Emotion Recognition from Speech: A Classroom Experiment

Rozalia Nikopoulou
Department of Informatics
Ionian University
Corfu, Greece
rnikopoulou@gmail.com

Ioannis Vernikos
Department of Computer Science
University of Thessaly
Lamia, Greece
imvernikos@gmail.com

Evaggelos Spyrou
Institute of Informatics and Telecomunnications
National Centre for Scientific Research – "Demokritos"
Athens, Greece
espyrou@iit.demokritos.gr

Phivos Mylonas
Department of Informatics
Ionian University
Corfu, Greece
fmylonas@ionio.gr

ABSTRACT

In this position paper we present an approach for the recognition of emotions from speech. Our goal is to understand the affective state of learners upon a learning process. We propose an approach that uses visual representations of the spectrum of audio segments, which are classified using the Bag-of-Visual Words model. Our approach is applied on a real-life dataset that contains interviews from middle-school students, collected upon a classroom experiment.

KEYWORDS

emotion recognition, speech, classroom,bag-of-visual words

ACM Reference Format:
Rozalia Nikopoulou, Ioannis Vernikos, Evaggelos Spyrou, and Phivos Mylonas. 2018. Emotion Recognition from Speech: A Classroom Experiment . In *Proceedings of PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA2018)*. ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Undoubtedly, the basic and the mostly used means of everyday human communication is the vocalized speech. Apart from its actual semantic meaning, it also carries emotions that - in the common case where audio is the only available modality - are not easily recognizable. In principle, information carried by speech is twofold [1]: a) linguistic (i.e., articulated patterns); and b) paralinguistic (i.e., variation in pronunciation). The former may be described in a qualitative way, while the latter in a quantitative one, i.e., one may extract related features and build her/his analysis on top of them. Both types can be used at a later stage to classify spoken content to a predefined set of emotions.

In this work, we present a paralinguistic approach, which is based on the extraction of visual features from spectrogram representations of spoken content. More specifically, an audio segment is transformed into a spectrogram, i.e., a visual representation of its spectral information. A Bag-of-Visual Words (BoVW) model is applied for classification of spectograms to emotions. We applied our proposed approach in a real-life classroom environemnt, where a group of middle-school students participated in the experiment. Upon the experiment's end, small interviews with the students were conducted and they were asked to express their unbiased opinion. The outcomes of the interviews (i.e., the resulting recordings) were annotated and used for emotion classification.

The rest of this paper is organized as follows: in section 2 we present related work regarding emotion recognition. Next, in section 3 we describe the classroom experiment and the BoVW methodology. Experiments are presented in section 4, while conclusions are drawn in section 5, where plans for future work are also presented.

2 RELATED WORK

The work of Martiinez [7] focuses mainly on moral emotions (guilt, remorse, shame), which differ from the basic ones (sadness, happiness, etc.). It points out that by understanding emotions, one may then raise the moral level of society and since education is the gate to society, it is of significant importance to focus on the emotions of the students, to promote skills such as self-knowledge and control. Tickle et al. [12] examined the use of virtual agents in the role of educator and the necessity for them to be provided with the ability to sense the emotional state of the students. Upon emotion recognition virtual agents could then make interaction more appealing for the students. Although they had satisfactory results, the concluded that recognizing facial expressions may need to further improvement. Bahreini et al. [2] examined the advantages of speech emotion recognition in e-learning, to facilitate smoother interaction between humans and computers. The introduced FILTWAM framework may be used in any e-learning setting, is able to identify both vocal and facial emotions and may give relevant feedback to the learner without the interference of the teacher. In general, typical approaches that aim to classify audio content to an a-priori known number of emotional states make use of well-known classifiers such as Hidden Markov Models and Support Vector Machines [9, 13]. Other works [4, 6] adopt the psychophysiology-based dimensional approach. Another recent trend is the use of deep-learning methods [10]. In previous work [11] experimented with BoVW and spectrograms, in well-known datasets.

PETRA2018, June 2018, Corfu, Greece
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METHODOLOGY

The classroom experiment was conducted in a computer laboratory of a middle school with the participation of students aged between 12-13 years old, by their ICT instructor. The students were familiar with their instructor which indicates that both their reactions and their expressions were authentic and not restrained. Students were divided into two teams. For both teams, the task was to build and program a robot using LEGO Mindstorm EVE 3 Educational kit[1] in combination LEGO Mindstorm EVE 3 software and Scratch[2]. The first team was asked to program the robot to be remotely controlled through a mobile phone, whereas the second team to follow a certain route using color sensors. During the experiment there was minimum interference by the instructor who at that time was documenting the students' reactions and facial expressions. Upon the completion of the experiment, each student was interviewed by the instructor and her/his voice was recorded and annotated based on both his vocal and facial appearence.

From each audio stream a single, randomly cropped segment of 2 sec length is extracted. For each segment, its spectrogram is extracted, using the Short-Time Fourier Transform at an 40 msec short-term window size and 20 msec step. From each spectrogram, we extract SURF features [3] from image patches surrounding pixels that have been sampled using a regular grid. The visual vocabulary of BoVW model is built by applying the k-means clustering algorithm on the SURF features from a training dataset. Then, we encode each feature to a visual word and calculate their frequencies. Finally, a spectrogram is represented by a feature vector comprising of the frequencies of all visual words of the dictionary. For classification, we train a multi-class SVM classifier.

4 EXPERIMENTAL RESULTS

For the sake of the aforementioned experiment, we involved 24 students (15 male and 9 female) and ended up with 42 recordings, with average duration 7.8 sec. These recordings were collected using a microphone of a personal computer and were post-processed in order to remove parts with silence and/or the voice of the instructor. Upon the annotation process, the dataset consisted of 24 samples with a positive emotion, 8 with a negative and 10 with a neutral. We split each sample into non-overlapping segments of $t = 1$ and $t = 2$ sec and extracted the corresponding spectrograms, using the pyAudioAnalysis opensource Python library[3] [5]. Their dimensions were set equal to 227×227. The BoVW model has been implemented using the Computer Vision Toolbox of Matlab R2016a [8]. We opted to create a balanced dataset, i.e., all three classes were represented by equal number of samples during training. Thus, the number of training samples per class was equal to the 80% of the samples of the smallest class. Remaining samples were used for testing. We used a grid size of 8×8 for the SURF features and varying size of $N = 100, 200, \ldots 1500$ for the visual vocabulary. For classification, we used an SVM with a gaussian kernel. The confusion matrices for the best vocabulary size are depicted in Tables 1 (1 sec) and 2 (2 sec), respectively. As it may be observed, the performance of our approach was satisfactory in case of 2−sec audio segments.

Table 1: Classification accuracy ($t = 1$sec, $N = 1200$ words).

	positive	neutral	negative
positive	**0.56**	0.22	0.22
neutral	0.22	**0.67**	0.11
negative	0.33	0.22	**0.45**

Table 2: Classification accuracy ($t = 2$sec, $N = 100$ words).

	positive	neutral	negative
positive	**0.75**	0	0.25
neutral	0	**1.00**	0
negative	0	0.25	**0.75**

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented an approach for emotion recognition from speech. We only used paralinguistic information and we did not extract audio spectral features. Instead, we opted to use visual representations of the spectrum of audio segments and classified them by applying the BoVW model. Our experiments were performed on a dataset consisting of recordings from students upon a classroom experiment and indicate the potential of our approach. Among our plans is to perform emotion recognition during the classroom experiments and adapt them accordingly based on the emotional state of the learners.

ACKNOWLEDGMENT

The work presented in this document is a result of MaTHiSiS project. This project has received funding from the European Union's Horizon 2020 Programme (H2020-ICT-2015) under Grant Agreement No. 687772.

REFERENCES

[1] C.N. Anagnostopoulos, T. Iliou and I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artificial Intelligence Review, 43(2), pp.155-177, 2015.

[2] K. Bahreini, R. Nadolski and W. Westera. Towards real-time speech emotion recognition for affective e-learning. Educ. and Inform. Techn. 21(5):1367-86, 2016.

[3] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, *Speeded-up robust features (SURF)*, Computer Vision and Image Understanding, 110(3), pp.346–359, 2008.

[4] T. Giannakopoulos, A. Pikrakis and S. Theodoridis, A dimensional approach to emotion recognition of speech from movies. In IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP), 2009.

[5] T. Giannakopoulos, *pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis*, PloS one, vol. 10(2), pp. e0144610, 2015.

[6] M. Grimm, K. Kroschel, E. Mower and S. Narayanan, Primitives-based evaluation and estimation of emotions in speech. Speech Comm., 49(10), pp.787-800, 2007.

[7] J.G. MartiÃŋnez, Recognition and emotions. A critical approach on education, Procedia-Social and Behavioral Sciences 46 pp. 3925-3930, 2012.

[8] MATLAB and Computer Vision Toolbox Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States.

[9] A. Nogueiras, A. Moreno, A. Bonafonte and J.B. Marino, Speech emotion recognition using hidden Markov models. In INTERSPEECH, 2001.

[10] M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos, Ph. Mylonas and F. Makedon. Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition, Computation 5(2), 26, MDPI, 2017.

[11] E. Spyrou, T. Giannakopoulos, D. Sgouropoulos and M. Papakostas, Extracting Emotions from Speech using a Bag-of-Visual-Words Approach. Int'l Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2017.

[12] A. Tickle, S. Raghu and M. Elshaw. Emotional recognition from the speech signal for a virtual education agent. Journal of Physics: Conference Series 2013 (Vol. 450, No. 1, p. 012053). IOP Publishing, 2013.

[13] Y. Wang and L. Guan, Recognizing human emotional state from audiovisual signals, IEEE Trans. on Multimedia, 10(5), pp.936-946, 2008.

[1] https://www.lego.com/en-us/mindstorms

[2] https://scratch.mit.edu/

[3] https://github.com/tyiannak/pyAudioAnalysis