the VAD points and the categories are the ones perceived by annotators. In the second row, the VAD points are the ones predicted by the network (Point A) and coloured according to the perceived categories. Finally, in the third row, the colour of the points corresponds to the predicted categories. Specifically, the output of the network at Point D (CLS D) has been used, since it outperforms CLS C in our experiments.

The first row of Figure 3 shows mixed VAD points. In fact, the subjectivity of human perception witnessed during the annotation procedure makes it difficult to obtain clear boundaries between classes. However, some patterns can still be observed. For example, "Angry" samples present higher arousal than "Calm" and "Happy", which can be seen in the first and second plots. In terms of valence, the distinction between the three categories is a bit more clear: "Happy" is the most positive emotion, followed by "Calm" which is neutral, and "Angry", which indicates negative arousal. This result is clearly aligned with the literature on emotion theory. Finally, we would like to mention that the dominance axis does not show clear boundaries between the categories.

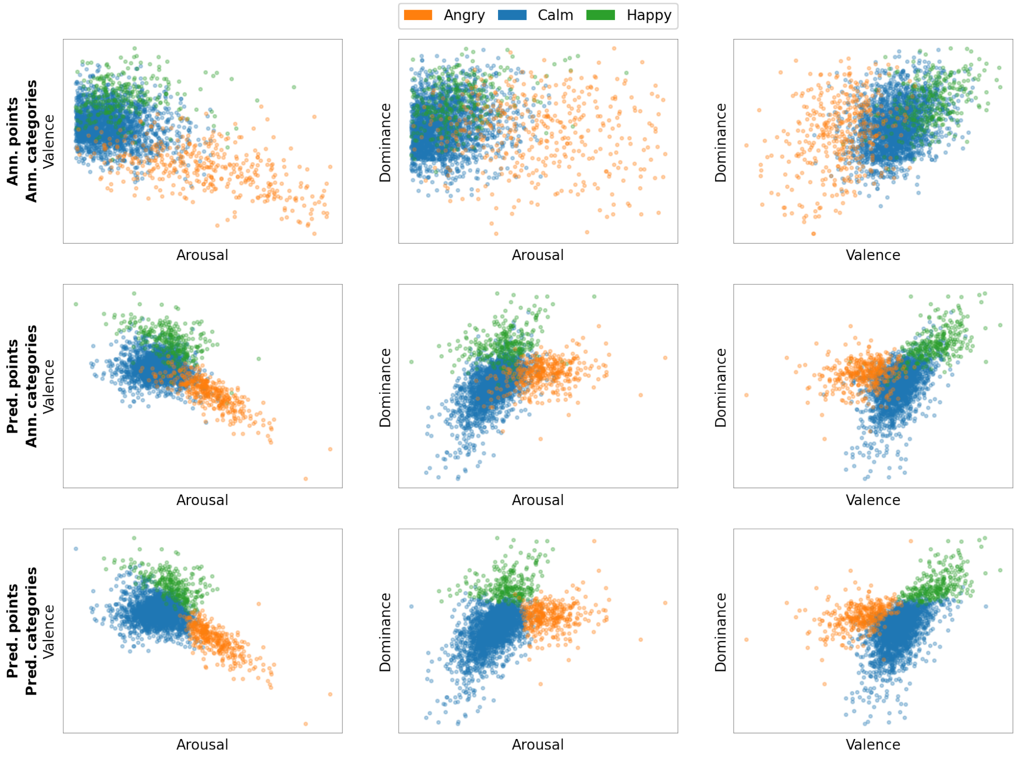

Figure 3. Correlation between the VAD and categorical models, according to both the annotated and predicted data.

A transformation of the space is observed in rows two and three; points are no longer located in the same place as in the first row of Figure 3. Instead, they correspond to the model's predictions. The difference between the second and third rows is that in the third one we can clearly see the boundaries learnt by the network to decide to which category a sample belongs to. In this new space, the categories can be better separated, even the annotated ones in the second row. It can be hypothesised that this fact is due to the joint training of the categories and VAD dimensions. The simultaneous VAD and categorical classifications result in the collaboration of both models in decision-making. Thus, the regions associated with categories are well-defined in the VAD projection space. Finally, similar relations of the categories and the VAD axes can be seen in the last two rows. Moreover, in this case, the dominance axis shows that, as expected, "Calm" is less dominant than "Happy" and "Angry".

4.2. Evolution of the Model

In this section, we show a representation of the work that each layer of the deep network is carrying out. To this end, the progress in the training stage can be explained as a fine-tuning process of the data representation, which can lead to a good classification. For this purpose, we present the output provided by each layer for each training sample in a bidimensional space, by applying a dimensionality reduction method such as PCA. Assuming that $X = \{x_1, x_2, \ldots, x_n\}$ is the set of training samples, where x_i is the spectrogram associated with each audio chunk presented in Section 3.1, the output of the first convolutional layer for each sample can be defined as $y_i = conv1(x_i)$. This y_i output is transformed into a flattened vector that can be visualised in a 2D space by applying a decomposition in Principal Components as $y_i' = (z_1, z_2)$, where z_1 and z_2 are the two first principal components in the PCA analysis of y_i. This representation can be replicated for the output for each convolutional network (conv1, conv2 and conv3) and also for the two dense layers: linear1 and linear2.

The visualisation of the aforementioned representations is displayed in Figure 4. Points in the picture stand for the decomposition in Principal Components of the outputs of each convolutional layer. The colour of each point represents the category it belongs to. In the first row, the colours of the categories correspond to the annotated labels (ground truth), whereas in the second row the colours represent the predicted categories.

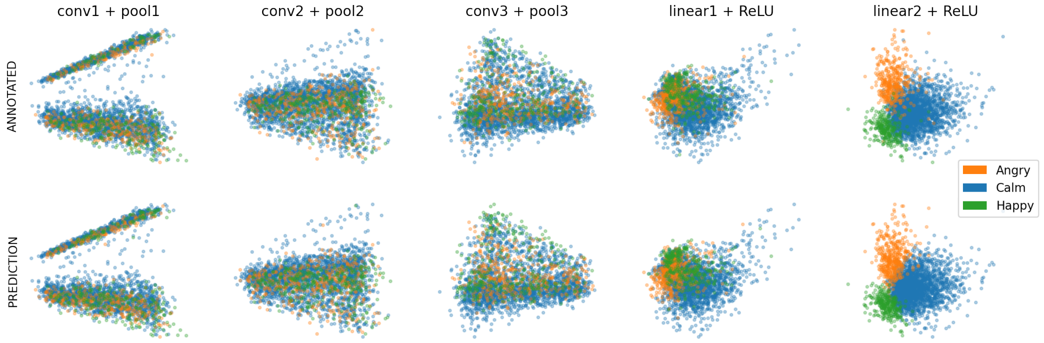

Figure 4. Two-dimensional representation obtained for each sample over different layers of the network using the PCA technique. The colour of each point represents the category to which it belongs, showing in the first row the annotated categories (ground truth) and in the second row the predicted categories.

Samples are mixed in the first stages, but as we go deep into the network, the categories are better-defined, so that the figure shows how the network learns how to differentiate them. Let us note that the dimensionality reduction in conv1 + pool1 (left image) is from 3416 to 2, whereas in linear2 + ReLU (right image) this is from 32 to 2, which could also lead to regions being better delimited, as the last picture of the right-hand side of Figure 4 shows.

In the second row, the colours are associated with the predicted categories. Thus, the regions of each category are well-defined at the final stages by clear boundaries, which supports the network's decision-making. It is interesting to note that, comparing the right-hand side pictures of annotated and predicted categories, a tendency toward "Angry" and "Happy" is shown in the predicted values. This correlates well with the values of the confusion matrices of the previous section, where the network estimates some "Calm" samples as "Angry" or "Happy". This seems to be due to the oversampling method that makes the minority classes more relevant. A more accurate sweep of the β coefficient might be useful in future work.

4.3. Class Model Visualisation

Class Model Visualisation is a global method within the Explainable Artificial Intelligence (XAI) framework, the goal of which is to generate image visualisations of each of the classes or categories the system is trying to predict [10]. We selected this method because it can provide insights into the features that the system takes into account when making those predictions in a visual way. At this point, we take advantage of the fact that we use the spectrogram to parametrise the voice signal by processing it as an image. This can be very useful for understanding the performance of the system and acting. For instance, it might be used for analysing the diversity of the samples in a category, which can be influenced by different factors such as the subjectivity in the annotation process. If a sample has a feature that the model relates to a different category that is not the predicted one, it might be due to a low agreement among the annotators, and it might be interesting to see what happens after a second annotation process.

Given a convolutional network f and a class of interest c, the goal is to generate an image visualisation I', which is representative of c. This is based on the scoring methods used to train f, which maximises the class probability score $S_c(I)$ for c, combined with a weighted (λ) L2 regularisation so that the image I' keeps regular values, such that:

$$I' = \arg\max_{I} S_c(I) - \lambda \|I\|_2^2 \tag{2}$$

Thus, the generated images (usually called Deep Dream) provide information related to what the black box model had learnt for a particular class or category in the dataset [83]. In this work, the Deep Dream images associated with the different classes, for both categorical and VAD models, are shown in Figure 5. A random spectrogram sample was selected for initialisation and an L2-regularisation method was employed to obtain the final images.

These deep dream images show different patterns for different categories. However, their interpretation is difficult since speech information is harder to interpret visually than deep dreams of images. First of all, it is worth noting that usually human speech is located in the 300–4000 Hz range, so the analysis will be focused on that interval. Focusing on the categorical model, it can be appreciated that, in "Calm", there is an intensity attenuation of around 1000 Hz, whereas, in "Happy", this is an intense interval, and the attenuation can be appreciated at lower frequencies, below 500 Hz. For the "Angry" category, the attenuation is observed below 1024 Hz, and above that frequency, there is an intense band (narrower than in the "Happy" class).

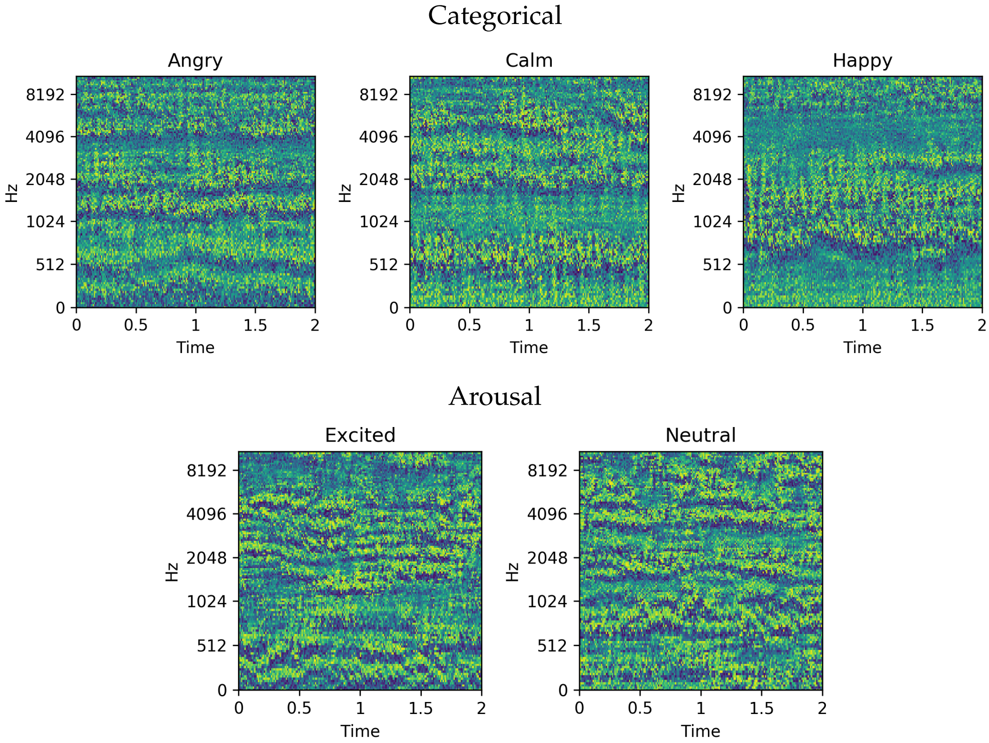

Figure 5. *Cont.*

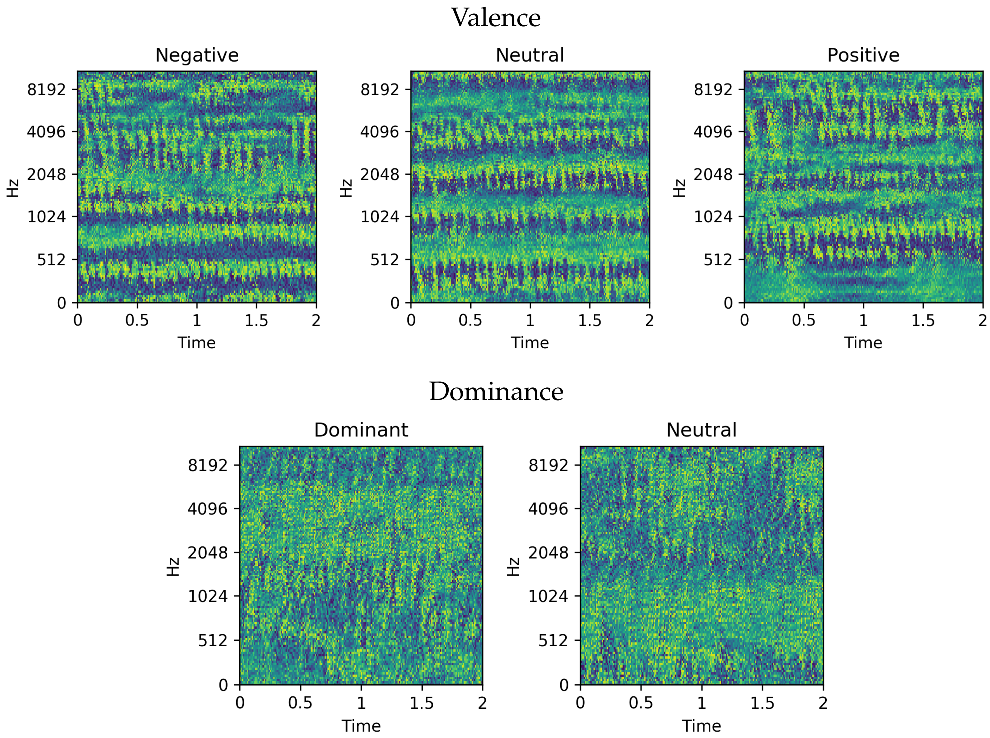

Figure 5. Extraction of the suitable spectrogram that maximises the classification output for each class using the DeepDream technique.

Switching to the VAD model, if we focus on Arousal, it can be concluded that, in “Neutral”, there is a mix of high and low frequencies that are activated. However, the “Excited” category seems to be more activated at high frequencies (above 1024 Hz). Regarding Valence, the patterns are better-defined, and there seems to be much less noise here. Clearly differentiated bands emerge in this dimension, and are located in different places for the three different values. If we compare “Negative” with “Neutral”, it can be seen that “Negative” has more defined blue bands between green or yellow bands, mainly at low frequencies. In “Neutral”, the separation among bands becomes vaguer, and a pattern is replicated all over the frequencies, which might be considered as complementary to the one appreciated in “Negative” (see frequency bands above 512 Hz, blue in “Negative” and green/yellow in “Neutral”). Finally, the vagueness among bands increases in “Positive”. Finally, the obtained images for the Dominance dimension are much noisier. However, the “Dominant” category seems to be activated at higher frequencies (2000–4000 Hz), while “Neutral” is activated at lower ones (below 1000 Hz).

In order to evaluate the effectiveness of these images to represent the categories, we tried to artificially transform all the samples in our test set into a specific category (using Deep Dream). To this end, the transformation consists of removing the profile of the deep dream image associated with the category it belongs to from each spectrogram, according to the system, and then adding the profile (deep dream) of the target category. We define the deep dream profile as a function that provides the average and standard deviation, over time, for each frequency, as Equation (3) shows. Assuming that x is the intensity associated with a time t and a given frequency f, the average and standard deviation for each frequency are computed as follows:

$$DD_{avg}(f) = \sum_{t \in \Delta t} \frac{x(f,t)}{\Delta t} \tag{3}$$

$$DD_{std}(f) = \sqrt{\left(\frac{\sum_{t \in \Delta t} (x(f,t) - DD_{avg}(f))^2}{\Delta t} \right)} \tag{4}$$

The transformation made to remove the profile of a specific category is described in Equation (5).

$$x'(f,t)_{\forall t \in \Delta t} = \frac{x(f,t) - DD_{avg}(f)}{DD_{std}(f)} \tag{5}$$

and the transformation made to add the profile of a new category is described in Equation (6):

$$x'(f,t)_{\forall t \in \Delta t} = (x(f,t) \cdot DD_{std}(f)) + DD_{avg}(f) \tag{6}$$

First, the samples that were correctly classified by the Neural Network were considered. These samples were transformed to a new category by applying the transformation in Equation (5) to each spectrogram, thus removing the profile of the category the sample belongs to. Then, the transformation on Equation (6) is applied to add the profile of the new category. These samples were firstly transformed to "Angry", then to "Calm" and finally to "Happy". Finally, the neural network classifies the transformed samples. The resulting confusion matrix is shown in Table 4.

Table 4. Confusion matrix with the percentage of correctly classified samples after profile transformation for each category (only correctly classified test samples). Each sample has been transformed to the profile of each class and therefore each row sums up to 100%.

	Angry	Calm	Happy
Angry	100	0	0
Calm	0	100	0
Happy	7.33	0	92.67

Table 4 shows that, when the transformations are applied, the system classifies the samples correctly in almost all the cases, i.e., only 7.33% of the samples transformed to "Happy" were wrongly classified as "Angry". Let us note that the transformation may introduce some noise that might lead to peak values that could be misinterpreted by the system, leading to errors. However, the good results suggest that the deep dream images are good representations of what the neural network learns for each category.

Then, all the samples of the test set, i.e., the ones not correctly classified, were considered and the process was replicated again. In this case, the predicted category was considered to remove the profile in the first step. The new results are shown in Table 5.

Table 5. Confusion matrix with the percentage of correctly classified samples after profile transformation for each category (all test samples). Each sample has been transformed to the profile of each class and therefore each row sums up to 100%.

	Angry	Calm	Happy
Angry	97.67	2.33	0
Calm	0	89.33	10.66
Happy	8.00	0	92.00

Table 5 shows some more misclassified samples for this experiment (8% of "Happy" that were classified as "Angry", 10% of "Calm" that were classified as "Happy" and 2% of "Angry" that were classified as "Calm"). In this case, there are more noisy samples because, when converting them to a new category, once the information about its prediction was removed, a higher error is achieved. For these samples, a reannotation process might be considered in order to see whether the noise comes from the subjectivity associated with the annotation procedure.

Moreover, the achieved results let us estimate which frequency bands are more relevant in each category. Let us focus, for instance, on the band above 512 Hz in the categorical model. We took a sample, labelled as “Happy”, which shows low-intensity values in the selected band. Then, the intensity values in that band were gradually increased until the spectrogram shown in Figure 6 was achieved. In this process, the values of the last layer of the network, from which the predictions were carried out, are represented in Figure 7. The figure shows how the prediction changes from “Happy” to “Calm”, which corresponds to a higher intensity band above 512 Hz in the Deep Dream image. This provides a hint to analyse samples that should be “Calm” and are predicted as “Happy” for instance. If the band above 512 Hz has low-intensity values, it might be a sample wrongly annotated as “Calm”.

Finally, it is worth mentioning that a qualitative comparison of the Deep Dream images shed some light on the achieved results. In fact, it is understandable why sometimes the system’s predictions are not accurate and some classes are mixed. If we focus on the categorical model, it is noticeable that Calm is something that is in between “Angry” and “Happy”, having similarities with both of them. However, “Happy” and “Angry” are much more different from each other. In the same way, regarding Valence, high similarity can be appreciated among “Negative”-“Neutral” and “Neutral”-“Positive”, while the differences between “Positive” and “Negative” are much more significant.

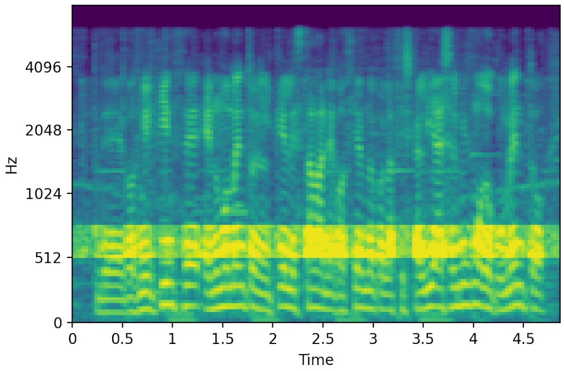

Figure 6. Spectrogram modified to alter the network prediction from “Happy” to “Calm”, intensifying frequencies above 512 Hz.

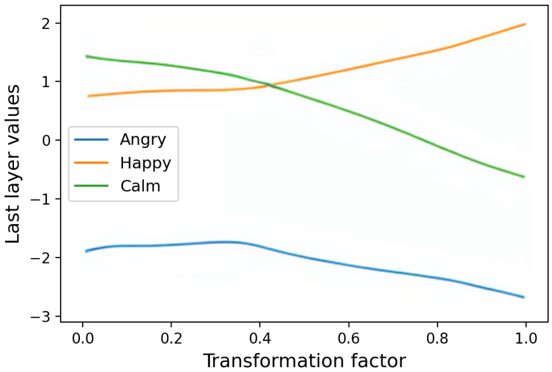

Figure 7. Representation of how the values of the network prediction change when applying changes by a factor in the spectrogram in Figure 6.

5. Conclusions

Here, we have presented a method to improve our understanding of the computational methods and decision-making involved in the identification of emotions in spontaneous speech. The selected task consists of Spanish TV debates, with a high level of complexity as well as additional subjectivity in the human perception-based annotation procedure.

Both categories of emotions and Valence–Arousal–Dominance dimensions were considered to represent emotional information. Then, a simple and light convolutional neural model was proposed to allow the joint identification of emotions using the VAD and the categorical model. The architecture of the model allows us to follow the decision-making process in order to understand where the outputs come from. The overall performance of our proposed model has also shown to be slightly higher than VGG16, a complex well-established CNN for image processing.

In this work, we focused on the understanding of the decision-making process—that is, where the errors come from and how the decisions are made. The evolution of the extracted patterns in the network layers that support their internal decisions was visualised and analysed. In addition, an XAI technique called Deep Dream was used to visualise the features related to the emotional categories. The experiments carried out show that the Deep Dream images might be an interesting tool when considering such complex neural network architectures for carrying out speech emotion detection over realistic tasks.

Author Contributions: Conceptualization, M.d.V., R.J., A.L.Z. and M.I.T.; methodology, M.d.V. and R.J.; software, M.d.V.; validation, M.d.V., A.L.Z. and R.J.; formal analysis, M.d.V., R.J., A.L.Z. and M.I.T.; investigation, M.d.V., A.L.Z. and R.J.; resources, M.d.V., R.J. and M.I.T.; data curation, M.d.V., R.J. and M.I.T.; writing—original draft preparation, M.d.V. and R.J.; writing—review and editing, M.d.V., R.J., A.L.Z. and M.I.T.; visualization, M.d.V. and R.J.; supervision, R.J. and M.I.T.; project administration, R.J. and M.I.T.; funding acquisition, R.J. and M.I.T. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this paper was conducted as part of the AMIC, AMIC-PdC, BEWORD and MENHIR projects, which received funding from the Spanish Minister of Science under grants TIN2017-85854-C4-3-R, PDC2021-120846-C43 and PID2021-126061OB-C42 and from the European Union's H2020-MSCA-RISE-2018 Research and Innovation Staff Exchange, under Grant No. 823907. The first author also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This corpus was developed by a consortium of Spanish Universities under the umbrella of AMIC, "Affective multimedia analytics with inclusive and natural communication" project. ATRESMEDIA, producer and owner of the copyright of LaSextaNoche program's contents, provided the consortium with the rights to use the audio files only for research purposes. The availability of the data are being considered by the consortium.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

VAD	Valence, Arousal, and Dominance
LSTM	Long Short-Term Memory
XAI	eXplainable Artificial Intelligence
NLP	Natural Language Processing
CNN	Convolutional Neural Network
PCA	Principal Component Analysis

References

1. Moors, A. Comparison of affect program theories, appraisal theories, and psychological construction theories. In *Categorical versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell*; John Benjamins: Amsterdam, The Netherlands, 2012; pp. 257–278.
2. de Velasco, M.; Justo, R.; Inés Torres, M. Automatic Identification of Emotional Information in Spanish TV Debates and Human-Machine Interactions. *Appl. Sci.* **2022**, *12*, 1902. [CrossRef]
3. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons: Hoboken, NJ, USA, 1999; Volume 98, p. 16.
4. Russell, J.A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **2003**, *110*, 145. [CrossRef]
5. Raheel, A.; Majid, M.; Alnowami, M.; Anwar, S.M. Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. *Sensors* **2020**, *20*, 4037. [CrossRef]
6. Egger, M.; Ley, M.; Hanke, S. Emotion recognition from physiological signal analysis: A review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55. [CrossRef]
7. Ekman, P.; Friesen, W.V.; Ellsworth, P. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 11.
8. Low, D.M.; Bentley, K.H.; Ghosh, S.S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig. Otolaryngol.* **2020**, *5*, 96–116. [CrossRef]
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
10. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
11. Brave, S.; Nass, C. Emotion in human-computer interaction. *Hum. Comput. Interact. Fundam.* **2009**, *20094635*, 53–68.
12. Richardson, S. Affective computing in the modern workplace. *Bus. Inf. Rev.* **2020**, *37*, 78–85. [CrossRef]
13. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
14. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
15. Alharbi, M.; Huang, S. A Survey of Incorporating Affective Computing for Human-System Co-Adaptation. In *Proceedings of the 2020 The 2nd World Symposium on Software Engineering*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 72–79. [CrossRef]
16. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
17. Piana, S.; Stagliano, A.; Odone, F.; Verri, A.; Camurri, A. Real-time automatic emotion recognition from body gestures. *arXiv* **2014**, arXiv:1402.5047.
18. Liu, B. Sentiment analysis and subjectivity. *Handb. Nat. Lang. Process.* **2010**, *2*, 627–666.
19. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl. Based Syst.* **2022**, *235*, 107643. [CrossRef]
20. Deng, J.; Ren, F. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
21. Li, W.; Shao, W.; Ji, S.; Cambria, E. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* **2022**, *467*, 73–82. [CrossRef]
22. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
23. Zhang, K.; Li, Y.; Wang, J.; Cambria, E.; Li, X. Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1034–1047. [CrossRef]
24. Prinz, J. Which emotions are basic. *Emot. Evol. Ration.* **2004**, *69*, 88.
25. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
26. Gunes, H.; Pantic, M. Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot. IJSE* **2010**, *1*, 68–99. [CrossRef]
27. Wöllmer, M.; Eyben, F.; Reiter, S.; Schuller, B.; Cox, C.; Douglas-Cowie, E.; Cowie, R. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In Proceedings of the 9th Interspeech 2008 Incorp 12th Australasian International Conference on Speech Science and Technology SST 2008, Brisbane, Australia, 22–26 September 2008; pp. 597–600.
28. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
29. Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* **2011**, *2*, 92–105. [CrossRef]
30. Fontaine, J.R.; Scherer, K.R.; Roesch, E.B.; Ellsworth, P.C. The world of emotions is not two-dimensional. *Psychol. Sci.* **2007**, *18*, 1050–1057. [CrossRef] [PubMed]
31. Scherer, K.R. What are emotions? In addition, how can they be measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [CrossRef]

32. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September, 2005.
33. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
34. Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; Pantic, M. AVEC 2012: The continuous audio/visual emotion challenge. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; pp. 449–456.
35. Vázquez, M.D.; Justo, R.; Zorrilla, A.L.; Torres, M.I. Can Spontaneous Emotions be Detected from Speech on TV Political Debates? In Proceedings of the 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Naples, Italy, 23–25 October 2019; pp. 289–294.
36. Sen, T.; Naven, G.; Gerstner, L.M.; Bagley, D.K.; Baten, R.A.; Rahman, W.; Hasan, K.; Haut, K.; Mamun, A.A.; Samrose, S.; et al. DBATES: Dataset of DeBate Audio features, Text, and visual Expressions from competitive debate Speeches. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
37. Blanco, R.J.; Alcaide, J.M.; Torres, M.I.; Walker, M.A. Detection of Sarcasm and Nastiness: New Resources for Spanish Language. *Cogn. Comput.* **2018**, *10*, 1135–1151. [CrossRef]
38. Justo, R.; Torres, M.I.; Alcaide, J.M. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. In Proceedings of the Pattern Recognition and Image Analysis—8th Iberian Conference, IbPRIA 2017, Faro, Portugal, 20–23 June 2017; Lecture Notes in Computer Science; Alexandre, L.A., Sánchez, J.S., Rodrigues, J.M.F., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10255, pp. 58–68. [CrossRef]
39. deVelasco, M.; Justo, R.; López-Zorrilla, A.; Torres, M.I. Automatic Analysis of Emotions from the Voices/Speech in Spanish TV Debates. *Acta Polytech. Hung.* **2022**, *19*, 149–171. [CrossRef]
40. Panda, R.; Malheiro, R.M.; Paiva, R.P. Audio Features for Music Emotion Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
41. Latif, S.; Cuayáhuitl, H.; Pervez, F.; Shamshad, F.; Ali, H.S.; Cambria, E. A survey on deep reinforcement learning for audio-based applications. *arXiv* **2021**, arXiv:2101.00240.
42. Huang, K.; Wu, C.; Hong, Q.; Su, M.; Chen, Y. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5866–5870. [CrossRef]
43. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *arXiv* **2017**, arXiv:1706.00612.
44. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
45. Marazakis, M.; Papadakis, D.; Nikolaou, C.; Constanta, P. System-level infrastructure issues for controlled interactions among autonomous participants in electronic commerce processes. In Proceedings of the Tenth International Workshop on Database and Expert Systems Applications, DEXA 99, Florence, Italy, 3 September 1999; pp. 613–617. [CrossRef]
46. Parthasarathy, S.; Tashev, I. Convolutional Neural Network Techniques for Speech Emotion Recognition. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 121–125. [CrossRef]
47. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
48. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013.
49. Tian, L.; Moore, J.D.; Lai, C. Emotion recognition in spontaneous and acted dialogues. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 698–704.
50. Ocquaye, E.N.N.; Mao, Q.; Xue, Y.; Song, H. Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *Int. J. Intell. Syst.* **2021**, *36*, 53–71. [CrossRef]
51. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-Based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484. [CrossRef]
52. Zheng, L.; Li, Q.; Ban, H.; Liu, S. Speech emotion recognition based on convolution neural network combined with random forest. In Proceedings of the 2018 Chinese Control In addition, Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 4143–4147. [CrossRef]

53. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.
54. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
55. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093. [CrossRef]
56. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, pp. 2449–12460.
57. Peyser, C.; Mavandadi, S.; Sainath, T.N.; Apfel, J.; Pang, R.; Kumar, S. Improving tail performance of a deliberation e2e asr model using a large text corpus. *arXiv* **2020**, arXiv:2008.10491.
58. López Zorrilla, A.; Torres, M.I. A multilingual neural coaching model with enhanced long-term dialogue structure. *ACM Trans. Interact. Intell. Syst.* **2022**, *12*, 1–47. [CrossRef]
59. Boloor, A.; He, X.; Gill, C.; Vorobeychik, Y.; Zhang, X. Simple Physical Adversarial Examples against End-to-End Autonomous Driving Models. In Proceedings of the 2019 IEEE International Conference on Embedded Software and Systems (ICESS), Las Vegas, NV, USA, 2–3 June 2019; pp. 1–7. [CrossRef]
60. LeCun, Y. Generalization and network design strategies. *Connect. Perspect.* **1989**, *19*, 143–155.
61. Weng, J.; Ahuja, N.; Huang, T.S. Cresceptron: A self-organizing neural network which grows adaptively. In Proceedings of the 1992 IJCNN International Joint Conference on Neural Networks, Baltimore, MD, USA, 7–11 June 1992; Volume 1, pp. 576–581.
62. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
64. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. [CrossRef]
65. Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. In *Proceedings of the 29th ACM International Conference on Information*; Knowledge Management; Association for Computing Machinery: New York, NY, USA, 2020; pp. 105–114. [CrossRef]
66. Zubiaga, I.; Menchaca, I.; de Velasco, M.; Justo, R. Mental Health Monitoring from Speech and Language. In Proceedings of the Workshop on Speech, Music and Mind, Online, 15 September 2022; pp. 11–15. [CrossRef]
67. Patel, N.; Patel, S.; Mankad, S.H. Impact of autoencoder based compact representation on emotion detection from audio. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 867–885. [CrossRef] [PubMed]
68. Senthilkumar, N.; Karpakam, S.; Gayathri Devi, M.; Balakumaresan, R.; Dhilipkumar, P. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. *Mater. Today Proc.* **2022**, *57*, 2180–2184. [CrossRef]
69. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]
70. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
71. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 210–215.
72. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
73. Zhang, W.; Lim, B.Y. Towards Relatable Explainable AI with the Perceptual Process. *arXiv* **2022**, arXiv:2112.14005v3.
74. Das, A.; Mock, J.; Chacon, H.; Irani, F.; Golob, E.; Najafirad, P. Stuttering speech disfluency prediction using explainable attribution vectors of facial muscle movements. *arXiv* **2020**, arXiv:2010.01231.
75. Anand, A.; Negi, S.; Narendra, N. Filters Know How You Feel: Explaining Intermediate Speech Emotion Classification Representations. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 756–761.
76. Esposito, A.; Marinaro, M.; Palombo, G. Children Speech Pauses as Markers of Different Discourse Structures and Utterance Information Content. In *Proceedings of the International Conference: From Sound to Sense*; MIT: Cambridge, MA, USA, 2004.
77. Ortega Giménez, A.; Lleida Solano, E.; San Segundo Hernández, R.; Ferreiros López, J.; Hurtado Oliver, L.F.; Sanchis Arnal, E.; Torres Barañano, M.I.; Justo Blanco, R. AMIC: Affective multimedia analytics with inclusive and natural communication. *Proces. Leng. Nat.* **2018**, *61*, 147–150.
78. Calvo, R.; Kim, S. Emotions in text: Dimensional and categorical models. *Comput. Intell.* **2012**, *Early view*. [CrossRef]
79. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]

80. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
81. Letaifa, L.B.; Torres, M.I. Perceptual Borderline for Balancing Multi-Class Spontaneous Emotional Data. *IEEE Access* **2021**, *9*, 55939–55954. [CrossRef]
82. Pastor, M.; Ribas, D.; Ortega, A.; Miguel, A.; Solano, E.L. Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation. In Proceedings of the IberSPEECH 2022, Granada, Spain, 14–16 November 2022; pp. 76–80.
83. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.