Can Spontaneous Emotions be Detected from Speech on TV Political Debates?

Mikel deVelasco Vázquez
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: mikel.develasco@ehu.eus

Raquel Justo
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: raquel.justo@ehu.eus

Asier López Zorrilla
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: asier.lopezz@ehu.eus

María Inés Torres
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: manes.torres@ehu.eus

***Abstract*—Decoding emotional states from multimodal signals is an increasingly active domain, within the framework of affective computing, which aims to a better understanding of Human-Human Communication as well as to improve Human-Computer Interaction. But the automatic recognition of spontaneous emotions from speech is a very complex task due to the lack of a certainty of the speaker states as well as to the difficulty to identify a variety of emotions in real scenarios.**
In this work we explore the extent to which emotional states can be decoded from speech signals extracted from TV political debates. The labelling procedure was supported by perception experiments where only a small set of emotions has been identified. In addition, some scaled judgements of valence, arousal and dominance were also provided. In this framework the paper shows meaningful comparisons between both, the dimensional and the categorical models of emotions, which is a new contribution when dealing with spontaneous emotions. To this end Support Vector Machines (SVM) as well as Feedforward Neural Networks (FNN) have been proposed to develop classifiers and predictors. The experimental evaluation over a Spanish corpus has shown the ability of both models to be identified in speech segments by the proposed artificial systems.

Index Terms: speech processing, emotion detection from Speech, human-AI, affective computing

I. Introduction and Context

During the last years the Scientific Community has shown an increasing interest in affective computing and its potential capability to change the way in which Human-Machine interaction is carried out by getting a better understanding of Human-Human Communication. This is an artificial system able to analyze the intra-cognitive communication [1][2][3] between humans in order to develop ICT applications aimed to cooperate in Human-Machine communication. As a consequence, this is a good example of Cognitive infocommunications [2], that deals with the idea of cognitive processes and ICT applications working together in order to take benefit of each other and go beyond their isolated capabilities [3].

One of the goals of affective computing is the study and development of systems that can detect emotions from multimodal signals. In this work we deal with video recordings but focusing on speech, since it is inseparably intertwined with the emotional status during the cognitive process in human communication. Furthermore, it seems to be a good indicator of depression [4], very related to the emotional status, or even parkinson disease [5].

Most of the research on the identification of emotional features from video recordings has been carried out with a reduced set of acted emotions [6] [7] [8]. To this end the basic set of emotions defined by Eckman [9] has been broadly used, mainly for facial expressions. At this point it is important to note that the choice of acted emotions was just based on the easiness to get them rather than supported by any hypothesis [8]. In contrast, current research focuses more on the identification of emotions in scenarios that implements realistic tasks [10], which is the framework of our research.

Nevertheless, the automatic recognition of spontaneous emotions from speech is a very complex task. To begin with, the intensity of spontaneous emotions is generally lower than the one of acted emotions resulting in a smaller emotional space where emotions are closer [11]. In addition, researchers of emotions have established that ordinary communication involves a variety of complex feeling states that cannot be characterized by a reduced set of categories, which does not cover the wide range of affect states [10]. Therefore a number of researchers [12][10] [8] propose a dimensional [13] representation where each affect state is represented by a point in a two-dimensional space, namely valence and arousal, which some authors extend to three by also considering dominance.

Another important drawback is that the surface realizations of the underlying spontaneous emotions are different to those associated to acted emotions [11] [14], which complicates the direct application of the results of the investigations carried out with acted emotions as well as the use of acted data for training purposes. Furthermore, the set of emotions that appear in each specific real scenario is very task dependent and, thus, also the related automatic detection is. For example, the goal may just be to recognize anger through a simple anger/no anger classification in call centers [15] or to identify annoyance activation levels [16] [17] in customer assistance calls.

An additional weak point of spontaneous emotions is the labelling procedure, since the current emotion of the speaker

cannot be unequivocally established. In fact, the emotional label assigned by a speaker to his own utterance might differ to the one assigned by a listener to the same utterance, being the first one closer to the current emotion [11]. However the speaker self annotation is not usually a realistic approach. As a consequence, the annotation of utterances in terms of spontaneous emotions is generally carried out through perception experiments, which are based on the particular judgement of every single annotator. Therefore, the disagreement amount annotators as well as the distance between the emotion expressed and the emotion perceived can be significant. In contrast, if emotions are expressed by professional actors, or just elicited, then the annotation procedure is not required [18]. Thus, the generated emotion is always labelled by the intent of the actor.

The previous framework shows spontaneous emotions generated and perceived to be very dependent of a variety of factors that make every data analysis and every automatic recognition task challenging and difficult for comparison. In this context, the main goal of this work is the analysis of the emotional content of speech produced by journalists and politicians on TV political debates. In summary, the problem addressed in this work is the analysis of the intra-cognitive communication between politicians and journalists during political debates. Additional contributions are the data annotation procedure and analysis as well as some baselines results of automatic detection of spontaneous emotions from the speech. The analysis carried out through different feature sets can be seen as an artificial cognitive capability that actually measures a human cognitive load, as defined in [16]. Section II describes the annotation procedure and data analysis in terms of both, categorical and dimensional models. Then some experiments aimed at the automatic detection of spontaneous emotions are shown in Section III. Finally Section IV present some concluding remarks.

II. Data Analysis

The specific task we are dealing with is described in this section. Then we summarize the labelling procedure along with the analysis of the annotated data.

A. Task

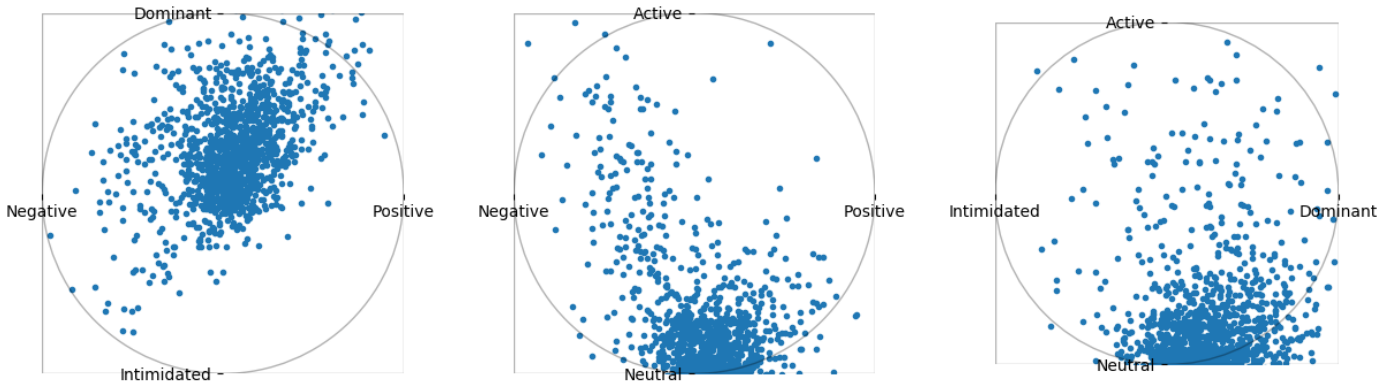
In contrast with acted emotions, spontaneous emotions show a high dependency of the specific environment in which they appear. As a consequence, the design of an appropriate corpus to develop automatic detectors of emotions as well as the choice of the specific set of emotions of interest are very linked to the particular task. In this work, the Spanish TV program "La Sexta Noche" was considered. This is a weekly broadcasted program in which a set of journalists and politicians talk about current issues. It is held as a round table discussion with a moderator that guides and conducts the debate. The program usually includes controversial topics to be discussed, so that emotional content could be expected. However, the participants are often used to speak in public, also to join TV debates, so it is not expected that they lose the control of the situation. In fact, we are in a realistic scenario in which emotions are subtle.

We first selected the broadcasts during the electoral campaign of the Spanish general elections in December 2015. Our goal was to design a corpus to train neural networks as well as to develop other data driven approaches of interest. To this end, the audios associated to the selected programs were split into smaller segments from two to five seconds. An algorithm was then designed to get audio chunks that included clauses, which are the smallest grammatical unit that can express a complete proposition. Accordingly, the algorithm uses silences and pauses, as well as the text transcriptions, to identify the compatible utterances. This procedure provided a set of 5500 audio chunks.

B. Representing Emotions by Categories and Dimensions

The audio chunks described in Section II-A, were labelled with emotional information. To this end, we considered both, the categorical model an the dimensional one, to represent the emotion associated to each chunk. The first one defines a set of discrete categories that ranks from the basic set defined by Ekman to larger sets defining more specific an realistic affect states. For this work we defined the set of categories of interest based on the selection provided in [19]. Then, it was adapted to the specific features of the task. For instance, *Sad* was not included since it is not expected to appear in political debates. The dimensional one is a psychological model that characterize affect states in terms of two or three dimensions, namely valence, arousal and dominance (VAD) [10], [20]. A crowd annotation using a crowdsourcing platform [21] was carried out to get emotional labels for both, VAD and categorical models. Each audio-clip was labeled by 5 annotators from the crowd, that were not previously trained. The goal was to pick up the diversity in people's perception in order to deal with the ambiguity associated to the interpretation of emotional information. Each annotator was asked to fill the following questionnaire for each audio:

- How do you perceive the speaker?
  - Excited
  - Slightly excited
  - Neutral
- His/her mood is:
  - Positive
  - Slightly positive
  - Neutral
  - Slightly negative
  - Negative
- How do you perceive the speaker in relation with the situation which he/she is in?
  - Rather dominant / controlling the situation
  - Rather intimidated / defensive
  - Neither dominant nor intimidated
- Select the emotion that you think describes better the speaker's mood:
  - Embarrassed
  - Bored/Tired
  - Disconcerted/Surprised
  - Angry
  - Interested
  - Satisfied/Pleased
  - Worried
  - Enthusiastic
  - Annoyed/Tense
  - Calm/Indifferent
- Quality of the audio:
  - Correct
  - Overlapping of several speakers
  - Advertisement
  - Other



(a) Valence & Dominance (b) Valence & Arousal (c) Dominance & Arousal

Fig. 1: Three projections of the data in the VAD space.

The first 3 questions are related to the VAD model and the fourth one to the categorical model. The fifth question was added to detect bad quality audios like music or overlappings. These audios were removed from the corpus.

The annotators' responses to the first three questions were used to provide a representation of each audio chunk according to the VAD model. Specifically, a conversion of the selected levels of valence, arousal and dominance into a real point in a 3D space was needed. To this end, a discrete value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited:1, Slightly excited: 0.5, Neutral: 0. Then the average value considering the 5 annotations was computed to represent each annotated chunk in the 3D space.

In order to analyse the annotated data, three projections of the points obtained in the 3D space were achieved and shown in Figures 1a, 1b and 1c, namely arousal/dominance, arousal/valence and dominance/valence. These figures show that in most of the audios speakers are neutral (not excited or not very active). Their mood is also neutral in terms of valence (no positive neither negative) but they look to be rather dominant. These results correlate well with the kind of audios we are dealing with, in which 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 with regarding their opinions (valence tends to be neutral or slightly positive).

For the categorical model, an agreement level of $>= 60\%$ was required in the annotations provided to each audio chunk to be considered. Thus, only five categories could be taken into account when the previous agreement thresholds were required, *Calm, Enthusiastic, Annoyed, Worried, Satisfied*. The rest of them were frequently mixed up with other ones so they were never associated to an audio chunk.

Then, the average of the valence and activation values of all the audios labeled with and specific category was computed. The corresponding point is represented in Figure 2 with a circle (each color represents a different category). The triangles

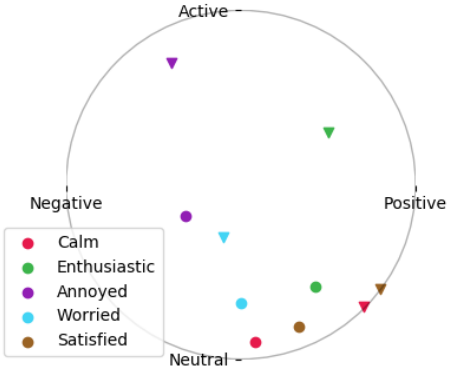


Fig. 2: Categorical average of the valence and arousal (dots) vs associated theoretical value [22], [23] (triangles).

are related to the position associated to the same category according to the map given in [22], [23]. This map shows the relationship among discrete categories and their representation in a valence/arousal space. From this figure it can be concluded that when comparing real vs. expected values the arousal is always lower in realistic emotions and the valence tends to be more neutral (more positive for annoyed and worried and more negative for calm, enthusiastic and satisfied). This means that most speakers in the proposed task tend to be more neutral than expected. In addition, their emotional expressions seem to be subtle, in accordance to spontaneous emotions and in contrast to simulated ones. Moreover, the whole space associated to realistic emotions (space occupied by circles) is smaller than the one related to the expected ones (space occupied by triangles).

Given that dominance is out of the previous representation the obtained dominance levels for each category were given on Figure 3. This figure shows that there are two different levels of dominance in the audios considered, a medium level for *Calm, Annoyed* and *Worried* and a higher level for *Enthusiastic* and *Satisfied*. Dominance seems to be relevant for this specific task since it is present in journalists' and politicians' speech and

it is worth considering it. Moreover, it seems to be higher for positive emotions like *enthusiastic* and a bit lower for negative ones like *Annoyed*.

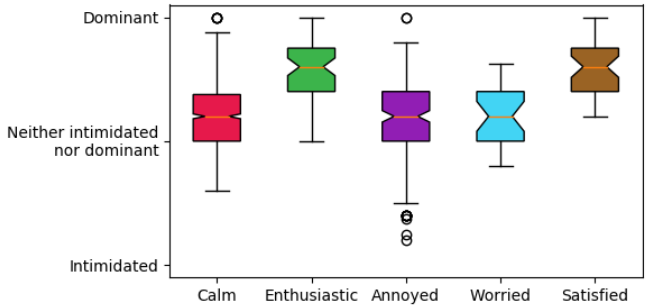


Fig. 3: Box-plot representation of most frequent categories over dominance dimension.

Finally, a confusion matrix with the number of audio chunks that follow a pattern of 3-2/2-3 annotations (3 annotations c_i/c_j and 2 annotations c_j/c_i) was built. From this matrix it was concluded that Annoyed and Worried are mixed up frequently and the same happens for Enthusiastic and Satisfied. This fact is also reflected in Figure 3, where their distributions are overlapped for dominance. Therefore, they were mixed and for the categorization experiments only three different categories were considered.

III. Exploring automatic detection of spontaneous emotions from speech

Two different and broadly used supervised learning paradigms were employed to carry out the automatic detection of spontaneous emotions from speech: Support-Vector Machines (SVM) [14] and Feedforward Neural Networks (FNN)[24], [25]. These models are capable of both classification and regression. Thus, they were used for classifying speech audios according to both the categorical and the dimensional VAD model, explained at Section II-B.

Our corpus consists in variable-length segments of speech, but SVMs can only process fixed-length inputs. This problem was overcome using the average and standard deviation of each acoustic feature over the whole audio chunk as input. Even though the FNN models can process variable-length audios using recurrent or convolutional layers, during our experiments only fully-connected layers have been used, in order to compare SVMs with FNNs fairly.

Six set of features were explored in this work according to previous experiments carried out with larger sets [26][27]:

- Set A: Pitch and Energy.
- Set B: Pitch, Energy and Spectral Centroid.
- Set C: Pitch, Energy, Spectral Centroid, ZCR and Spectral Spread.
- Set D: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 12 MFCC coefficients.
- Set E: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 16 LPC coefficients.
- Set F: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 21 Bark features.

A. Experimental set up

Four kind of SVMs and FNNs were trained, with little variation in their hyper-parameters. The first was a classifier that will predict an emotion from the categorical model. Each of the three remaining models were regressors devoted to predict the value associated to each of the three axis of the VAD model: valence, arousal and dominance.

All the SVMs for both categorical and dimensional models used the Radial Basis Function (RBF) kernel. Then, SVMs were trained until convergence. Furthermore, two-layered multilayer perceptrons FNN, with sigmoidal activation functions were also trained. However, FFN used different output layers and loss functions in the classification and regression tasks. A softmax activation function was selected for categorical classification whereas a sigmoidal activation function was applied to deal with the dimensional regressions. Then, the cross-entropy loss was employed for categorical classification, and the batch-level coefficient of determination (R^2) as the regression loss function. In addition, this coefficient was also proposed as evaluation metric in the test partition.

B. Experimental results

During a first series of experiments, the classifiers performed poorly and tended to predict the majority class, as shown in the first two rows of Table I. In order to avoid this problem, an oversampling method was employed (second row of Table I) to equalize the number of examples per category. The oversampling technique improves most of the results regardless of the set and the model used.

Table I shows that FNN models obtained slightly more accurate results than SVMs. The best FNN model was obtained using the oversampling method trained with the Set B of features.

		Set A	Set B	Set C	Set D	Set E	Set F
Without oversampling	Net	0.283	0.283	0.283	0.350	0.283	0.283
	SVM	0.336	0.283	0.324	0.279	0.281	0.283
With oversampling	Net	0.243	**0.400**	0.238	0.339	0.271	0.358
	SVM	0.284	0.290	0.221	0.280	0.287	0.214

TABLE I: Macro F1 Score on categorical experiments.

Independent models for each axis of the dimensional model were trained, because different sets of features might extract better the information required to build accurate models for each dimension. To this end, the same 6 set of features were used along with both FNN and SVM paradigms.

Two different metrics are proposed for evaluation purposes. The first and more frequently used in the literature is the mean squared error (MSE). However MSE scores as good models those models which predict the mean of the training data. Alternatively, we found out that R^2 (R squared), also known as the coefficient of determination, was instead a better metric. In fact, R^2 punishes harder regressors, which always predict the same value. Table II shows that those sets with low MSE (supposed to be best models) are not the ones with the highest R^2 score, which scores the strength of the relationship between the predictors and response.

Set	Model	Valence		Arousal		Dominance	
		MSE	R^2	MSE	R^2	MSE	R^2
A	Net	0.019	0.172	0.011	0.003	0.018	0.085
	SVM	0.020	0.111	0.011	0.016	0.018	0.089
B	Net	0.023	0.177	0.011	-0.021	0.019	0.033
	SVM	0.025	0.104	0.010	0.035	0.017	0.074
C	Net	**0.016**	0.086	0.010	0.005	**0.016**	0.088
	SVM	0.019	0.050	0.010	0.045	**0.016**	0.103
D	Net	0.020	0.275	0.012	**0.116**	0.018	0.095
	SVM	0.023	0.160	0.013	0.065	0.020	0.073
E	Net	0.022	**0.357**	**0.008**	0.095	0.019	0.069
	SVM	0.028	0.202	**0.008**	0.082	0.018	**0.119**
F	Net	0.022	0.330	0.012	0.077	0.018	0.070
	SVM	0.018	0.215	0.011	0.086	0.018	0.081

TABLE II: Results of 3 dimensional models tested with MSE and R^2 Score.

Table II shows that sets D, E and F achieve better results than sets A, B and C. This can be explained due to the information given by LPC, Bark and MFCC features, that provide similar information. This table also shows significantly better R^2 scores for valence whereas the ones for arousal looks lowers. These founds match the expectations according to Figure 1.

There is not a learning paradigm that seems to fit best the three dimensions. While the feed-forward nets are more suitable methodologies in the valence and arousal dimensions, SVM regressions seem to be more suitable for the dominance. The best feature set seems to be Set E, with the exception of the arousal dimension, where Set D is the best performing.

IV. Concluding remarks

In this work we have analyzed the emotional content of speech produced by journalists and politicians on TV political debates. Spontaneous emotions have been labelled through perception experiments carried out over a crowdsourcing platform. These experiments reported a very reduced emotional map for this task where only a few emotions clearly appeared. The dimensional model showed distributions around neutral values, including some positive tendency towards dominance and positive valence. The dimension averages of identified categories define a reduced dimensional map where spontaneous emotions occupy a short space towards low values of arousal. The work has also explored the automatic detection of the spontaneous emotions for this task. The experiments carried out did not result in impressive accuracies, which matches the outcomes of the data analysis and also due to the fixed-length strong constraint for FNN. Better results are expected using both recurrent and convolutional networks due to the fact that some information is lost just after computing the average and standard deviation of each acoustic feature over the whole audio. Finally, it has to be outlined that R^2 score seems to be a more accurate evaluation metric for these tasks than the broadly used MSE, which provides very optimistic results when samples are close. This system develop an artificial cognitive system that represents the human decision process of the annotators when analyzing the intra-cognitive communication between politicians and journalists in political debates.

V. Acknowledgements

This work has been partially funded by the Spanish Government under grant TIN2017-85854-C4-3-R (AEI/FEDER, UE) and conducted in the project EMPATHIC (Grant n 769872) funded by the European Union's H2020 research and innovation program.

References

[1] P. Baranyi and A. Csapó, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.

[2] P. Baranyi, A. Csapó, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer International, 2015.

[3] P. Baranyi, "Special issue on cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 15, no. 5, pp. 7–10, 2018.

[4] K. Gbor and K. Vicsi, "Comparison of read and spontaneous speech in case of automatic detection of depression," in *8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 09 2017, pp. 213–218.

[5] D. Sztah, M. G. Tulics, K. Vicsi, and I. Vallik, "Automatic estimation of severity of parkinson's disease based on speech rhythm related features," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017, pp. 000 011–000 016.

[6] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 371–384, Oct 2015.

[7] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, March 2016, pp. 2274–2278.

[8] 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.

[9] P. Ekman, *Handbook of Cognition and Emotion*. Sussex, U.K.: John Wiley and Sons, Ltd., 1999, ch. Basic Emotions.

[10] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010. [Online]. Available: http://dx.doi.org/10.4018/jse.2010101605

[11] R. Chakraboty, M. Pandharipande, and S. K. Kopparapu, *Analyzing Emotions in Spontaneous Speech*. Singapore: Springer Nature, 2017.

[12] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH*, 2008.

[13] J. A. Russel, "A circumflex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 116–1178, 1980.

[14] 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 and Language*, vol. 53, pp. 156 – 180, 2019.

[15] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Anger detection in call center dialogues," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Oct 2015, pp. 139–144.

[16] J. Irastorza and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in *Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on*. IEEE, 2016, pp. 000 103–000 106.

[17] J. Irastorza and M. Torres, *Cognitive Infocommunications, Theory and Applications. Topics in Intelligent Engineering and Informatics*. Springer, 2019, ch. Tracking the Expression ofAnnoyance in Call Centers.

[18] T. Bnziger, M. Mortillaro, and K. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 156 – 180, 2012.

[19] 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 of the United States of America*, vol. 114, no. 38, p. E7900E7909, September 2017.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 2014, pp. 3–10.

[21] R. Justo, J. M. Alcaide, and M. I. Torres, "Crowdzientzia: Crowdsourcing for research and development," in *Proceedings of IberSpeech*, November 2016, pp. 403–410.

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

[23] J. A. Russell, "Pancultural aspects of the human conceptual organization of emotions," *Journal of Personality and Social Psychology*, vol. 45, pp. 1281–1288, 12 1983.

[24] Z. Zhang, J. Han, E. Coutinho, and B. W. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *CoRR*, vol. abs/1810.05507, 2018.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5200–5204.

[26] M. de Velasco, R. Justo, J. Antn, M. Carrilero, and M. I. Torres, "Emotion Detection from Speech and Text," in *Proc. IberSPEECH 2018*, 2018, pp. 68–71.

[27] A. Lopez-Zorrilla, M. deVelasco Vazquez, S. Cenceschi, and M. Ines Torres, "Corrective focus detection in italian speech using neural networks," *ACTA POLYTECHNICA HUNGARICA*, vol. 15, no. 5, pp. 109–127, 2018.