Realtime depression estimation using mid-term audio features

Theodoros Giannakopoulos, Christos Smailis, Stavros Perantonis, and Constantine Spyropoulos

National Center for Scientific Research DEMOKRITOS, Athens, Greece
{tyianak,xsmailis,sper,costass}@iit.demokritos.gr
http://www.iit.demokritos.gr

Abstract. This paper presents a method towards estimating a clinical depression-specific score, namely the Beck Depression Inventory (BDI) score, based on analysis of mid-term audio features. A combination of support vector machines and semi-supervised learning has been applied to map the mid-term features to the BDI score. The method has been evaluated on the AVEC 2013 depression dataset. The overall system has been implemented in Python achieving a $20x$realtime computational complexity on an average computer.

Keywords: Depression; Audio Analysis; Regression; Support Vector Machines

1 Introduction

One of the most common mood disorders is clinical depression. Therefore, automatic estimation or detection of its presence has gained research interest during the last years, through means of audio-visual signal analysis. This work focuses on the audio domain in order to estimate the Beck Depression Inventory (BDI) score of an individual, given a long audio recording of his/her voice.

Some of the previous studies in the field of depression analysis focus on utilizing acoustic features for the task of either predicting the Beck Depression Inventory (BDI) score of an individual, or classifying a person as depressive or not depressive. For example in [1] the performance of different audio features is explored for the task of classifying audio files contained within the AVEC 2013 dataset [2], as either depressed or nondepressed. In [3] several novel Canonical Correlation Analysis (CCA) based feature selection methods are presented, in order to reduce the massive dimensionality introduced for the AVEC 2013. Results revealed that using only 17% of original features, a relative improvement of 30% decrease of RMSE over the baseline on challenge test set was obtained. In [4] audio features that reflect changes in coordination of vocal tract motion have been utilized.

Other studies have adopted different approaches by attempting to fuse audio as well as video modalities in order to boost their performance. For example in [5] the authors made use of Motion History Histogram (MHH) in the video

modality, to capture the movement of each pixel (texture variation) within the face area. In the audio modality the authors used a set of spectral Low-Level Descriptors (LLD). MHH was once again used for extracting change information of the vocal expression. For each modality, the Partial Least Square (PLS) regression algorithm is applied and predicted values of visual and vocal clues were further combined at decision level. Another multimodal approach is presented in [6]. This study presents a multimodal approach using a GMM-UBM system with three different kernels for the audio subsystem and Space Time Interest Points in a Bag-of-Words approach for the vision subsystem. These are then fused at the feature level to form the combined AV system. Key results include the strong performance of acoustic audio features and the bag-of-words visual features in predicting an individuals level of depression.

2 Feature Extraction

In most audio analysis and processing methods, it is rather common that the input audio signal is divided into short-term frames before feature extraction. In particular, the audio signal is broken into (non-)overlapping frames and a set of features is extracted for each frame. The result of this procedure is a sequence of feature vectors per audio signal. Another common technique used as a second step in audio feature extraction is the processing of the feature sequence on a mid-term basis. The audio signal is first divided into mid-term segments and then, for each segment, the short-term processing stage is carried out. At a next step, the feature sequence, which has been extracted from a mid-term segment, is used for computing feature *statistics*. In practice, the duration of mid-term windows typically lies in the range $1 - 10$ secs, depending on the application domain. In this work, extensive experimentation has led to selecting a 5 second mid-term window and a 50 msec short-term frame. In both cases, 50% overlap has been adopted.

2.1 Short-term audio features

In this section we describe the adopted short-term features. These features have been used in several general audio analysis methods and speech processing applications [7], [8], [9] and they cover a wide range of audio signal properties achieving discrimination abilities in several classification and regression tasks.

Energy The energy feature is computed as a sum of squared signal values (in the time domain), normalized by the window length. Short-term energy usually exhibits high variation over successive speech frames, since speech signals contain weak phonemes and short periods of silence between words.

Zero Crossing Rate The Zero Crossing Rate (ZCR) of an audio signal is the rate of sign-changes of the signal divided by the duration of that signal. ZCR has been interpreted as measure of the noisiness of a signal, therefore it usually exhibits higher values in the case of noisy signals.

Entropy of Energy The entropy of energy is a measure of abrupt changes in the energy level of an audio signal. It is computed by firstly dividing each frame in sub-frames of fixed duration. Then, for each sub-frame j its energy is computed and divided by the total energy. Finally, the entropy of that sequence of (normalized) sub-energies e_j is computed as the final feature value. This feature has lower values if there exist abrupt changes in the energy envelop of the respective signal.

Spectral Centroid and Spread The spectral centroid and the spectral spread are two basic spectral domain features that quantify the position and shape. The spectral centroid is the center of gravity of the spectrum, while spectral spread is the second central moment of the spectrum.

Spectral Entropy Spectral entropy is computed in a similarly to the entropy of energy, however it is applied on the frequency domain.

Spectral Flux Spectral flux is a measure of spectral change between two successive frames and it is computed as the squared difference between the normalized magnitudes of the spectra of the two successive frames.

Spectral Rolloff Spectral rolloff is the frequency below which a certain percentage of the magnitude distribution of the spectrum is concentrated. It can be treated as a spectral shape descriptor of an audio signal and it has been used for discriminating between voiced and unvoiced sounds.

MFCCs The Mel-Frequency Cepstrum Coefficients (MFCCs) have been very popular in the field of speech analysis. In practise, MFCCs are the discrete cosine transform coefficients of the mel-scaled log-power spectrum. MFCCs have been widely used in speech recognition, speaker clustering and many other audio analysis applications.

Chroma vector This is a 12-dimensional representation of the spectral energy of an audio signal. This is a widely used descriptor, mostly in music-related applications, however it has also been used in speech analysis. The chroma vector is computed by grouping the spectral coefficients of a frame into 12 bins representing the 12 equal-tempered pitch classes of western-type music.

2.2 Final feature vector extraction

The process described in 2.1 leads to a sequence of short-term feature vectors of 21 dimensions (this is the total number of short-term features described above). As a next step, statistics are calculated in a mid-term basis as described in the beginning of this Section. In particular, the following statistics are computed:

(a) Average value μ, (b) Standard deviation σ^2 and (c) σ^2/μ ratio. This leads to several mid-term feature vectors of 63 elements. The number of these mid-term vectors depends on the overall duration of the audio signal. Each of these vectors are fed as input in the next regression step that produces the final depression estimate decision.

3 Depression Estimation

Each mid-term feature vector described in the previous section is used in the context of a Support Vector Machine regression technique to estimate the Beck Depression Index. Note that since one decision is calculated per feature vector, the final decision (per audio recording) is extracted by averaging the mid-term BDIs. This rationalle helps in generating a sufficient number of samples for the SVM regression model training phase and in addition it manages to handle mid-term vocal characteristics. Finally, a semi-supervised dimensionality reduction technique has been adopted in order to give weight to feature dimensions that are discriminative in terms of depression estimation. In particular, recordings that share similar BDIs have been grouped together using a simple k-Means clustering approach. In the sequel, a Linear Discriminant Analysis step has been performed on the initial feature space to extract a depression-discriminant subspace. During this process, the BDI clusters have been used as indices in the LDA process.

4 Experiments

4.1 Dataset

In order to evaluate the presented method, we have used the AVEC 2013 depression dataset [2]. This includes 340 video recordings of 292 subjects performing a human-computer interaction tasks while being recorded by a audio-visual sensors. The average age is 31.5 years and a range of 18 to 63 years. The length of each recording varies from 20 to 50 minutes, with an average duration of 25 minutes per recording. The total duration of all recordings is 240 hours. The behaviour within the clips consisted of different tasks such as: sustained vowel phonation; speaking loud while solving a task, counting from 1 to 10, read speech, singing, telling a story from the subject's own past, and telling an imagined story. The 16-bit audio was recorded at a sampling rate of 41KHz.

4.2 Results

A cross-validation procedure on the AVEC dataset has been conducted in order to evaluate the presented scheme via the following performance measures: (a) Root mean square error (RMSE) (b) 4-class classification accuracy: in order to calculate this measure, the final decisions are discretized to four "depression levels" according to the following categorization: 0-13: minimal depression, 14-19: mild depression, 20-28: moderate depression and 29-63: severe depression.

Then the overall accuracy of the classification task is computed. (c) 2-class classification accuracy: similarly to the 4-class case, the problem is binarized using $BDI = 14$ as a threshold. The cross-validation procedure estimated an RMSE of 9.5, 4-class accuracy rate equal to 52% and binary accuracy rate of 70%.

5 Conclusions and future work

We have presented a method for detecting a subject's clinical depression score using audio information. A wide range of audio features has been extracted in a mid-term basis, while a combination of SVMs and a semi-supervised dimensionality reduction method has been used in the recognition process. Results demonstrate a 20% drop in the error regarding the baseline performance of the adopted dataset. We conduct ongoing research on using temporal analysis techniques to enchance the signal representation process by selecting representative areas of the recording with high discrimination ability (in terms of depression analysis). In addition, we plan to implement visual shape modelling methods in the context of a fusion system. Finally, it is rather important to conduct collaboration with a specialist in mental health in order to extract useful correlations between low-level audio features and mid-level depression-related characteristics.

References

1. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: (A study of acoustic features for the classification of depressed speech)
2. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM (2013) 3–10
3. Kaya, H., Eyben, F., Salah, A.A., Schuller, B.: (Cca based feature selection with application to continuous depression recognition from acoustic speech features)
4. Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., Mehta, D.D.: Vocal biomarkers of depression based on motor incoordination. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM (2013) 41–48
5. Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM (2013) 21–30
6. Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., Epps, J.: Diagnosis of depression by behavioural signals: a multimodal approach. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM (2013) 11–20
7. Giannakopoulos, T., Pikrakis, A.: Introduction to Audio Analysis: A MATLAB Approach. Academic Press (2014)
8. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Fourth Edition. Academic Press, Inc. (2008)
9. Hyoung-Gook, K., Nicolas, M., Sikora, T.: MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. John Wiley & Sons (2005)