IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Emotion Detection from Speech and Text

Mikel deVelasco, Raquel Justo, Josu Antón, Mikel Carrilero, M. Inés Torres

Universidad del Pais Vasco UPV/EHU

mikel.develasco@ehu.eus, raquel.justo@ehu.eus, josuantonsanz@gmail.com,
mcarrilero001@ikasle.ehu.eus, manes.torres@ehu.eus

Abstract

The main goal of this work is to carry out automatic emotion detection from speech by using both acoustic and textual information. For doing that a set of audios were extracted from a TV show were different guests discuss about topics of current interest. The selected audios were transcribed and annotated in terms of emotional status using a crowdsourcing platform. A 3-dimensional model was used to define an specific emotional status in order to pick up the nuances in what the speaker is expressing instead of being restricted to a predefined set of discrete categories. Different sets of acoustic parameters were considered to obtain the input vectors for a neural network. To represent each sequence of words, a models based on word embeddings was used. Different deep learning architectures were tested providing promising results, although having a corpus of a limited size.

Index Terms: Emotion Detection, Speech, Text transcriptions

1. Introduction

The emotion recognition is the process of identifying human emotions, a task that is automatically carried out by humans considering facial and verbal expressions, body language, etc. However, this is a challenging task for an automatic system. In recent years, the great amount of multimedia information available due to the extensive use of the Internet and social media, along with new computational methodologies related to machine learning, have led to the scientific community to put a great effort in this area [1, 2].

Emotion recognition from speech signals relies on a number of short-term features such as pitch, vocal tract features such as formants, prosodic features such as pitch loudness, speaking rate, etc. Surveys on databases, classifiers, features and classes to be defined in the analysis of emotional speech can also be found in [3]. Regarding methodology, statistical analysis of feature distributions has been traditionally carried out. Classical classifiers such as the Bayesian or Super Vector Machines (SVM) have been proposed for emotion features from speech. The model of continuous affective dimensions is also an emerging challenge when dealing with continuous rating emotion labelled during real interaction [1, 4]. In this work, recurrent neural networks have been proposed to integrate contextual information and then predict emotion in continuous time using a three-dimensional emotional model.

Speech transcripts have also been demonstrated to be a powerful tool to identify emotional states [5]. Over the last decade, there has been considerable work in sentiment analysis [6]. Moreover, the detection of emotions such as anger, joy, sadness, fear, surprise, and disgust have also been addressed [7]. However, spoken language is informal and provides information in an unstructured way so that developing tools to select and analyse sentiments, opinions, etc. is still a challenging topic [8]. In early systems dealing with emotion detection in text, knowledge-based approaches were applied making use of emotion lexicons, such as Sentiwordnet [9]. Other methods, employ machine learning based approaches [10], where statistical classifiers are trained using large annotated corpora and the emotion detection can be seen as a multi-label classification problem. In this work we propose to use neural networks to solve the regression problem given the 3-dimensional emotional model.

The main contribution of this work is the appropriate selection of a neural network architecture for emotion detection considering both acoustic signals and their corresponding transcription. Additionally, the proposed architecture has been adapted to the 3-dimensional VAD (Valence, Arousal and Dominance) emotional model [11].

Section 2 describes the two proposed approaches for the automatic emotion recognition from used features to the common network architectures basics. Experiments carried out are fully described in Section 3. Section 3.1 aims to explain the difficulties to find out a Spanish corpus and it has led us to create our own corpus. Section 3.2 mentions the used baselines methods and which measure has been used for testing. Section 3.3 shows the experiments carried out under the regression problem of emotional status with acoustic features whereas Section 3.4 deals with the experiments achieved at the regression problem of emotional status with language features. Finally some concluding remarks are reported in Section 4.

2. Emotion Detection from Speech and Language

Emotion detection from speech is based on the extraction of relevant features from the acoustic signal, that can be seen as hints of the emotional status. That is, a numerical vector that represent the specific information related to emotional status and embedded in an acoustic signal is needed. There are numerous acoustic parameters that can be obtained using the free software *Praat*[1] or the free python library *pyAudioAnalysis*[2]. Considering these tools the following set of 72 parameters could be considered: Pitch, Zero Crossing Rate (ZCR), Energy, Entropy of the energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral flux, Spectral Rolloff, Chroma vector (12 coefficients), Chroma deviation, MFCC coefficients (12), LPC coefficients (16), Bark features (21). However, some of these parameters provide the same or very similar kind of information. For instance, LPC, Bark and MFCC coefficients provide similar information about the phonemes without considering the vocal tract. Thus, such a big set of parameters is useless since it will complicate the learning procedure requiring more training data. In this work different subsets of the aforementioned parameters were explored:

[1] http://www.fon.hum.uva.nl/praat/

[2] https://pypi.org/project/pyAudioAnalysis/

10.21437/IberSPEECH.2018-15

- 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.

The first set was selected according to the studies performed in [12] where the arousal state of the speaker affects the overall energy and pitch. In addition to time-dependent acoustic features such as pitch and energy, spectral features were selected for Sets B and C as a short-time representation for speech signal [13]. For Sets D, E and F different Cepstral-based features were added, proven that they are good for detecting stress in speech signal [14].

When regarding emotion detection from language the same procedure has to be carried out. First of all a vectorial representation of the transcribed text is needed. In this case, we hope to capture some meaning of the utterance that might help in the detection of specific emotional status. An appropriate representation should consider some semantic information like the word embeddings *word2vec*[15], *doc2vec* [16] or *GLOVE* [17]. *Word2vec* embeddings, the most simple model, are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. *Word2vec* takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. However, this technique represents each word of the vocabulary by a distinct vector, without parameter sharing. In particular, they ignore the internal structure of words, which is an important limitation for morphologically rich languages. For example, in Spanish, most verbs have more than forty different inflected forms and this leads to a vocabulary where many word forms occur rarely (or not at all) in the training corpus, making it difficult to learn good word representations. Thus, [15] proposes to learn representations for character n-grams, and to represent words as the sum of the n-gram vectors. The model (known as *FastText*) can be seen as an extension of the continuous skip-gram model [18] which takes into account subword information.

Additionally, a way of representing the emotional status is needed in order to establish a machine learning problem. A categorical emotion description (e.g. six basic emotions) is an easy way to procedure but it provides a quite constrained model. Affective computing researchers have started exploring the dimensional representation of emotion [19] as an alternative. Dimensional emotion recognition, aims to improve the understanding of human affect by modelling affect as a small number of continuously valued, continuous time signals. It has the benefit of being able to: (i) encode small changes in affect over time, and (ii) distinguish between many more subtly different displays of affect, while remaining within the reach of current signal processing and machine learning capabilities [20]. In our work, we represent the problem of dimensional emotion recognition as a regression one, where each emotional status is represented as three-dimensional real-valued vector. The dimensions of this vector correspond to Valence (corresponding to the concept of polarity), Arousal (degree of calmness or excitement), and Dominance (perceived degree of control over a situation): the VAD model.

In order to solve the regression problem of emotional status detection, we propose to use deep learning. When considering emotion detection from speech Long-Short Term Memory (LSTM) neural networks were tested. The underlying idea is to be able to learn the relationship among present and past information although existing a big distance among them. That is, they have memory and they can manage with temporal sequences of data like the sequence of vectors extracted from an acoustic signal. When regarding emotional status detection from text a classical feedforward network was considered because of simplicity. Such networks have proven to be efficient for problems in similar tasks, like sentiment analysis [21].

3. Experiments

We have carried out two series of experiments for the evaluation of the regression processes. In the first one we present the most interesting results related to the emotion detection from speech, and in the second we show the most interesting results on emotion detection from text.

3.1. Corpus

As far as we know there is no Spanish three-dimensional corpus within the literature, so for the experiments, we have created a small corpus using the VAD model. The corpus consists of 120 fragments between 3 and 5 seconds taken from the Spanish TV program "La Sexta Noche". This TV program consist of political debate, news and events, and discussions commonly appear. Each fragment has been transcribed manually and tagged using crowdsourcing (the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people) techniques. In this case each fragment has been labeled by 5 different annotators, following the next questionnaire:

1. In order to address the Valence: "How do you perceive the speaker?"
   - Excited
   - Slightly Excited
   - Neutral
2. In order to address the Arousal: "His mood is . . . "
   - Positive (nice / constructive)
   - Slightly Positive
   - Slightly Negative
   - Negative (unpleasant / non colaborative)
3. In order to address the Dominance: "How do you perceive the speaker about the situation in which he or she is in?"
   - More dominant / controlling the situation / . . .
   - He or she does not dominate the situation neither is he or she cowed.
   - More coward / defensive / . . .

Once the tags were generated by crowdsourcing, the answers collected from all the annotators were transferred to the three-dimensional model, making the average of each answer

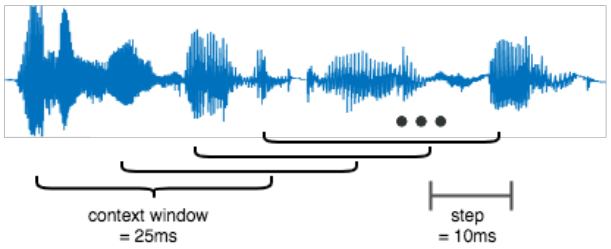

Figure 1: *Schematic diagram of speech production.*

for all fragments where the first answer of each question was assigned the value 0, the last answer was assigned the value 1, and the rest of the answers a midpoint. Then the corpus was split into two sets, 70% of the fragments were used for training purposes and the remaining 30% for test.

3.2. Baselines models and Evaluation Metrics

Both emotion detections problems, from speech and from text, has been tested first with Linear Regression (LR) [22] and Super Vector Regression (SVR) [23] (with three different types of kernels *linear*, *poly*, and *rbf*), in order to compare with Neural Networks.

Regarding the input of these baselines models, two different approaches have been analysed to fix the problem of time sequence. On the one hand, we fit the models with full information, considering each feature in on each time-step independent (*full* models), and on the other hand, calculating the mean of each feature over time-steps (*mean* models).

In relation to the evaluation metric, the Mean Square Error (MSE) has been used, because it seems to provide a good interpretation of how far the prediction and the true label are. In this problem, MSE can be described as the mean of the distances between the points of the true label on the three-dimensional model and the predicted points on the same three-dimensional model.

3.3. Experiments with acoustic features

In order to obtain the acoustic parameters, each audio has been divided into individual frames using a context window of 25 milliseconds and a step of 10 milliseconds (as shown in Figure 1), obtaining 300 frames per audio. A vector made up of the selected acoustic features was associated to each frame. Different experiments were carried out using the different feature sets (A, B, C, D, E, F) described in Section 2. Additionally, for each set, different experiments were also performed including both the first and the first and the second derivatives.

The network proposed to address the regression problem of emotional status with acoustic features is a Recurrent Neural Network (RNN). The network is composed with an LSTM layer (the architecture proposed by [24]) of 10 cell memory blocks, to get a representation of the audio along the time. Subsequent layers are two Dense layers which aim to infer the VAD model from the representation given by the LSTM. The fist Dense layer consist of 15 units and *ReLU* activation function while the second and last consist of 3 units and *sigmoid* activation function.

The output layer contains a sigmoidal activation function to take advantage of the output limitation benefits, it is bounded between 0 and 1. On the other hand, the hidden layer contains the *ReLU* activation function because it provides good results and avoids the vanishing gradient problem [25].

The layers of the proposed network consist of a small number of units or cells not to build a large network architecture, since we have a limited sized training corpus.

Table 1: *Best result obtained with baseline models and Recurrent Neural Network (RNN) in the regression problem of emotional status with acoustic features. Each pair of set and model has been tested with acoustic features itself, with first derivatives and with first and second derivatives, but only best performance is shown. MSE error has been used in order to compare.*

	LR	**SVR**			**RNN**
		linear	poly	rbf	
Set A	0.1682	**0.1660**	0.1661	0.1691	0.1670
Set B	0.1686	0.1653	0.1685	0.1710	**0.1565**
Set C	0.1690	0.1679	0.1718	0.1709	**0.1576**
Set D	0.1742	0.1894	0.1703	0.1710	**0.1665**
Set E	0.1699	0.2007	0.1842	0.1711	**0.1664**
Set F	0.1733	0.2256	0.1898	0.1710	**0.1413**

As shown in Table 1, the proposed network slightly improves the results of the baseline models in almost all the sets in the corpus. It can also be concluded that by selecting a smaller set of parameters, better results are obtained with the baseline model. However, the set A seems to have insufficient information and Bark features are of great help in the case of networks providing the best results.

3.4. Experiments with language features

Regarding the word representation, *FastText* embeddings from SBWC[3] has been used in the experiments. The mentioned embeddings are a Skipgram model of 300 dimensions and 855380 different word vectors, trained with Spanish Billion Word Corpus[4] with more than 1.4 billion words.

The regression problem of emotional status with language features has been addressed with a small Deep Neural Network (DNN). This network consist of three similar layers; the first two layers are composed of 5 units, a sigmoidal activation function and followed by Dropout layer with 0.5 of keep-probability in order to prevent to the overfitting problem [26]; while the last layer, the output layer, is a Dense layer of 3 units and the sigmoidal activation function (same as the network proposed for the regression problem of emotional status with acoustic features).

Table 2: *Best result obtained with baseline models and Deep Neural Network (DNN) in the regression problem of emotional status with language features. MSE error has been used.*

	LR	**SVR**			**DNN**
		linear	poly	rbf	
Mean	0.1906	0.1350	**0.1165**	0.1197	0.1203
Full	0.1356	0.1292	0.1199	0.1229	**0.1196**

As shown in Table 2, the proposed network achieves similar results when comparing it to the baselines models. It is an interesting result given the small size of the training set and the great

[3]https://github.com/uchile-nlp/spanish-word-embeddings/blob/master/README.md

[4]http://crscardellino.me/SBWCE/

impact it has when building neural networks. The obtained results suggest that increasing the annotated training corpus neural networks might improve the baseline models.

4. Conclusions

The main goal of this work was to develop an automatic emotion detection system from speech and language. The system acted over acoustic fragments extracted from a TV show and their corresponding transcriptions. Each fragment was annotated by means of a crowdsourcing platform using a 3-dimensional VAD model. Different neural networks architectures were tested and the obtained results show that RNN can outperform baseline systems when considering emotion detection from speech. Moreover, using a simple feedforward neural network with a very small training corpus (84 sentences) similar results to those obtained with baseline models can be achieved.

For further work we propose to get a bigger annotated corpus by using crowdsourcing tools to better train the proposed neural networks. Additionally, the two knowledge sources (acoustic and text) might be merged to provide a more accurate emotion detection system.

5. Acknowledgements

This work has been partially founded by by the Spanish Government (TIN2014-54288-C4-4-R and TIN2017-85854-C4-3-R), and by the European Commission H2020 SC1-PM15 program under RIA 7 grant 69872.

6. References

[1] J. Irastorz and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in *7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2016, p. 103106.

[2] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. B. 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, Shanghai, China, March 20-25, 2016*, 2016, pp. 2274–2278.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, pp. 1162–1181, 2006.

[4] A. Mencattini, E. Martinelli, F. Ringeval, B. W. Schuller, and C. D. Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Trans. Affective Computing*, vol. 8, no. 3, pp. 314–327, 2017. [Online]. Available: https://doi.org/10.1109/TAFFC.2016.2531664

[5] C. Clavel, G. Adda, F. Cailliau, M. Garnier-Rizet, A. Cavet, G. Chapuis, S. Courcinous, C. Danesi, A.-L. Daquo, M. Deldossi *et al.*, "Spontaneous speech and opinion detection: mining call-centre transcripts," *Language resources and evaluation*, vol. 47, no. 4, pp. 1089–1125, 2013.

[6] S. Mohammad, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 599–608.

[7] J. R. Bellegarda, "Emotion analysis using latent affective folding and embedding," in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 2010, pp. 1–9.

[8] R. Justo, T. Corcoran, S. M. Lukin, M. Walker, and M. I. Torres, "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web," *Knowledge-Based Systems*, vol. 69, pp. 124–133, 2014.

[9] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[10] S. Volkova and Y. Bachrach, "Inferring perceived demographics from user emotional tone and user-environment emotional contrast," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1567–1578.

[11] R. A. Calvo and S. Mac Kim, "Emotions in text: dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[12] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[13] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[14] S. E. Bou-Ghazale and J. H. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 8, no. 4, pp. 429–442, 2000.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[16] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1188–1196.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*. ACL, 2014, pp. 1532–1543.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[19] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[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. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann." *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, 2013.

[22] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[23] C.-C. Chang, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology, 2: 27: 1–27: 27, 2011 "http://www. csie. ntu. edu. tw/˜ cjlin/libsvm"*, vol. 2.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.