# A Non-Linguistic Approach for Human Emotion Recognition from Speech

Evaggelos Spyrou[1,2,3], Ioannis Vernikos[3], Rozalia Nikopoulou[4] and Phivos Mylonas[4]
[1]*Institute of Informatics and Telecommunications National Center for Scientific Research – "Demokritos,"* Athens, Greece
[2]*Department of Computer Engineering, Technological Education Institute of Sterea Ellada,* Lamia, Greece
[3]*Department of Computer Science, University of Thessaly,* Lamia, Greece
[4]*Department of Informatics*, Corfu, Greece
email:espyrou@iit.demokritos.gr, imvernikos@gmail.com, rnikopoulou@gmail.com, fmylonas@ionio.gr

***Abstract*—One of the most important issues in several aspects of human-computer interaction is the understanding of the users' emotional state. In several applications such as monitoring of humans in assistive living environments, or assessing students' affective state during a course, it is imperative to use an unobtrusive method, so as to avoid discomforting or distracting the user. Thus, one should opt for approaches that use either visual or audio sensors which may observe users, without any kind of direct contact. In this work, our goal is to recognize the emotional state of humans using only the non-linguistic aspect of speech information, i.e., the acoustic properties of speech. Therefore, we propose an emotion classification approach that is based on the bag-of-visual words model that has been previously applied in many computer vision tasks. A given audio segment is transformed to a spectrogram, i.e., a visual representation of its spectrum. From this representation we first extract SURF features and using a previously constructed visual vocabulary, we quantize them into a set of visual words. Then a histogram is constructed per image; These feature vectors are used to train SVM classifiers. Our approach is evaluated using a) 3 publicly available datasets that contain speech from different languages and b) a custom dataset that has been constructed during a real-life classroom experiments, involving middle-school students.**

## I. Introduction

Human emotion recognition [1] is a research area that aims at recognizing humans' emotional states. The recognized emotions may be exploited in several applications within the broad field of human-computer interaction. To this goal, there exist many approaches that use several types of sensors, such as cameras and microphones, or even body sensors. More specifically, visual-based approaches try to recognize emotions based on the users' visual appearance, e.g., by recognizing facial features, gestures, postures, motion etc. [2], [3]. Audio-based approaches rely on speech, and extract either low- or mid-level features [4], or even process the actual spoken content, i.e., apply a natural language understanding method [5]. Body sensors measure physiological parameters such as temperature, heart rate, muscle activity, skin conductance response or even brain activity [6].

Undoubtedly, among the least obtrusive approaches is the use of microphone sensors on vocalized speech. Often people feel that their privacy is violated when observed by cameras or microphones. However, when body sensors are used for long time periods they typically feel discomfort. Thus, a totally unobtrusive approach is non-feasible. Therefore, in this work we opt to use speech-related audio features to recognize the emotional state of humans. It is our belief that such an approach is the least obtrusive among the aforementioned. Note that it does not rely on the linguistic content of speech, i.e., on the set of spoken phrases. Instead, it is based on low-level spectral audio features, not considering the semantics of spoken content.

Apart from the aforementioned linguistic content of speech, i.e., the articulated patterns that are pronounced by the speaker, the other component of speech is its non-linguistic content [7]. This consists of the variation of the pronunciation of the articulated patterns, i.e., the acoustic aspect of speech. Typically, description of linguistic patterns is qualitative. On the other hand, low-level features may be extracted from the non-linguistic ones, such as the rhythm, the pitch, the intensity etc. Many emotion classification approaches from speech exploit both types of information.

Linguistic patterns may be extracted using an automatic speech recognition system (ASR). However, such approaches do not easily provide language-independent models, as there exists a plethora of different sentences, speakers, speaking styles and rates [8]. On the other hand, approaches based on non-linguistic features may be more robust to different languages. However, even in this case, emotion recognition still remains a challenging problem, since non-linguistic content may severely depend on factors such as cultural particularities, or even potential chronic emotional state of a given speaker.

The majority of approaches that use the non-linguistic component of speech typically rely on the extraction of spectral features from the raw audio speech signals [7]. In this work and contrary to previous ones, we do not extract such audio features. Instead, we transform an audio segment to a spectrogram, which is a visual representation of the spectrum of its frequencies, as they vary with time. A spectrogram is an image, thus we opt to apply a computer vision algorithm on it, known as the Bag-of-Words (BoW) or Bag-of-Visual Words (BoVW) model [9]. In previous work [10] we first experimented with this model and obtained promising results using several well-known datasets.

In this work we present an investigation of the BoVW model on audio spectrograms, using datasets of several languages. Moreover, apart from using datasets that have been created

using actors, we also present a real-life experiment. More specifically, we have applied the proposed methodology in a real-life classroom environment, using a group of middle-school students and conducted an experiment where they were asked to freely express their unbiased opinion regarding the exercise they participated. The outcomes of the interviews (i.e., the resulting recordings) were annotated and used for emotion classification.

The structure of the rest of this paper is as follows: in Section II we present related work on emotion classification from audio focusing on applications in education. Then, in Section III we describe the Bag-of-Visual Words model and the proposed emotion classification strategy. Data sets used and experiments are presented in Section IV. Finally, results are discussed and conclusions are drawn in Section V, where plans for future work are also presented.

## II. Related Work

Emotion recognition from the non-linguistic component of multimedia documents containing speech has been typically based on the extraction of several low-level features which were then used to train models and map them to the underlying emotions. Such research efforts [11]–[13] in many cases included fusion of multimodal features. In case of music signals, several efforts have turned to the recognition of their affective content [14], [15] with applications to affect-based retrieval from music databases. Well-known classifiers have been used, e.g., Hidden Markov Models or Support Vector Machines, to extract features to emotional states. Such states either include fear, happiness, anger etc. [11], [12], or in other cases [16]–[18] follow the dimensional approach [19] that originates from psychophysiology. Spectrograms have also been previously used for other audio analysis-related tasks, such as content classification [20] and segmentation [21], for stress recognition [22] and more recently for emotion recognition with convolutional neural networks [4], [23]. As for applications of emotion recognition in education, focus is typically given on moral emotions (guilt, remorse, shame), which differ from the basic ones (sadness, happiness, etc.) [24]. Also, virtual agents have been used in the role of educator [25] which were provided with the ability to sense the emotional state of the students. Upon emotion recognition these agents could then make interaction more appealing for the students. Bahreini et al. [26] examined the advantages of speech emotion recognition in e-learning, to facilitate smoother interaction between humans and computers.

## III. Emotion Recognition from Spectrograms using BoVW

In this section we present the Bag-of-Visual Words model, on which the presented approach is based. Also, we describe the spectrogram generation process and we present the proposed emotion recognition methodology.

### A. The BoVW Model

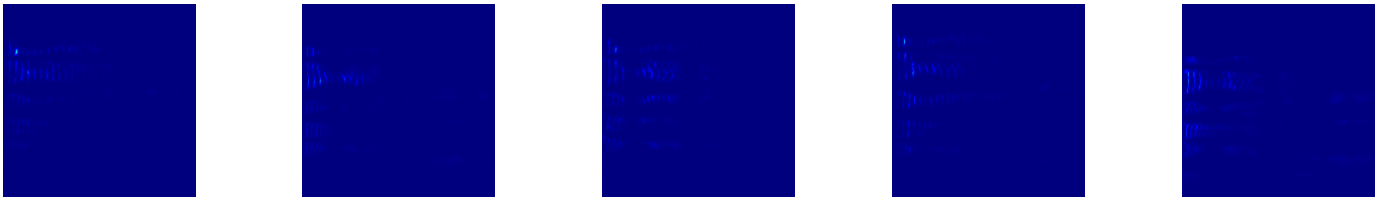
The origin of the Bag-of-Words (BoW) model goes back to the 1950s and the field of text document analysis [27], where the main idea was to describe a (part of a) text document using a histogram on word frequencies. During the 2000s it was adopted accordingly to fit the needs of several computer vision-related tasks such as concept detection, image classification and object/scene recognition problems [28] and is referred to as the "Bag-of-Visual Words" (BoVW) model.

BoVW is a weakly supervised model, built upon the notion of visual vocabularies. A visual vocabulary is actually a set of "exemplar" image patches, which are commonly referred to as "visual words." Using such a vocabulary, a given image may be described based on these words. To built an appropriate visual vocabulary, one should use a large corpus of representative images of the domain of interest, so that they would be closely related to the problem at hand. Typically, a clustering approach such as the well-known k-means algorithm is applied on the extracted features. The centroids (or in some cases the medoids) are then selected as the words that comprise the visual vocabulary. The size $N$ of the visual vocabulary is often determined heuristically. Note that the visual vocabulary acts as a means of quantization of the feature space which consists of the locally-extracted descriptors, which are accordingly quantized to their nearest word.

More specifically, any given image is described by a feature vector (histogram) consisting of frequencies of visual words from the vocabulary within it. To this goal, features are extracted using the exact method that was used during the vocabulary construction process. Each feature is then translated (coded) to the most similar visual word of the vocabulary, using an appropriate similarity metric such as the Euclidean distance. This way, a histogram of visual words is extracted and is used for the description of the whole image and is represented using a $1 \times N$-dimensional feature vector; each component of the feature vector corresponds to a visual word, while its value denotes the appearance frequency of this word within the whole image. Note that BoVW provides a fixed-size representation of the image, i.e., independent of the number of features that have been originally extracted. This property is important, since in several feature extraction approaches such as salient point extraction, the number of feature varies depending on the image content which often makes their use difficult. Contrast to this case, the fixed size feature vectors generated by BoVW may be easily used to train several well-known classifiers and models such as neural networks and support vector machines.

### B. Spectrogram Generation

For the generation of spectrograms we first extract a single segment of length $t_s$ sec from any given audio sample. This segment is randomly selected from the entire sample. Then, we apply the short-time Fourier transform (STFT) on the original signal. We use short-term windows of fixed size $t_w$ and step $t_s$. Pseudocolored images of spectrograms from 5 emotions that have been used within the experimental evaluation of this work are illustrated in Fig. 1.



(a) Anger (b) Fear (c) Happiness (d) Neutral (e) Sadness

Fig. 1: Example spectrogram images per emotion that have been generated from the EMOVO [32] dataset (Figure best viewed in color).

### C. Emotion Recognition using BoVW

For the extraction of visual features from spectrograms, we chose to follow a grid-based approach. Contrary to salient point approaches that first extract a set of points and extract descriptions from patches that surround them, a grid-based approach is based on a rectangular grid. Each crossing is used to define a point (pixel) of interest. More specifically, we use a regular, i.e., square grid. From each resulting pixel we extract the well-known and widely used Speeded-Up Robust Features (SURF) [29] which combine fast extraction speed and robustness to several transformations and illumination change and have been adopted in many real-life computer-vision problems. Note that in principle, SURF is both a salient point extraction and description scheme, yet we only use the latter. Although SURF salient regions have been effectively used into BoVW problems [30], in our case the extracted number was rather small, and insufficient to produce effective histograms. Moreover, it has been demonstrated in [31] that the regions that result from grid-based image sampling may also carry useful information, which is in general adequate to provide a content description that can be used in classification schemes. For classification, we trained support vector machines on the extracted feature vectors. In Fig. 2 we illustrate a visual overview of the whole process, i.e., from raw speech to the final recognized emotion.

## IV. Experiments

In this section we describe the artificial datasets that we have used and also the real-life one that has been created within a classroom experiment. Moreover, we present several implementation details and also the results of the performed experiments.

### A. Artificial Datasets

The first part of our experimental evaluation consisted of experiments using 3 widely known and freely available datasets of 3 different languages. More specifically, we used: a) EMOVO [32] which is an emotional speech dataset in Italian. 6 actors performed 14 sentences, representing *disgust*, *fear*, *anger*, *joy*, *surprise* and *sadness*; b) SAVEE [33] which is a larger emotional speech dataset in English. 4 actors performed 15 sentences per emotion, representing the same emotions as in EMOVO; and c) EMO-DB [34] which is an emotional speech dataset in German. 10 male and 5 female actors performed 493 sentences in total, representing *anger*, *boredom*, *disgust*, *fear*, *happiness*, *sadness* and *neutral*. For our task, we chose 5 of the common emotion classes, namely *Happiness*, *Sadness*, *Anger*, *Fear* and *Neutral*.

### B. Real-life Classroom Experiment

The aforementioned real-life classroom experiment was conducted in a computer laboratory of a middle school in Corfu, Greece. It involved 24 students (15 male and 9 female), aging between 12-13 years old and their ICT instructor. Note that all students were familiar with their instructor. This way, we expected that their reactions and their facial and vocal expressions were authentic and not restrained, as it would have been with a total stranger. Students were randomly divided into two teams. Each team was assigned with a different task, yet both tasks were quite similar and involved building and programming of an educational robot. More specifically, the well-known LEGO Mindstorm EVE 3 Educational kit[1] was used. Programming involved both the Lego Mindstorm software and Scratch.[2]

The task of the first team was to program the robot so that it could be controlled using a mobile phone. The task of the second team was to program the robot to follow a predefined colored route, using color sensors. The role of the instructor was to carefully observe and document the students' reactions and facial expressions, at an effort to assess their emotions. Note that the instructor was not interfering at any of the two tasks. Upon the completion of both tasks, each student was separately interviewed by the instructor. She/he was asked to freely describe her/his experience and opinion regarding the task she/he was not involved. All students were encouraged to express their real feelings. Also, note that they have all volunteered to take this course among several other options and were in general satisfied with the use of Mindstorm kits in many previous activities. The voice of each student was recorded with a computer microphone and annotated based on both her/his vocal and facial appearance, according to the instructor's experience. Special care had been taken so that all recordings took place within an environment free of ambient noises. Also, recordings were post-processed in order to remove parts with silence and/or the voice of the instructor. This way, we ended up with 42 recordings, with average

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

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

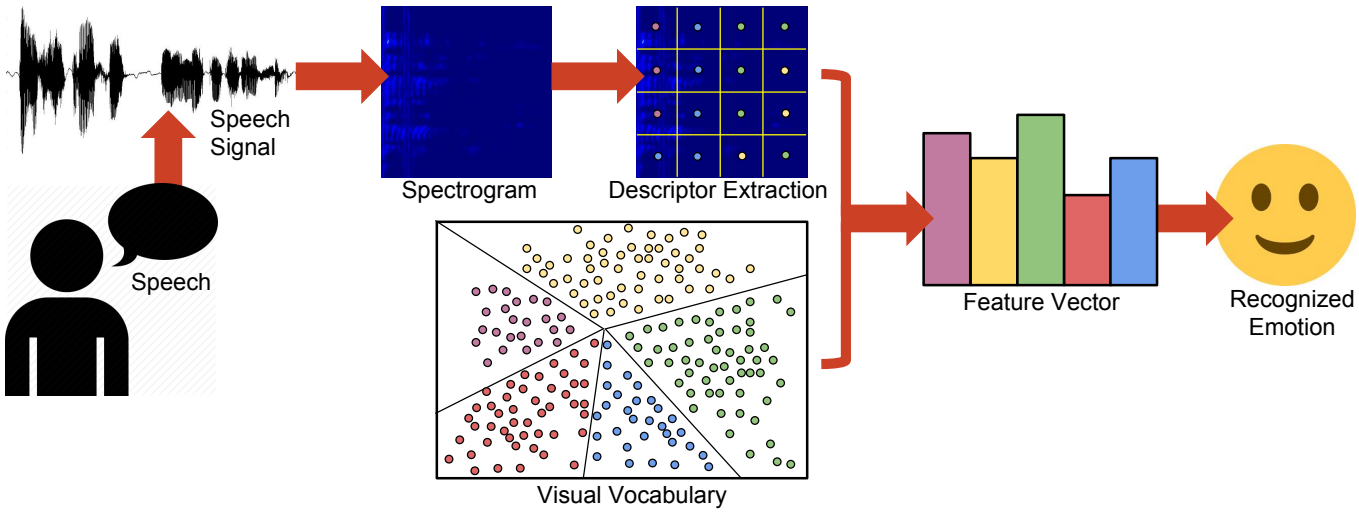



Fig. 2: A visual overview of the proposed emotion recognition scheme (Figure best viewed in color).

duration 7.8 sec. Upon the annotation process, the dataset consisted of 24 samples with a positive emotion, 8 with a negative and 10 with a neutral. We will refer to this dataset as "kids."

### C. Results

In both cases, we experimented with segments with $t_S = 1$ and $T_S = 2$ sec, $t_w = 40$ msec and step $t_s = 20$ msec. For the vocabulary size, we used $N = 100, 200, \ldots, 1500$. The BoVW model has been implemented using the Computer Vision Toolbox of Matlab R2016a [36]. We used a grid of size $8 \times 8$. We opted to create a balanced dataset, i.e., in each dataset, all 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. The spectrograms have sizes $227 \times 227$ px and have been extracted using the pyAudioAnalysis open source Python library [35].

We compared our approach with a baseline one (which will be referred to as "baseline 1") that used, an SVM classifier using an early fusion approach on standard features. More specifically we used Histograms of Oriented Gradients (HOGs) [37], Local Binary Patterns [38] and histograms of color coefficients. Feature extraction has been performed in a $2 \times 2$ grid rationale, i.e., all features have been computed on 4 grids and then the resulting feature vectors were merged to a single feature vector that represents the whole image. We also compared our approach to one that was based on early fusion of several short-term audio features (which will be referred to as "baseline 2"), i.e., features extracted on the speech signal. These features have been extracted using [35] and are zero-crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, mel frequency cepstral coefficients, a chroma vector and the chroma deviation. As it may be observed in Table I in almost all cases the proposed BoVW scheme outperformed the baseline approaches.

TABLE I: F1 score for all datasets.

| | *proposed* | *baseline 1* | *baseline 2* |
|---|---|---|---|
| **EMOVO** | **0.63** | 0.42 | 0.45 |
| **SAVEE** | **0.54** | 0.32 | 0.30 |
| **EMO-DB** | 0.74 | 0.68 | **0.80** |
| **KIDS** | **0.83** | 0.73 | 0.75 |

## V. Conclusions

In this paper we presented an approach for recognizing the emotional state of humans, relying only on audio information. More specifically, our approach extracts non-linguistic information from spectrograms that represent audio segments. From each spectrogram we overlaid a sampling grid and extracted a set of interest points. From each, we extracted SURF features and used them to train a BoVW model, which was then used for emotion classification with SVMs. We evaluated the proposed approach on 3 publicly available artificial datasets. Furthermore, we created a dataset that consisted of real-life recordings from students. In almost all cases, our approach outperformed two baseline approaches, one that also worked on visual features from spectrograms and another that relied on audio spectral features. Note that the proposed approach does not rely at any step on spectral features. As a next step, we plan to perform multilingual experiments and further use our approach in other applications such as audio content classification and retrieval.

## 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] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, *Emotion recognition in human-computer interaction*. IEEE Signal processing magazine, 18(1), 32-80, 2001.

[2] T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson and R. Picard, *Real-time inference of mental states from facial expressions and upper body gestures*. In Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE International Conference on (pp. 909-914). IEEE, 2011.

[3] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, *Real-time automatic emotion recognition from body gestures*. arXiv preprint arXiv:1402.5047, 2014.

[4] 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, 2017.

[5] S. G. Koolagudi and K. S. Rao. *Emotion recognition from speech: a review.* International journal of speech technology, 15(2), 99-117, 2012.

[6] A. Haag, S. Goronzy, P. Schaich, and J. Williams, *Emotion recognition using bio-sensors: First steps towards an automatic system*. In Tutorial and research workshop on affective dialogue systems (pp. 36-48). Springer, Berlin, Heidelberg, 2004.

[7] 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), 155-177, 2015.

[8] M. El Ayadi, M. S. Kamel and F. Karray, *Survey on speech emotion recognition: Features, classification schemes, and databases.* Pattern Recognition, 44(3), 572-587, 2011.

[9] L. Fei-Fei and P. Perona, *A bayesian hierarchical model for learning natural scene categories*. In Computer Vision and Pattern Recognition, IEEE Computer Society Conference on (Vol. 2, pp. 524-531). IEEE, 2005.

[10] E. Spyrou, T. Giannakopoulos, D. Sgouropoulos and M. Papakostas, *Extracting emotions from speech using a bag-of-visual-words approach.* In Semantic and Social Media Adaptation and Personalization (SMAP), 12th International Workshop on (pp. 80-83). IEEE, 2017.

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

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

[13] A. Hanjalic, *Extracting moods from pictures and sounds: Towards truly personalized TV*. IEEE Signal Proc. Magazine, 23(2), pp.90–100, 2006.

[14] L. Lu, D. Liu and H.J. Zhang, *Automatic mood detection and tracking of music audio signals*. IEEE Trans. on audio, speech, and language processing, vol. 14(1), pp.5–18, 2006.

[15] Y.H. Yang, Y.C. Lin, Y.F. Su, and H.H. Chen, *A regression approach to music emotion recognition*. IEEE Trans. on audio, speech, and language processing, vol.16(2), pp.448-457, 2008.

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

[17] 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, vol. 2008, pp. 597–600. 2008.

[18] 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.

[19] R. Plutchik and H. Kellerman (eds.),*Emotion: theory, research and experience. Vol. 1, Theories of emotion*. Academic Press, 1980.

[20] H. Lee, P. Pham, Y. Largman and A.Y. Ng, *Unsupervised feature learning for audio classification using convolutional deep belief networks*. In Advances in neural information processing systems, 2009.

[21] T. Zhang and C.C.J. Kuo, *Audio content analysis for online audiovisual data segmentation and classification*, IEEE Trans. on speech and audio processing, vol.9(4), pp.441–457, 2001.

[22] L. He, M. Lech, N. Maddage and N. Allen, *Stress and emotion recognition using log-Gabor filter analysis of speech spectrograms*. In Proc. of Int'l Conf. on Affective Computing and Intelligent Interaction, IEEE, 2009.

[23] Q. Mao, M. Dong, Z. Huang and Y. Zhan, *Learning salient features for speech emotion recognition using convolutional neural networks*. IEEE Transactions on Multimedia, 16(8), pp.2203–2213, 2014.

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

[25] 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.

[26] 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.

[27] Z.S. Harris, *Distributional structure*. Word, 10(2–3), pp.146–162, 1954.

[28] J. Sivic and A. Zisserman, *Efficient visual search of videos cast as text retrieval*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 31(4), pp. 591606, 2009.

[29] 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.

[30] A. Alfanindya, N. Hashim and C. Eswaran, *Content Based Image Retrieval and Classification using speeded-up robust features (SURF) and grouped bag-of-visual-words (GBoVW)*. In Technology, Informatics, Management, Engineering, and Environment (TIME-E), International Conference on (pp. 77-82). IEEE, 2013.

[31] T. Tuytelaars, *Dense interest points*, in proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[32] G. Costantini, I. Iaderola, A. Paoloni and M. Todisco, *EMOVO Corpus: an Italian Emotional Speech Database*. In Proc. of the Ninth Int'l Conf. on Language Resources and Evaluation (LREC), 2014.

[33] S. Haq, P.J. Jackson and J. Edge, *Speaker-dependent audio-visual emotion recognition*. In Proc. of the Int'l Conf. on Audio-Visual Speech Processing, 2009.

[34] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier and B. Weiss, *A database of German emotional speech*. In Interspeech, vol.5, pp. 1517–1520, 2005.

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

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

[37] N. Dalal and B. Triggs. *Histograms of oriented gradients for human detection.* In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[38] Z. Guo, L. Zhang, and D. Zhang. *A completed modeling of local binary pattern operator for texture classification.* IEEE Trans. on Image Processing 19(6), pp.1657–1663, 2010.