Recognizing Emotional States Using Speech Information

Michalis Papakostas, Giorgos Siantikos, Theodoros Giannakopoulos, Evaggelos Spyrou, and Dimitris Sgouropoulos

Abstract Emotion recognition plays an important role in several applications, such as human computer interaction and understanding affective state of users in certain tasks, e.g., within a learning process, monitoring of elderly, interactive entertainment etc. It may be based upon several modalities, e.g., by analyzing facial expressions and/or speech, using electroencephalograms, electrocardiograms etc. In certain applications the only available modality is the user's (speaker's) voice. In this paper we aim to analyze speakers' emotions based solely on paralinguistic information, i.e., not depending on the linguistic aspect of speech. We compare two machine learning approaches, namely a Convolutional Neural Network and a Support Vector Machine. The former is trained using raw speech information, while the latter is trained on a set of extracted low-level features. Aiming to provide a multilingual approach, training and testing datasets contain speech from different languages.

Keywords Emotion recognition • Speech information • Convolutional neural networks • Transfer learning • Support vector machines

1 Introduction

It is common sense that the basic means of human communication is the vocalized speech. Apart from meaning, speech also carries emotions. Although the latter are more easily recognized through visual channels, e.g. facial features, gestures,

M. Papakostas
Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX, USA
e-mail: michalis.papakostas@mavs.uta.edu

G. Siantikos • T. Giannakopoulos • E. Spyrou (✉) • D. Sgouropoulos
Institute of Informatics and Telecommunications, National Center for Scientific Research–"Demokritos", Athens, Greece
e-mail: dickos@iit.demokritos.gr; tyiannak@iit.demokritos.gr; espyrou@iit.demokritos.gr; dsgou@iit.demokritos.gr

P. Vlamos (ed.), *GeNeDis 2016*, Advances in Experimental Medicine and Biology 989, DOI 10.1007/978-3-319-57348-9_13

etc., in many practical applications, e.g. in human-computer interaction through voice-user interfaces (VUIs), speech may be the only available modality for emotion recognition. The latter comprises probably the most challenging speech-related task, e.g., when compared to automatic speech recognition (ASR), speaker identification etc.

In general, one may argue that speech carries two distinct types of information [1]: *explicit* or linguistic information, which concerns articulated patterns by the speaker; and *implicit* or paralinguistic information, which concerns the variation in pronunciation of the linguistic patterns. The former may be qualitatively described, while the latter may be quantitatively measured, using certain spectral features and also features such as the pitch, the intensity etc. Using either or both types of information, one may attempt to classify an audio segment that consists of speech, based on the emotion(s) it carries. However, emotion recognition from speech appears to be a significantly difficult task even for a human, no matter if he/she is an expert in this field (e.g. a psychologist).

Many approaches are assisted by ASR aiming to fuse linguistic and paralinguistic information. The main disadvantage of these is that they are not able to provide language-independent models. Of course, another disadvantage is that there exists a plethora of different sentences, speakers, speaking styles and rates [2]. Thus, most approaches that aim to be language-independent tend to rely on paralinguistic speech information. Nevertheless, even in this case, such information may be significantly diverse, depending on cultural particularities. Additionally, a speaker's potential chronic emotional state may suppress the expressiveness of several emotions. Still, relying solely on paralinguistic information is probably the most appealing approach, when dealing with speakers emotion recognition.

Typically, the set of extracted features is labelled by an expert and learned using a machine learning approach, e.g., the well-known Support Vector Machines (SVMs), which shall be discussed later. However, during the last few years a new trend in the field of machine learning are the "Deep Neural Networks." One of their main advantages over traditional approaches is that they do not need to be trained using specific features. Instead, they translate raw data into compact intermediate representations, while they remove any redundancy. In this work we aim to compare both the aforementioned approaches. We will investigate whether a Convolutional Neural Network (CNN) (i.e., a deep learning approach) is able to replace the traditional approach of feature extraction, model training and classification and render paralinguistic features obsolete. Also we shall use several datasets from various languages, at an attempt to provide a language-independent approach.

The remaining of this paper is as follows: In Sect. 2 we present related work within the broader research area of emotion recognition from speech. In Sect. 3 we provide theoretical background concerning SVMs and CNNs. The application of the aforementioned techniques for the problem at hand is discussed in Sect. 4. Experimental results are presented and discussed in Sect. 5. Finally, conclusions are drawn in Sect. 6, where plans for future work are also discussed.

2 Related Work

Besides extracting information regarding events, structures (e.g., scenes, shots) or genres, a substantial research effort of several multimedia characterization methods has focused on recognizing the *affective* content of multimedia material. These methods try to map low-level audio-visual features to the *emotions* that underlie the respective multimodal information [3–5]. Automatic recognition of emotions in multimedia content can be very important for various multimedia applications. For example, recognizing affective content of music signals [6, 7] can be used in a system, where the users will be able to retrieve musical data with regard to affective content.

The most common approach to affective audio content recognition, so far, is to apply well-known classifiers (Hidden Markov Models, Support Vector Machines, etc.) for classifying signals into an *a-priori known number of distinct* categories of emotions, e.g., fear, happiness, anger [5, 8]. An alternative way to emotion analysis is the dimensional approach, according to which, emotions can be represented using specific dimensions that stem from psychophysiology [7, 9–11]. In [10], Valence-Arousal representation is used for affective video characterization. Towards this end, visual cues, such as motion activity, and simple audio features, e.g., signal energy are used for modelling the emotion dimensions. Finally, in [12] an SVM regressor has been used to recognize valence and arousal in speech segments from movies.

3 Machine Learning Approaches

3.1 Support Vector Machines

Support Vector Machines (SVMs) [13] are well-known supervised learning models, which have been extensively used in classification and regression problems. Their goal is to find the optimal hyperplane separating data in a feature space. More specifically, an SVM model is built in a way that the margin between the mappings of the examples of the categories is maximized. Then, a hyperplane is constructed, which is used to separate (i.e., classify) unknown examples, based on the side they fall on. Although they are linear models, using appropriate kernels (i.e., a technique called “kernel trick”) they are able to handle non-linearly separable data in features spaces of higher dimensionality than the one of the original problem.

3.2 Convolutional Neural Networks

During the last few years, deep neural networks have lead to breakthrough results on a variety of pattern recognition problems. Research fields such as computer vision

[14] and voice recognition [15] have benefitted. One of the most recognizable and effective deep architectures is an architecture called Convolutional Neural Network (CNN) [16].

Briefly, CNNs may be regarded as a special type of neural network that uses many identical copies of the same neuron. This allows the network to comprise of a significantly large number of neurons, thus being able to express computationally large models, also offering the advantage of keeping the number of actual network parameters (i.e., the set of values describing the neurons' behaviour) that actually need to be learned, relatively small.

CNNs can consist of an arbitrary number of layers depending on both the application and the choices of the designer of the network. These layers can be categorized into three distinct types:

- *Convolutional*: these layers consist of a rectangular grid of neurons. This requires that the previous layer is also a rectangular grid of neurons. Each neuron takes inputs from a rectangular section of the previous layer; the weights for this rectangular section are the same for each neuron in the convolutional layer. Thus, the convolutional layer is just an image convolution of the previous layer, where the weights specify the convolution filter. These weights are the parameters of the network and they are shared among multiple neurons as it has been previously mentioned. In addition, there may be several grids in each convolutional layer; each grid takes inputs from all grids in the previous layer, using potentially different filters.
- *Max-Pooling*: After each convolutional layer, a pooling layer may follow. This layer type takes small rectangular blocks from the convolutional layer and subsamples it to produce a single output from that block. There are several ways to do this pooling, such as taking the average, the maximum, or a learned linear combination of the neurons in the block. In this work, pooling layers will always be max-pooling layers, i.e., they shall take the maximum of the block they are pooling.
- *Fully-Connected*: Finally, after several convolutional and max-pooling layers, the high-level reasoning in the CNN is actuated via fully connected layers. A fully connected layer takes all neurons in the previous layer (be it fully connected, pooling, or convolutional), connecting them to each of its single neurons. Fully connected layers are not spatially located anymore (you can visualize them as one-dimensional), so there can be no convolutional layers after a fully connected layer.

4 Proposed Methodology

4.1 Data Representation

Each audio sample (speech segment) is represented through the respective spectrogram. A spectrogram is a frequency-time representation of the signal, stemming

from the application of the short-time Fourier Transform on the original signal. The spectrogram representation when as input to the Convolutional Neural Network, is therefore handled as image. A 20 ms window length and 15 ms window step has been adopted (i.e. 25% overlap). The spectrograms have been extracted using the pyAudioAnalysis opensource Python library[1] [17]. Figure 1 illustrates an example of a spectrogram for each of the 4 classes that are included in the experiments, taken from the SAVEE dataset [18]. The spectrogram is always resized to 227×227 size to "fit" the adopted network structure. This enforces a default segment length. For varying signal lengths, the CNN is applied to several overlapping mid-term segments and a post-processing step is also applied to merge temporally successive decisions.

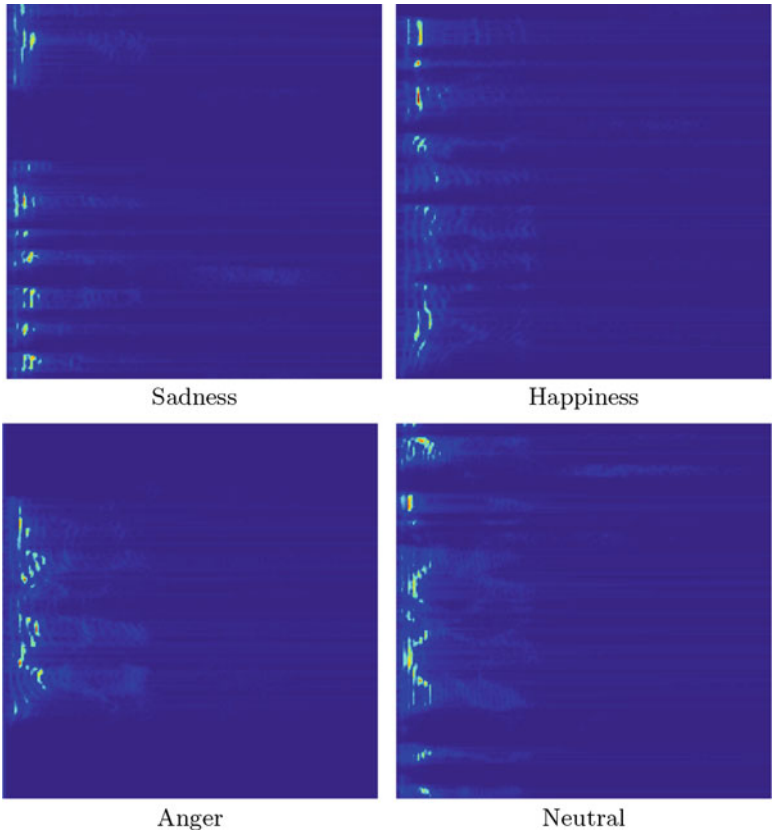

Fig. 1 Example spectrograms from each of the 4 included classes

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

Table 1 Adopted short-term audio features

Index	Name	Description
1	Zero crossing rate	Rate of sign-changes of the frame
2	Energy	Sum of squares of the signal values, normalized by frame length
3	Entropy of energy	Entropy of sub-frames' normalized energies. A measure of abrupt changes
4	Spectral centroid	Spectrum's center of gravity
5	Spectral spread	Spectrum's second central moment of the spectrum
6	Spectral entropy	Entropy of the normalized spectral energies for a set of sub-frames
7	Spectral flux	Squared difference between the normalized magnitudes of the spectra of the two successive frames
8	Spectral rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated
9–21	MFCCs	Mel frequency cepstral coefficients: a cepstral representation with mel-scaled frequency bands
22–33	Chroma vector	A 12-element representation of the spectral energy in 12 equal-tempered pitch classes of western-type music
34	Chroma deviation	Standard deviation of the 12 chroma coefficients

4.2 SVM-Based Classification

The audio signal is transformed to a sequence of feature vectors which are used for the SVM models' training. Features are extracted from 20 ms windows and afterwards the final feature vectors are formed by concatenating the mean and variance values of the features over a mid-term window of 1 s. Table 1 summarizes the features that are computed. There are 34 features in total which result in a 68-dimensional feature vector for each mid-term window. An SVM with RBF kernel has been used in our experiments.

4.3 CNN-Based Classification

For recognizing the emotions, we decided to utilize a CNN classifier that performs upon pseudo-RGB-colored frequency images. As recent literature has shown, deep hierarchical visual feature extractors can significantly outperform shallow classifiers trained on hand-crafted features and are more robust and generalizable when countering problems that include significant levels of inherent noise. The architecture of our deep CNN was initially proposed in [19]. The model is mainly based on the Caffenet [20] reference model, which is similar to the original AlexNet [14] and the network proposed in [21]. For our experiments we used the *BVLC Caffe*[2] deep-learning framework.

[2]https://github.com/BVLC/caffe.

The network architecture consists of two convolution layers with stride of 2 and kernel sizes equal to 7 and 5, respectively, followed by max pooling layers. As a next step, a convolution layer with three filters of kernel size equal to 3 is applied, followed again by a max pooling layer. The next two layers of the network are fully connected layers with dropout, followed by a fully connected layer and a softmax classifier, that shapes the final probability distribution. All max pooling layers have kernel size equal to 3 and stride equal to 2. For all the layers we used the ReLu as our activation function. The output of the network is a distribution on our three target classes, while the output vector of the semifinal fully connected layer has size equal to 4096. The initial learning rate of 0.001, which decreases after 700 epochs by a factor of 10.

Since training a new CNN from scratch would require big loads of data and high computational demands, we used transfer learning to fine-tune the parameters of a pre-trained architecture. The notion of "transfer" learning refers to the transfer of knowledge from a "learned" task (source task) at a given domain (source domain) to a related, yet unsolved task (target task), at the same or possibly at another domain (target domain), aiming to improve the learning process [22].

The original CNN was trained on the 1.2 M images of the ILSVRC-2012 [23] classification training subset of the ImageNet [24] dataset. Following this approach, we managed to decrease the required training time and to avoid overfitting our classifier, by ensuring a good weight initialisation, given the relatively small amount of available data. Finally, the data are preprocessed by augmenting the frame dimensionality to 240×320. The input to the network corresponds to the 227×227 center crops and their mirror images.

5 Experimental Results

For training and evaluation of the aforementioned approaches, we used 3 widely known emotional speech datasets, all of which are freely available from their authors. More specifically, these datasets are:

- **EMOVO** [25] is an emotional speech corpus, containing speech in Italian language from 6 actors who performed 14 sentences. The emotions represented here are *disgust*, *fear*, *anger*, *joy*, *surprise* and *sadness*.
- **SAVEE** [18] is a larger dataset, since besides speech, it contains video of the participating actors while expressing the same 6 emotions as in the EMOVO case. The data consists of 15 TIMIT sentences per emotion played by 4 English male speakers.
- **EMO-DB** [26] is a German acted database, consisting of 493 utterances performed by 10 (5 male and 5 female) actors expressing the emotions of *anger*, *boredom*, *disgust*, *fear*, *happiness*, *sadness* and *neutral*.

Table 2 Accuracy (%) when using each single dataset for testing

Test dataset	SVM	CNN
SAVEE	30	25
EMOVO	45	29
EMO-DB	**80**	51

Each time, training has been performed using both the remaining datasets. Overall best performance is indicated in bold.

For our task, we chose 4 of the common emotion classes, namely *Happiness*, *Sadness*, *Anger* and *Neutral*. A major difficulty resulting from the choice of datasets, is the great differences between languages, since besides the linguistic differences, there is also big variability in the way each emotion is expressed. For each classification method 3 different experiments were carried out where a single dataset is used for testing and the remaining 2 for training.

Accuracies of all experiments are summarized in Table 2. Best result was achieved when using SAVEE and EMOVO databases for training and EMO-DB for testing. In all cases the SVM outperformed the CNN. We feel that the poor classification performance of the CNN was due to the lack of generalization, which is due to the unsuccessful transfer of learning at a different domain.

6 Conclusions and Future Work

In this paper we investigated the use of deep learning in emotion recognition from speech, aiming to assess whether it may be used over traditional machine learning approaches. More specifically, we trained both a Convolutional Neural Network on raw spectrograms and a Support Vector Machine, using a set of low-level features. As it is indicated by the experimental results, the performance of the SVM was by far superior. Moreover, the SAVEE and EMOVO databases have proven to be adequate, when tested when the EMO-DB database. However, the overall conclusion is that multilingual emotion recognition remains one of the most challenging problems. Our plans for future work include the usage of more datasets for training and evaluation. We also aim to investigate other pre-trained deep learning networks, since we feel that deep learning may significantly contribute to the problem at hand. Finally, among our plans is to apply such approaches into real-life problems, e.g., emotion recognition within training and/or educational programs.

Acknowledgements 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. Anagnostopoulos, C.N., T. Iliou, and I. Giannoukos. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* 43(2):155–177.
2. El Ayadi, M., M.S. Kamel, and F. Karray. 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition* 44(3):572–587.
3. Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18:32–80.
4. Hanjalic, A. 2006. Extracting moods from pictures and sounds: towards truly personalized tv. *IEEE Signal Processing Magazine* 23:90–100.
5. Wang, Y., and L. Guan. 2008. Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia* 10:936–946.
6. Lu, L., D. Liu, and H. Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing* 14:5–18.
7. Yang, Y.-H., Y.-C. Lin, Y.-F. Su, and H. Chen. 2008. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing* 16:448–457.
8. Nogueiras, A., A. Moreno, A. Bonafonte, and J.B. Marino. 2001. Speech emotion recognition using hidden Markov models. In *Proceedings of Eurospeech*, 2679–2682.
9. Grimm, M., K. Kroschel, E. Mower, and S. Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication* 49(10–11):787–800.
10. Hanjalic, A., and X. Li-Qun. 2005. Affective video content representation and modeling. *IEEE Transactions on Multimedia* 7:143–154.
11. Wollmer, M., F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie. 2008. Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the 9th Interspeech*, 597–600.
12. Giannakopoulos, T., A. Pikrakis, and S. Theodoridis. 2009. A dimensional approach to emotion recognition of speech from movies. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 65–68. Piscataway, NJ: IEEE.
13. Vapnik, V. 1998. *Statistical Learning Theory*, vol. 1. New York: Wiley.
14. Krizhevsky, A., I. Sutskever, and G.E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.
15. Hinton, G., L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T.N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97.
16. Simard, P.Y., D. Steinkraus, and J.C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, vol. 3, 958–962.
17. Giannakopoulos, T. 2015. Pyaudioanalysis: an open-source python library for audio signal analysis. *PloS One* 10(12):e0144610.
18. Haq, S., and P. Jackson. 2009. Speaker-dependent audio-visual emotion recognition. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, Norwich, UK.
19. Donahue, J., L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634.
20. Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell. 2014. Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 675–678. New York: ACM.
21. Zeiler, M.D., and R. Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 818–833. Berlin: Springer.

22. Torrey, L., and J. Shavlik. 2010. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 242–264. Hershey, PA: IGI Global.
23. Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.
24. Deng, J., W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, 248–255. Piscataway, NJ: IEEE
25. Costantini, G., I. Iaderola, A. Paoloni, and M. Todisco. 2014. Emovo corpus: an Italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, ed. N.C.C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA).
26. Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of German emotional speech. In *Proceedings of Interspeech, Lissabon*, 1517–1520.