L. B. Letaifa, M. I. Torres and R. Justo, "Adding dimensional features for emotion recognition on speech," **2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)**, Sousse, Tunisia, 2020, https://doi.org/10.1109/atsip49331.2020.9231766

Adding dimensional features for emotion recognition on speech

Leila Ben Letaifa
University of the Basque Country
Spain
leila.benletaifa@ehu.eus

Maria Inés Torres
University of the Basque Country
Spain
manes.torres@ehu.eus

Raquel Justo
University of the Basque Country
Spain
raquel.justo@ehu.eus

***Abstract*—Developing accurate emotion recognition systems requires extracting suitable features of these emotions. In this paper, we propose an original approach of parameters extraction based on the strong, theoretical and empirical, correlation between the emotion categories and the dimensional emotions parameters. More precisely, acoustic features and dimensional emotion parameters are combined for better speech emotion characterisation. The procedure consists in developing arousal and valence models by regression on the training data and estimating, by classification, their values in the test data. Hence, when classifying an unknown sample into emotion categories, these estimations could be integrated into the feature vectors. It is noted that the results using this new set of parameters show a significant improvement of the speech emotion recognition performance.**

***Index Terms*—Speech emotion recognition, dimensional parameters, feature extraction.**

I. Introduction

Despite the great progress made in speech recognition, we are still far from having a natural interaction between man and machine because the machine does not understand the emotional state of the speaker [2]. Hence recently, researchers have been attempting to automatically recognize the speaker emotions and make information technology more accessible and credible for human users [3] [5].
An important issue in speech emotion recognition is the extraction of efficient speech features that characterize the emotional content of speech. Inspired by the source filter model of the the physical speech production, researchers have used a large number of acoustic parameters [1] [3] [4] including parameters in the time, frequency, amplitude, and spectral distribution domains [3]. In order to reduce the dimension of the large sets of the descriptors, two kinds of practise are widespread [4] : projecting features in a reduced dimension space (i.e. linear discriminant analysis or principal component analysis) and selection of a subset of discriminant features (i.e. Fisher selection algorithm or genetic algorithm). Recently a new "minimalistic" acoustic feature set is proposed in [3] [6] in order to make results between the different emotion recognition systems comparable.

Correlations between acoustic parameters and, categorical emotions [1] [8] in one hand, and dimensional parameters [7] [1] in the other hand have been investigated and agreed. Hence, state of the art emotion recognition systems extract acoustic features to identify an emotional state either in a discrete domain (i.e. emotional categories) or in a continuous domain (by assigning it values in the 3 dimensional arousal-valence-dominance space).
Correlation between categorical emotions and dimensional parameters is also assessed [10] and recent studies [9] has shown the usefulness of dimensional parameters ground-truth in emotion detection. So in this paper, we explore the contribution of dimensional parameters estimation to automatic speech classification into emotional categories. It is important to highlight that the proposed parametrisation is rather based on the human perception of emotions and not only in the speech production model.

The remaining of the document is organized as follows: in Section 2 we introduce the proposed architecture of valence/arousal estimation and we describe how the obtained parameters are integrated into the emotion recognition system. Next, the database annotations analysis are described with a particular attention to dimensional parameters relevance in Section 3. Finally we provide and discuss results of regression and emotion recognition using a baseline system and the proposed method in Section 4 before concluding our study in Section 5.

II. Proposed architecture

The main contribution of this paper concerns the automatic insertion of the Valence-Arousal-Dominance (VAD) model parameters in the characteristic's vector in order to increase the emotion recognition system performance. The procedure is carried out in three steps:

- Regression (Fig 1): the annotated train data is used to develop arousal and valence regression models.
- Estimation (Fig 2): the automatic classifier uses the previous models to estimate the regression variables (i.e. arousal and valence) for the test data.
- Classification (Fig 3): the estimated values are concatenated with the acoustic features in order to perform emotion recognition.

III. Database description

The corpus consists of 134 sessions of 10 minutes each spontaneous spoken interactions between Spanish elderly people and a simulated virtual coach. Eighty participants talked

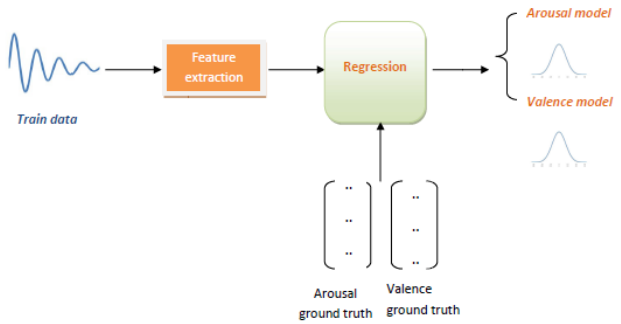


Fig. 1. Arousal and valence models estimation by regression

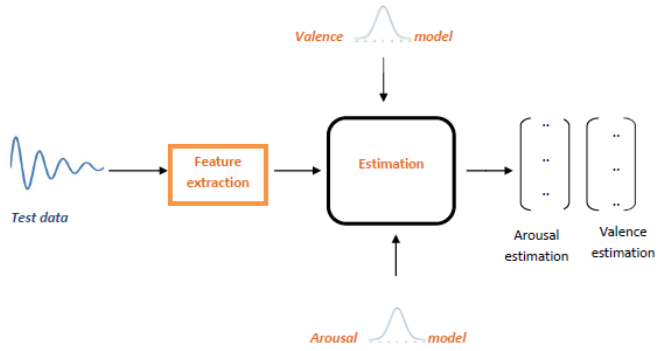


Fig. 2. Estimation of arousal/valence values in the test data

to a visual agent in two different scenarios, a general one related to leisure preferences and another one developing a nutrition coaching session [5] [9]. In each session both audio and video recordings were stored. Here only the audio part of the conversations (which duration is about 7 hours) is considered.The audio files were then labelled in terms of emotions. To this end, a manual annotation from scratch was carried out by three native collaborators. The perceived emotion was labelled in terms of both the categorical and the three-dimensional (arousal, valence and dominance) models. The three-dimensional model parameters are assigned numerical labels among -1, 0 , 1.

Thus, each emotion is described by four parameters:

- its category : calm, sad, happy, puzzled and tense,
- its arousal level : excited (1), slightly excited (0) and neutral (-1),
- its polarity: positive (1), not positive and not negative (0) and negative (-1)
- its dominance level : rather dominant (1), neither dominant nor intimidated (0) and rather intimidated (-1).

The time limits that indicate changes in emotional state were also set by the annotators.

A. Annotation intersection

The audio annotation procedure aims to get agreement scores high enough to develop machine learning based systems. So, first, a set of files was chosen to be annotated separately. Then the inter-annotator agreement was computed. If the agreement is less than a predefined threshold, the annotators discuss together and re-annotate the files. Then a new set of files is given to the annotators. The inter-annotator agreement concern each label. It can be computed by the evaluation measures:

- Per event agreement (V): each event from one annotator is assigned only one event from the other annotator and two events result in a hit if the overlap between them is above a predefined threshold. Hence : V = hits / (hits + misses)

- Global agreement (G): This kind of agreement takes into account only the intersection oven union of the annotated labels. G = intersections/unions

For these experiments we used the weighted sum measure W= $\alpha * V + \beta * G$
$\alpha = \beta = 0.5$

In order to decide which annotation to use in the emotion recognition system, we report for each pair of annotations (1-2, 2-3 and 3-1), the per label mean agreement value. The mean agreement values (over all the files) are computed for the arousal, valence, dominance and categories (Fig 4, Fig 5, Fig 6, and Fig 7).

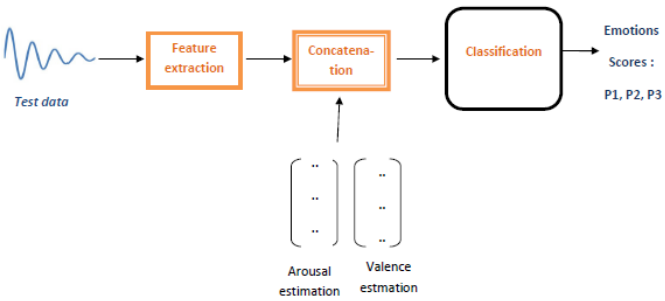


Fig. 3. Emotion recognition using acoustic features and estimated values of arousal/valence

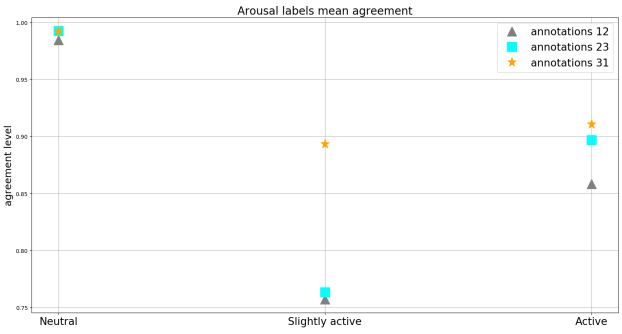


Fig. 4. Mean arousal agreement

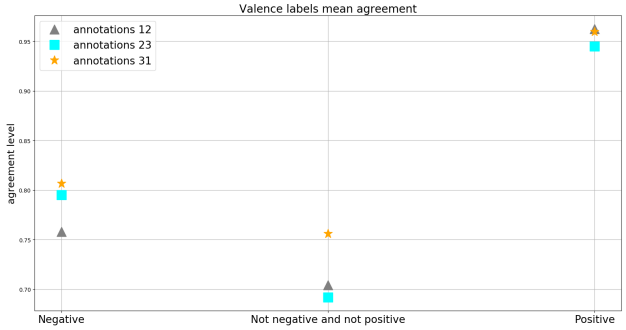


Fig. 5. Mean valence agreement

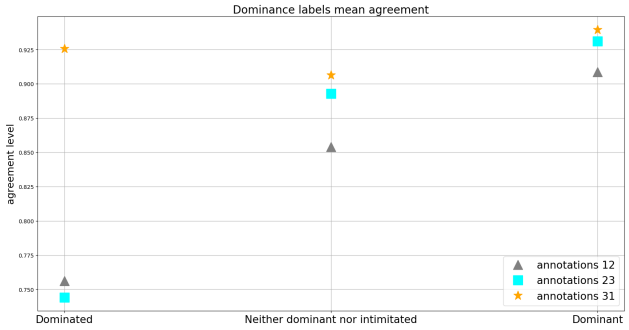


Fig. 6. Mean dominance agreement

For almost all the labels, the higher inter annotators agreement is between the annotators 1 and 3. So the intersection of the annotations 1 and 3 is computed. We decided to keep the segments with minimum duration = 0.8s.
The intersection is computed according to the categorical labels. For the dimensional labels means values between annotation1 and annotation3 is considered.
In the rest of this document, only the intersection parts of the annotations are considered.

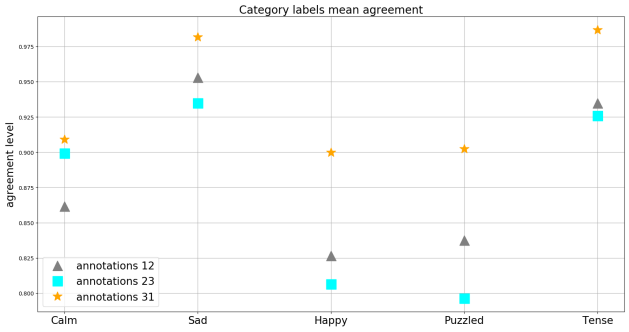


Fig. 7. Mean categories agreement

B. Categorical labels distribution

In the following figure, we report the time proportion of each categorical label present in the intersection.

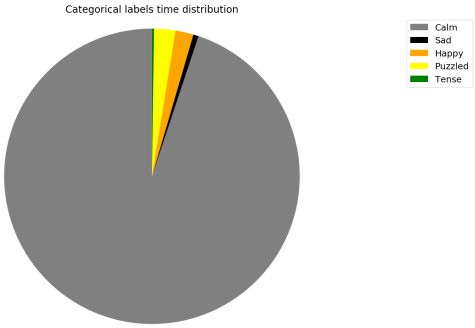


Fig. 8. Time distribution for the categorical labels

We notice that the most frequent label is “calm”. It represents 94% of the audio database which correspond to about 6 hours and 10 minutes. The label “sad” is quasi absent, so elder people seem to be happy with the dialog system. Happy and puzzled labels are present in only 4% of the database with respective duration’s of 9 and 8 minutes.
For these reasons, in the emotion recognition system only the frequent labels that are calm, happy and puzzled are considered.

C. Dimensional parameters relevance

The objective of this paragraph is to analyse the arousal/valence/dominance of the participants for each emotion expression. The three-dimensional model parameters of the intersection belong to the set (-1, -0.5, 0, 0.5 and 1). Hence, we report the number of segments labeled:

- neutral (-1) or slightly active (-0.5, 0 or 0.5) or active (1), for the arousal dimension
- positive (1) or neutral (-0.5, 0 or 0.5) or negative (-1), for the valence dimension
- dominant (1), neutral (-0.5, 0 or 0.5) or dominated (-1), for the dominance dimension.

These statistics are computed for each category label which are calm, sad, happy, puzzled and tense.

We notice that the arousal is mostly neutral for calm and puzzled and sometimes active for happy. This is quite concordant with the theory. The valence is rather positive for happy and calm and more neutral for puzzled. The dominance is mainly neutral. We conclude that:

- the arousal could discriminate calm and puzzled from happy
- the valence helps to distinguish between happy / calm and puzzled
- the dominance is not informative in our database, so we decide not to use it for emotional labels characterisation.

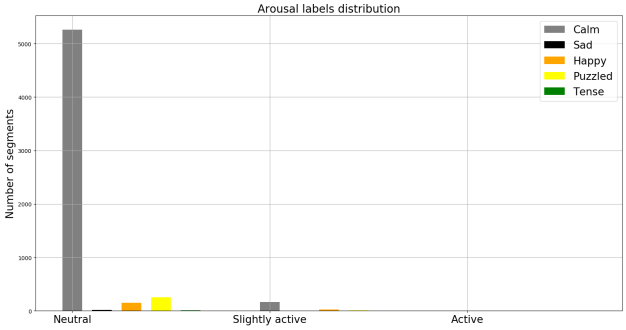


Fig. 9. Arousal segments per categorical label

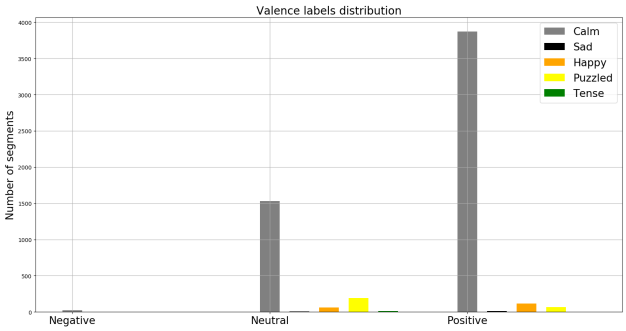


Fig. 10. Valence segments per categorical label

IV. Emotion recognition

The emotion recognition system classify an input speech segment into one emotional label among three labels that are calm, happy and puzzled.

A. Experimental protocol

The experimental protocol consists in using a set of 90% of the data for the training and another set (10%) for the tests. Support vector machines are used for the classification and F measure for the performance measure.

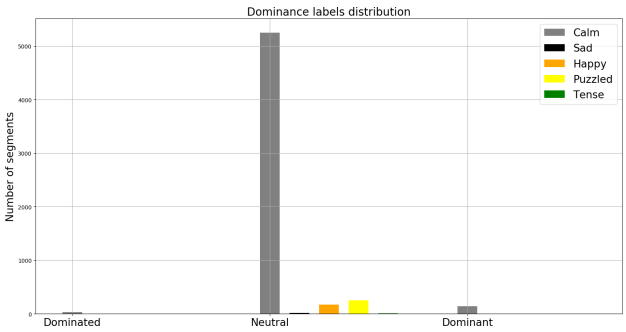


Fig. 11. Dominance segments per categorical label

The number of segments for the SVM training and test are reported in the following table. For the calm emotional category, only a subset of data is used in order to have a balanced database.

TABLE I
Categorical labels : number of segments

	happy	***puzzled***	***calm***
Train segments	160	234	234
Test segments	18	26	26

Feature extraction is carried out in two steps. In the first step 34 short term acoustic parameters are extracted for each frame. They are:

- Zero crossing rate
- Energy and Entropy of energy
- Spectral parameters: Centroid, Spread, Entropy, Speed and rolloff
- Mel frequency Cepstral Coefficients (13)
- Chroma Features (13)

In a second step, a mid term analysis is performed for each speech segment by computing the mean and the standard deviation for the short term acoustic features.

B. Arousal and valence estimation

In the following figures, we report for each emotion category, the distribution of the arousal and valence in :

- the train set: ground-truth
- the test set: ground-truth
- the test set2 : estimation

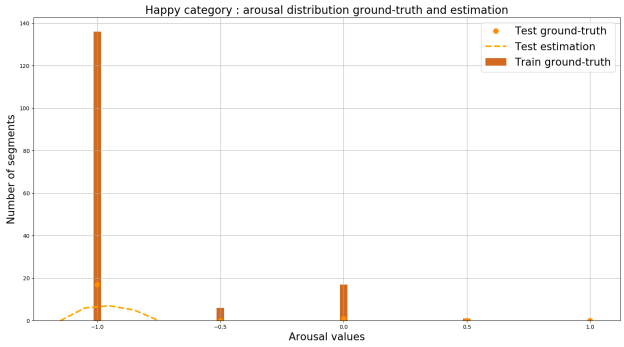


Fig. 12. Happy category : arousal distribution ground-truth and estimation

For the three emotions, the arousal ground-truth in the train and test data is mostly neutral (i.e. equal to -1). All the estimated values of the arousal are between -0.86 and -1.16 with a mean square error of 0.04. So the estimation of the arousal is quite concordant with its ground-truth.

For the valence, we notice that :

- Most ground-truth valence labels are neutral or positive in both train and test sets.

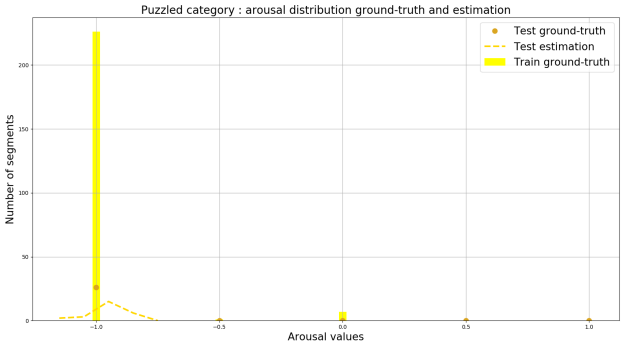


Fig. 13. Puzzled category : arousal distribution ground-truth and estimation

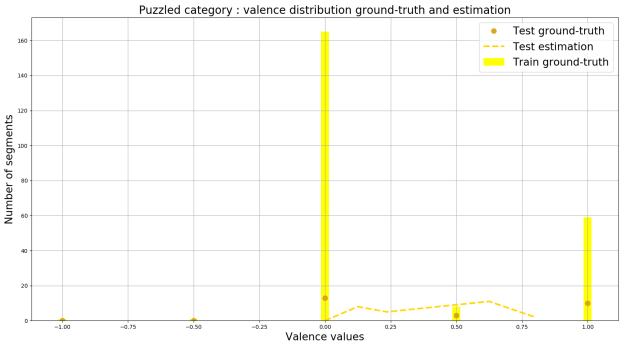


Fig. 16. Puzzled category : Valence distribution ground-truth and estimation

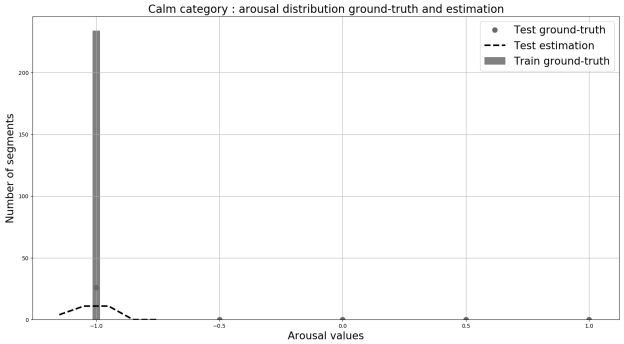


Fig. 14. Calm category : arousal distribution ground-truth and estimation

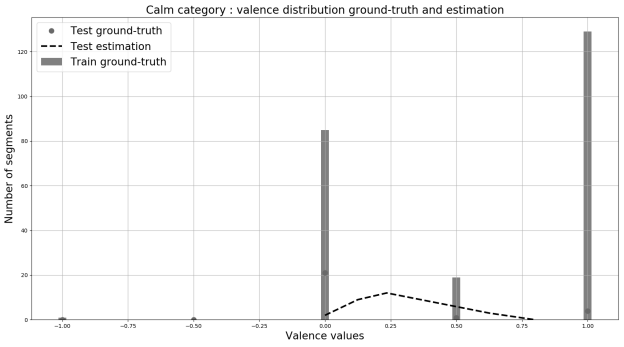


Fig. 17. Calm category : Valence distribution ground-truth and estimation

- Calm and puzzled do not have the same labels proportions in train and test sets.
- The estimation (with mean square error is 0.24) of the valence is rather positive than neutral.

From all of these analysis we can conclude that the estimations of the arousal and the valence are enough relevant to be considered in the emotion speech recognition system.

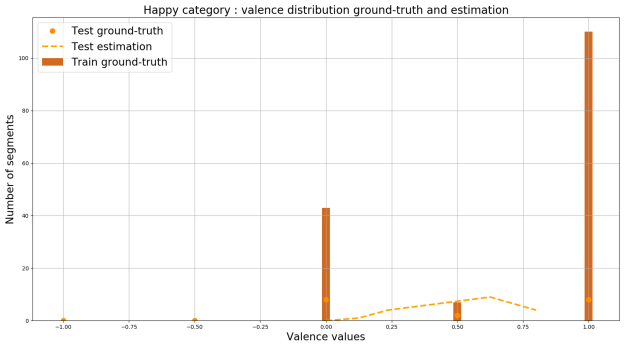
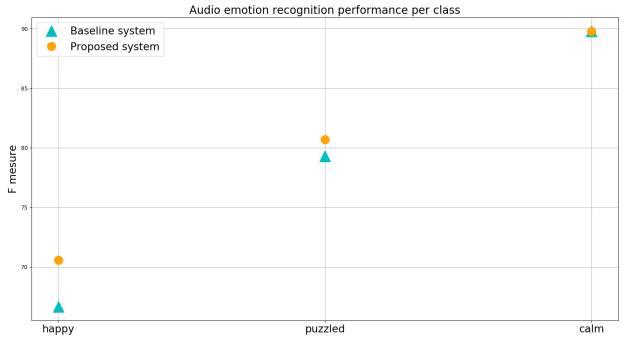


Fig. 15. Happy category : Valence distribution ground-truth and estimation

C. Recognition results

The F measure of the baseline system is 79.15%. The proposed system is evaluated in the same conditions as the baseline system. So the same parameters set is used and arousal and valence estimated values are added. The F measure of the proposed system is 80.75%

The F measure of the baseline and the proposed system per emotion label are reported in the following figure.



The proposed system outperform the baseline one for all the emotion categories.

V. Conclusion

In speech emotion recognition, the goal of the feature extraction module is to select distinctive speech parameters that contribute to accurate detection of emotion. Although many speech features have been explored by the scientific and phonetic communities, researchers have not yet identified the best speech features for this task. In this paper, we propose to exploit the correlation between categorical emotions and dimensional emotion model to improve the parametrisation. More practically, arousal and valence dimensional parameters are estimated by regression and integrated into the acoustic features. Primary experiments on Empathic database, have been conducted for a baseline system and the proposed one using the new parameters in the same conditions. They show a significant improvement of the emotion recognition system.

References

[1] K. Hartmann, I. Siegert, D. Philippou-Hubner and A. Wendemuth, "Emotion Detection in HCI: From Speech Features to Emotion Space," 12th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems. August 11-15, 2013. Las Vegas, NV, USA.

[2] M. El Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: features, classification schemes and databases," Pattern recognition, vol. 44. Elsevier Ltd2011, pp.572–587.

[3] F. Eyben, K. Scherer, B. Shuller, J. Sundberg, E. Andre, C. Busso, L. Devilliers, J. Epps, P. Laukka, S. Nrayanan and K. Thuong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPs) for Voice Research and Affective Computing", IEEE Transactions on Affective Computing, Vol. 7, Issue: 2, April-June 2016.

[4] C. Clavel and G. Richard,"Système d'intéraction émotionnelle,", Chapter 5: Reconnaissance acoustique des émotions. Ed. C. Pelachaud, Hermès, 2010

[5] R. Justo, L. Ben Letaifa, C. Palmero, E. Gonzalez-Fraile, A.T. Johansen, A. Vazquez, G. Cordasco, S. Schlogl, B. Fernandez-Ruanova, M.R. Silva, S. Escalera, M. De Velasco, J. Tenorio-Laranga, A. Esposito, M. Kornes, and M. Ines Torres, "Analysis of the Interaction between Elderly People and a Simulated Virtual Coach". In Journal of Ambient Intelligence and Humanized Computing.

[6] F. Eyben, F. Weninger, and B. Shuller, "Affect recognition in real-life acoustic conditions - A new perspective on feature extraction," Interspeech, August 25-29, 2013. Lyon, France.

[7] D. P. Szameitat, C. Widgruber, K. Alter, and A. J. Szmaeitat, "Acoustic correlates of emotional dimensions in laugher: Arousal, dominance, and valence" , Cognition and emotion.pp. 1-13: Ed Psychology Press 2010.

[8] M. Thibeault, "Les emotions: une étude articulatoire, acoustique et perceptive", Phd Thesis. Quebec university. Montréal 2011.

[9] L. Ben Letaifa, M. de Velasco, R. Justo and M. Inés Torres, "First Steps to Develop a Corpus of Interactions between Elderly and Virtual Agents in Spanish with Emotion Labels", 7th International Conference on Statistical Language and Speech Processing. October 14-16, 2019. Ljubljana, Slovenia.

[10] M. de Velasco, R. Justo, A. López Zorrilla and M. Inés Torres, "Can Spontaneous Emotions be Detected from Speech on TV Political Debates? ", 10th IEEE International Conference on Cognitive Infocommunications, October 23-25, 2019. Naples, Italy.