IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Contrasting the Emotions identified in Spanish TV debates and in Human-Machine Interactions

Mikel deVelasco, Raquel Justo, Leila Ben Letaifa, M. Inés Torres

Speech Interactve Research Group, Universidad del País Vasco UPV/EHU
{mikel.develasco, raquel.justo, leila.benletaifa, manes.torres}@ehu.es

Abstract

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. To this end we developed a crowd annotation procedure to label the speech signal in terms of both, emotional categories and Valence-Arousal-Dominance models. The analysis of these data showed interesting findings that allowed to profile both the speakers and the task. Then, Convolutional Neural Networks were used for the automatic classification of the emotional samples in both tasks. Experimental results drew up a different human behavior in both tasks and outlined different speaker profiles.

Index Terms: emotions recognition from speech, perception, communication, human-machine interaction, crowd annotation, speech processing.

1. Introduction

Speech signal includes information about the personal characteristics of the speaker, the content of the message delivered or the language used to code it, among others [1]. The analysis of the speech also allows to estimate, to some extent, the current emotional status of the speaker [2, 3, 4], even the basal mood, or the probability to be suffering a particular mental disease [5]. However, speech may also be influenced by several other variables, such as the habits of the speaker, his personality, culture or the particular task being performed [6, 7]

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. Thus, we focus on spontaneous emotions appearing in each task that show significant differences to the six basic emotions [8] that have been many times simulated by professional actors [9, 10, 11] and recorded in the lab [12]. In fact, spontaneous emotions have been hypothesized to be extremely task dependent [2, 3, 7, 4, 6]. Further to this, emotions cannot be unambiguously identified. As a consequence, not even expert labelling procedure can lead to a ground truth for learning. As an alternative, crowd annotation implementing perception experiments has also been proposed as a way to establish the ground truth [13]. However, human perception of emotions does not usually show a high agreement. As a consequence, a certain ambiguity and uncertainty always remains, which adds an stochastic component to the emotion identification problem.

In order to verify whether actually the task plays a significant role when dealing with emotion detection, a preliminary comparison of the emotional content in two very different Spanish tasks was carried out in this research work. To this end, we chose the following set of features to be analysed: agreement in crowd annotation, perceived emotions and significance in the particular task, distribution of categories in both tasks, distribution of dimensional axes of emotions, namely Valence, Arousal and Dominance (VAD), and the representation of the categories into the 3D VAD model. An additional contribution is the comparative analysis of the results in terms of categories of the automatic classification of the samples based on Convolutional Neural Networks (CNN).

Section 2 describes the two tasks addressed as well as the annotation procedure and its outcomes. Then Section 3 develops the analysis of emotional content of the corpora and Section 4 describes the preliminary classification experiments carried out. Finally, Section 5 summarizes the concluding remarks and future work.

2. Perception of emotions

2.1. Description of the tasks

TV Debates Firstly, a data-set that gathers real human-human conversations extracted from TV debates, specifically the Spanish TV program "La 6 Noche", was selected. In this weekly broadcasted show, hot news of the week are addressed by using social and political debate panels that were led by two moderators. There is a very wide range of talk-show guests (politicians, journalists, etc.) who analyse, from their perspective, social topics. Given that the topics under discussion are usually controversial it is expected to have emotionally rich interactions. However, the participants are used to speak in public so they do not lose control of the situation and even if they might overreact sometimes, it is a real scenario, when emotions are subtle. The spontaneity in this situation makes a great difference from scenarios with acted emotions as shown in [2]. The selected programs were broadcasted during the electoral campaign of the Spanish general elections in December 2015.

Elder interaction with simulated virtual agent Empathic is a European Research & Innovation Project [1] [14, 7] that implements personalized virtual coaching interactions to promote healthy and independent aging. As a part of the project, a series of spontaneous conversations between elderly and a Wizard of OZ (WOZ) have been recorded in three languages: Spanish, French and Norwegian. WOZ's technique allows users to believe that they are communicating with a human (and not a machine) in order to make their reaction more natural [7]. The conversations are related to four main topics: leisure, nutrition, physical activity and social and family relationships [14, 7]. In this work we focused on the Spanish dialogues that were recorded by 79 speakers resulting in 7 hours and 15 minutes of audio extracted from the recordings [3].

[1] www.empathic-project.eu

10.21437/IberSPEECH.2021-11

2.2. Crowd perception

TV Debates and Spanish Virtual agent interaction data were labeled in terms of emotions using the crowd annotation technique. To begin with, we automatically extracted segments of audio that we estimated to match a clause. A clause can be defined as "a sequence of words grouped together on semantic or functional basis" [15]. Thus, we can hypothesize that the emotional status does not change inside a clause. This procedures allowed to get 4118 chunks from the TV Debate corpus and 2000 from the Virtual agent corpus. Then, all these segments were crowd annotated by native speakers. To this end, both categorical and VAD model of emotions were considered. For the categorical model we first consider the categories proposed in [16] and then we reduce and adapt the list of each of the tasks. For TV Debates task we selected a list of ten labels to be considered by annotators. Then we added three questions to annotate the perception of each of the axes of the dimensional model, namely Valence, Arousal and Dominance

Three of them are related to the arousal: Excited, Slightly excited and Neutral.Valence is annotated as Positive or Slightly positive or Neutral or Slightly negative or Negative in TV Debates. For Virtual Agent task, valence is assigned one of only three labels that are Positive, Neutral and Negative for valence. The dominance labels are: Rather dominant / controlling the situation, Rather intimidated / defensive, and Neither dominant nor intimidated.. The whole questionnaire is reported in [2]. For Virtual Agent task we also selected list of ten categories adapted to the task that differs from the previous one. As an example *Sad* was only included in this task whereas *Annoyed* was only proposed to TV Debates annotators.

Annotators agreement Each audio segment was annotated by 5 different annotators. Table 1 shows the statistics of agreement per audio chunk for the categorical model. This table shows that for about 70% of the data and in both tasks, the agreement is 3/5 or 2/5. This confirms the ambiguity and subjectivity of the task. Moreover the Krippendorff's $alpha$ coefficient was also low for both tasks resulting in 0.11 and 0.13 values respectively. This coefficient reflects the agreement degree but is very dependent on the number of labels, which was high and sometimes difficult to be perceived.

In the rest of the document, we do not consider samples with agreement below 0.6, which means we have used the 64.13% of the corpus for the TV debates task and the 66.20% of the Virtual Agent task.

Table 1: *Statistics of the agreement per audio chunk*

	TV Debates		**Virtual Agent**	
Agr	No. audios	% audios	No. audios	%. audios
5/5	197	4.72%	149	7.45%
4/5	799	19.40%	421	21.05%
3/5	1645	39.95%	754	37.7%
2/5	1431	34.75%	636	31.8%
1/5	46	1.18%	40	2%
Tot. audios	4118		2000	

Annotation labels The defined sets of labels were then reduced by merging overlapping categories that we selected for the tag pairs with high level of confusion among them in the annotation procedure. Then, a minimum agreement of 0.6 (3/5) was requested for each sample as a well as a minimum number of samples. Table 2 shows the resulting list of categories considered for each task along with the percentage of samples. This Table shows that different categories appear in each corpus. Some of them could be equivalent, such as *Calm/Indiferent* and *Calm/relaxed* but *annoyed/tense* does not appear in Virtual Agent task whereas puzzled is not in the list for TV debates.

Table 2 also shows that both data-sets are imbalanced, being the *Calm* category the majority class with around 75% of the samples. This reflects the spontaneous nature of the data. There are more positive emotions in the Virtual Agent annotations and more negative emotions in TV Debates. This difference comes from the tasks characteristics. During political debates, people try to convince or even impose their opinions on other interlocutors. However, during the coaching sessions, people speak with a machine. They are quiet and paying attention to the answers to their expectations.

For the dimensional model we got a set of scale values for each axe. For for Arousal we proposed Neutral, Slightly excited and Excited in both databases. For Dominance we proposed Rather intimidated / defensive, Neither dominant or intimidated, Rather dominant / controlling the situation fro both databases. For Valence we got Negative, Slightly Negative, Neither negative or positive, Slightly Positive and Positive for TV Debates whereas we reduced the scale to Rather Negative, Neither negative or positive, Rather Positive for the Virtual Agent task. These dimensions are considered in Section 3

Table 2: *Categories more frequent in the corpora*

TV Debates		**Virtual Agent**	
Category	% audios	Category	% audios
Calm/Indiferent	73.64	Calm/Relaxed.	78.32
Annoyed/Tense	14.32	Happy/Pleased	8.76
Enthusiast	4.72	Interested	5.66
Satisfied	3.23	Puzzled	2.95
Worried	2.12		
Interested.	1.57		
Others	0.40	Others	4.31

3. Analysis of emotions

Figure 1 shows the probability density function of each variable (Valence, Arousal, Dominance) of VAD model that has been obtained by a Gaussian kernel density estimator (upper row). Figure 1 also shows different 2D projections of sample distribution in the 3D space (row below), representing each scenario in a different colour. When regarding Arousal, Virtual Agent seems to work in a very neutral scenario where excitement is almost absent. In TV debates, although neutrality is also predominant, some excitement is perceived, due to the debate nature of the conversations. Valence distribution shows a clear deviation towards positive values when considering Virtual Agent scenario, a sign of the good acceptance of the system among the users, whereas in TV debates neutrality is predominant with only a slight nuance towards positiveness. On the contrary Dominance is shifted towards Dominant values, in TV debates, but keeps

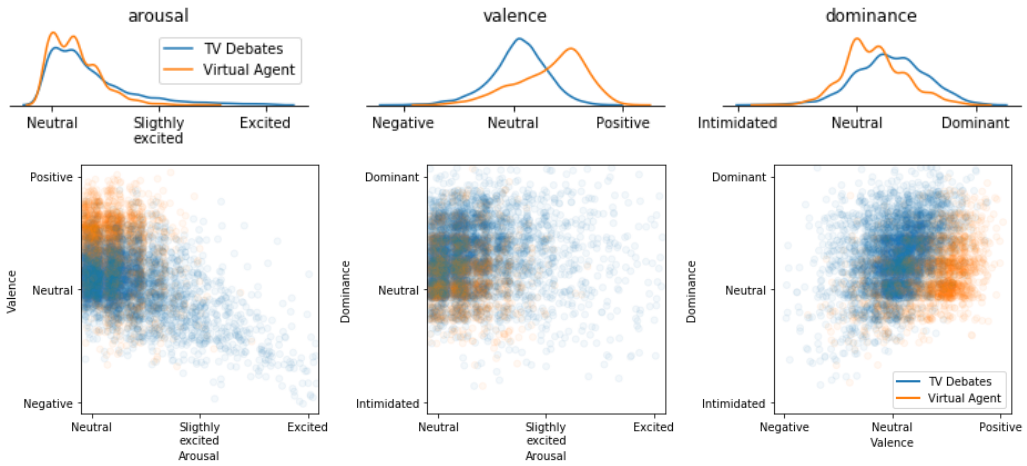

Figure 1: *VAD representation.*

neutral when users interact with the Virtual Agent. These results correlate well with the kind of audios we are dealing with in the two scenarios. In TV debates, people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally they appear to be neutral when communicating their opinions (valence tends to be neutral or slightly positive). In the Virtual Agent scenario the users are volunteers, with a good predisposition, and thus they seem to be pleased with the system (Positive Valence values). They are relaxed talking to the agent (levels of excitement tend to neutrality) and although they do not have to convince anyone they know well what they are talking about and are not intimidated (dominance values are around neutrality with a slight shift to the right).

The categorical model is also considered in this work and each category is represented in the 3D VAD space for comparison purposes. Specifically, the average of the Valence, Arousal and Dominance values of all the audios labeled within a specific category was computed and the resulting value was represented as a point in the 3D space. Figure 2 shows 2D projection of the resulting representation. If we focus on TV Debates, it can be noticed that *Interested* and *Worried*, the least representative categories, according to Table 2, are very close to the category with the highest number of samples, *Calm/Indifferent*, in all the 2D projections (purple, orange and deep blue points), so they were merged in an only one category. The same happens with Enthusiastic and Satisfied (light blue and green points). When considering Virtual Agent scenario although the category *Calm/Relaxed* is the most relevant one with more than the 75% of the samples we decided to keep the remaining categories because the fusion is not as clear as in the previous case, as shown in Figure 2. Thus the final set of categories used for the classification experiments reported in this work (Sectiion **??**) is the following one for TV Debates: 1) *Annoyed/Tense*, 2) *Enthusiastic + Satisfied*, 3) *Calm/Indifferent + Interested + Worried* and for as the Virtual Agent the list is: 1) *Calm/Relaxed*, 2) *Happy/Pleased*, 3) *Interested*, 4) *Puzzled*.

As shown above, there are some categories that are not in both sets due to the nature of the different tasks, like Annoyed/Tense that is only In TV Debates or Puzzled that only appears in the interaction with the Virtual Agent. Moreover, Figure 2 shows that there is not any point in Virtual Agent scenario around the location of the red point (Annoyed/Tense) of TV Debates (higher excitement levels and negative values of Valence), which is in fact quite separated from the other categories. The same happens with *Puzzled* represented by the brown point (low levels of Valence and Dominance) that has not any representation in TV Debates and it is a bit separated from the other categories in Virtual Agent scenario. This correlates well with the idea that people interacting with the Virtual Agent are not in general annoyed or tense, while this is a quite common feeling in a debate. Furthermore, speakers in the debates do not usually show that they are in an unexpected situation, since it can be interpreted as a weak point, while it is quite easy to imagine it in the interaction with a machine. There are also categories, like *Calm* that has a similar location in both scenarios but with higher values of Valence for Virtual Agent interactions. That is, the users interacting with the Virtual Agent perceived as calm tend to be more positive than the ones in TV Debates. The same happens with *Enthusiastic + Satisfied* from TV Debates and *Happy/Pleased* from Virtual Agent, that although they are very close in their location in both scenarios (with a very similar meaning) *Happy/Pleased* seems to have more positive Valence values than *Enthusiastic + Satisfied*, but a bit lower Dominance and Arousal values.

4. Experiments and results

To complete the work, some classification problems were carried out in both tasks described in Section 2.1. For TV Debates, 4118 chunks were selected distributed in the 3 classes mentioned avobe (*Annoyed/Tense*, *Enthusiastic + Satisfied*, and *Calm/Indifferent + Interested + Worried*) and for the Virtual Agent, 2000 samples were selected divided into 4 classes (*Calm/Relaxed*, *Happy/Pleased*, *Interested*, and *Puzzled*).

One of the challenges of both data-sets is the different length of each audio sample. Some kind of Neural Networks are specifically well suited to deal with this problem and given that deep learning is the state of the art in many AI areas, including emotion recognition, a Convolutional Neural Network architecture was designed for this work. Let us note that in [17] a neural network architecture provided promising results when comparing ot to classical Support Vector Machines, for a regression problem over the task related to TV debates.

The number of samples in both data-sets are also a challenge. It makes nonsense to try to identify the emotions from

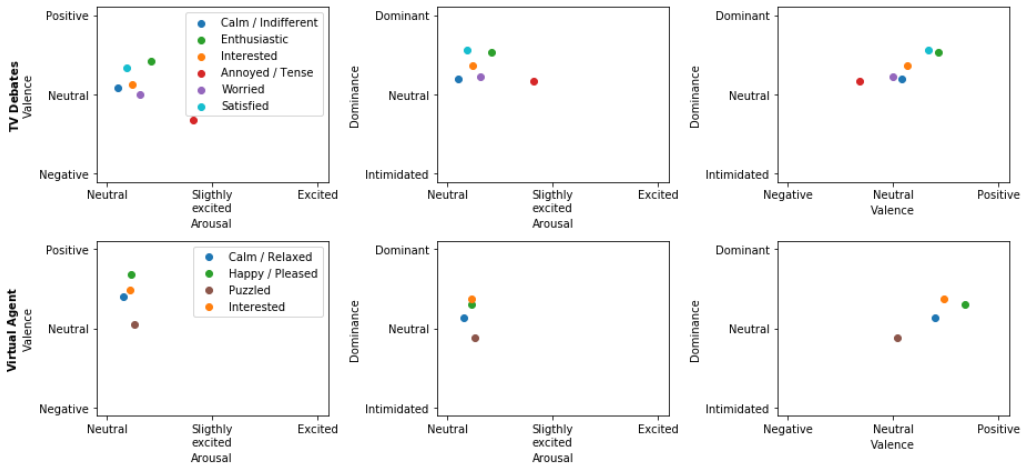

Figure 2: *Categories in dimensional representation.*

raw-audio. Different works suggest that there is not a standard audio feature-set that works well for all emotion recognition corpora [18, 19, 20]. In this context, we decided to use the audio Mel-frequency spectrogram as the classifier's input. It is known that the spectrogram encodes almost all audio information and should be possible to identify from that.

Figure 3 shows the architecture of the network used in this work. It takes the mel-spectrogram input and reduce both mel-frequency and time dimensions using 2D convolutions and max-poolings (red boxes). This sub-network reduces time dimension but creates richer audio representation. Then, the network takes the new representation and try to classify each time step. After classifying all time steps, the network averages it in order to provide an output for the input audio.

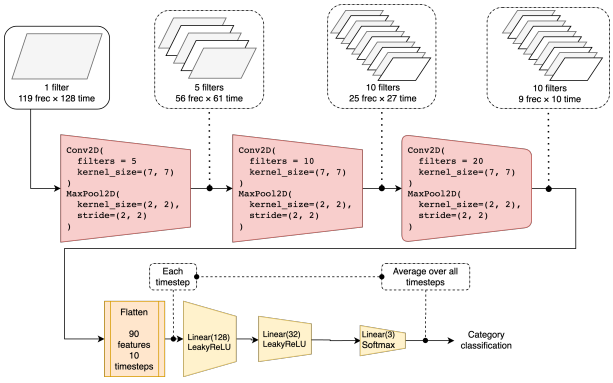

Figure 3: *Architecture of the Network used*

In the training process, several decisions were chosen. On the one hand, the network will only see a sub-part of the full audio. Thus, the training process is easier if all the batches work with the same input length, which can be considered as a dropout mechanism. On the other hand, an repetition over-sampling method was chosen, where all the non-majority class samples were provided 5 times. It helps the network to avoid the exclusive prediction of the majority class. Adam optimizer is used with a learning rate of 0.001 and 150 epochs were training on all database. These experiments were carried out over a 10-fold cross-validation procedure.

Classification results are given in Table 3. Most promising results come from TV Debates, in fact, the model guesses 72% of the test samples, and achieves a F1 Score of 0.59, that can be considered a good result taking into account the ambiguity and subjectivity of the task.

As expected, the category *Calm/Indifferent + Interested + Worried* got better results since it is the majority class with a F1 Score of 0.82.In contrast, *Annoyed/Tense* and *Enthusiastic + Satisfied* perform a little bit worse, with a 0.56 and 0.43 in F1 Score.

Table 3: *Evaluation of the classification results for the categorical model*

TV Debates				**Virtual Agent**			
Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
72%	0.56	0.66	0.59	74%	0.32	0.27	0.27

Nevertheless, Virtual Agent experiments obtained lower results. Table 3 show a very high accuracy (74% of the samples) along with low values of F1, precision and recall values. The majority class, i.e.*Calm*, achieved an F1 score of 0.88 whereas all minority classes remain under 0.32 fro F1. This is mainly due to the huge imbalance of this data-set along with the very reduced number of samples.

5. Conclusions

This work provides a comparison of the emotional content in two different Spanish corpora dealing with very different tasks. The emotional labels, associated to spontaneous emotions, were achieved by means of perception experiments using crowd annotation. The agreement among the annotators was considered to build the ground truth. The analysis carried out shows the main differences associated to each task, in terms of both, the emotional category distribution and the level of Valence, Arousal and Dominance and brings out the relevance of the task when addressing an emotion recognition problems. This analysis also highlights that the perception experiments carried out were able to outline a different speaker profile for each of the tasks. Thus, crowd annotation seems to be valid approach for emotions. Finally, some preliminary classification experiments

were also conducted showing very promising results for TV Debate task whereas the Virtual Agent task needs more samples and a more sophisticated oversampling method. Future work includes a deeper and interrelated analysis of the data as well getting a higher number of annotated samples for the Virtual Agent classification task.

6. Acknowledgements

The research presented in this paper is conducted as part of the AMIC and EMPATHIC projects project that have received funding from the Spanish Minister of Science under grant TIN2017-85854-C4-3-R and from the European Union's Horizon 2020 research and innovation program under grant agreement No 769872. First author has also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

7. References

[1] A. López-Zorrilla, N. Dugan, M. Torres, C. Glackin, G. Chollet, and N. Cannings, "Some asr experiments using deep neural networks on spanish databases," in *IberSpeech*, Lisbon, 2016, pp. 149–158.

[2] M. deVelasco, R. Justo, A. López-Zorrilla, and M. Torres, "Can spontaneous emotions be detected from speech on tv political debates?" in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*, Naples, 2019.

[3] L. B. Letaifa, M. I. Torres, and R. Justo, "Adding dimensional features for emotion recognition on speech," in *IEEE International Conference on Advanced Technologies for Signal and Image Processing*, Tunisia, 2020.

[4] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech*, 2018.

[5] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," 2020.

[6] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech Language*, vol. 53, pp. 156 – 180, 2019.

[7] R. Justo, L. B. Letaifa, C. Palmero, E. G. Fraile, A. Johansen, A. Vazquez, G. Cordasco, S. Schlogl, B. F. Ruanova, M. Silva, S. Escalera, M. D. Velasco, J. T. Laranga, A. Esposito, M. Kornes, and M. I. Torres, "Analysis of the interaction between elderly people and a simulated virtual coach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 6125–6140, 2020.

[8] P. A. Davidson. R. J., Ekman, *Nature of emotion: Fundamental questions*, ser. Oxford University Press, P. E. . R. J. Davidson, Ed. New York: Springer, 1994.

[9] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions, IJSE*, pp. 68–99, 2010.

[10] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[11] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: How does an automated system compare to naive human coders?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2274–2278.

[12] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition."

[13] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English," in *Proc. Interspeech 2019*, 2019, pp. 316–320. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1413

[14] M. I. Torres, J. M. Olaso, C. Montenegro, R. Santana, A. Vázquez, R. Justo, J. A. Lozano, S. Schlögl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. P. Cantariño, O. Deroo, O. Gordeeva, J. Tenorio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto, "The empathic project: Mid-term achievements," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 629–638.

[15] A. Esposito, V. Stejskal, and Z. Smékal, "Cognitive role of speech pauses and algorithmic considerations for their processing," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, pp. 1073–1088, 2008.

[16] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, pp. E7900 – E7909, 2017.

[17] M. de Velasco, R. Justo, J. Antón, M. Carrilero, and M. I. Torres, "Emotion detection from speech and text," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 68–71. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-15

[18] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 698–704.

[19] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," 2017.

[20] S. Parthasarathy and I. Tashev, "Convolutional neural network techniques for speech emotion recognition," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 121–125.