Automated depression analysis using convolutional neural networks from speech

By Lang He, Cui Cao

Review by Farhad Javanmardi

Introduction

- According to the World Health Organization (WHO), depression is the fourth most mental disorder by 2020.
- The effective treatments for depression can be aided by the detection of the problems at its early stages.
- The diagnosis of depression is mostly based on patient self-report or clinical judgments of symptom.
- Evaluations by clinicals depends on their expertise and diagnosis methods.

Introduction

- Depression, stress or emotion affect:
  - the process of speech production
    - Prosodic features
    - Vocal tract
    - Glottal source
- The voice patterns have a close relationship with depression.
- In recent years, deep learning methods utilized to predict depression severity by learning a lot of valuable information from the voice patterns.
- Among these different deep learning methods, Convolutional Neural Networks (CNN) has been widely used to achieve state-of-the-art performance in many communities, especially for
  - Audiovisual signals
- This paper explores how the depression severity prediction can benefit from the adoption of CNN in learning spectrogram patterns of the speech.

Databases utilized

- AVEC2013 depression database is a subset of the Audio-visual depressive language corpus (AVDLC).
  - 340 videos from 292 subjects (average duration of 25 min)
  - Performing a Human–Computer Interaction task while being recorded by a webcam and a microphone.
  - For this study, 150 videos from 82 subjects were used:
    - Training set: 50 recordings
    - Development set: 50 recordings
    - Test set: 50 recordings

- AVEC2014 depression database is a subset of the AVEC2013 corpus.
  - 300 videos (duration ranging from 6s to 4 min)
  - Performing two different Human–Computer Interaction tasks
  - For this study, they used 300 videos:
    - Training set: 100 recordings
    - Development set: 100 recordings
    - Test set: 100 recordings

- The depression levels were labeled per each video using Beck Depression Inventory-II (BDI-II).

Depression level	Label
0 -13	Minimal depression
14 - 19	Mild depression
20 - 28	Moderate depression
29 - 63	Severe depression

Methodology used

- Hand-crafted and deep-learned features used for estimating the severity of depression.
- For hand-crafted features:
  - Low level descriptors are extracted from the raw audio.
  - Median Robust extended Local Binary Patterns (MRELBP) features are extracted from the spectrograms of audio.
- For deep-learned features:
  - DCNN used to directly learn the deep-learned features from the the raw audio and spectrogram images.
- The joint fine-tuning method used to combine the four streams for the final depression prediction.

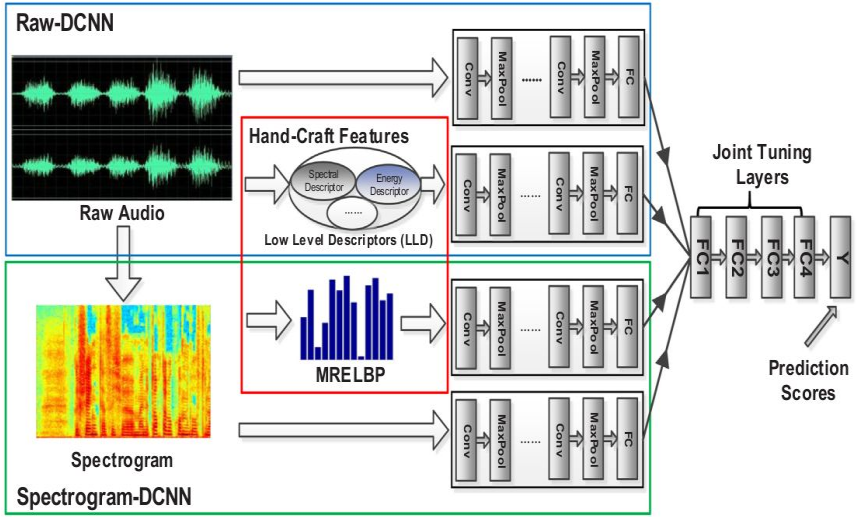


Fig. 1. Illustration of the proposed method for depression recognition using deep neural networks. The Raw-DCNN (Top) takes raw audio signals and low level descriptors (LLD) as input, while the Spectrogram-DCNN (Bottom) uses texture features as input. The red box in Fig. 1 is Hand-Crafted features. Other two arrows are Deep-Learned features. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from four DCNNs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Hand-crafted features

- For audio features:
  - 2268 feature vectors extracted by openSMILE toolkit.
    - 42 functionals on 32 energy and spectral related low-level descriptors (LLD).
    - 32 functionals on 6 voicing related LLD.
    - 19 functionals on 6 delta coefficients of the voicing related LLD.
    - 10 voiced/unvoiced durational features.

Table 1
38 low-level descriptors.

Energy&Spectral (32)
loudness (auditory model based), zero crossing rate, energy in bands from 250–650 Hz, 1 kHz-4 kHz, 25%, 50%, 75%, and 90% spectral roll-off points spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, flatness, MFCC 1–16
Voicing related (6)
F0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

Table 2: Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]Not applied to voicing related LLD.

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %–99 %, percentage of frames contour is above: minimum + 25%, 50%, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[1,3], standard deviation of segment length[1,3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude range of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1 – 5

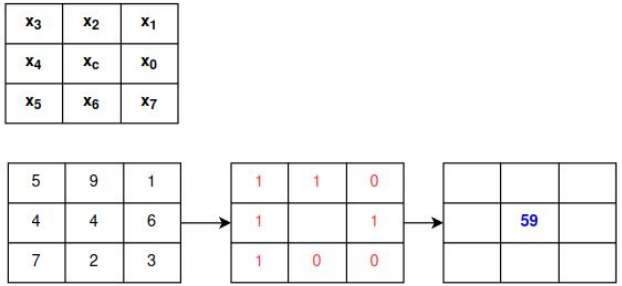
Tables source Valstar, Michel, et al. "Avec 2013: the continuous audio/visual emotion and depression recognition challenge." *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 2013.

Hand-crafted features

- The LBP is computed as:

$$LBP_{R,P}(x_c) = \sum_{n=0}^{P-1} s(x_{R,P,n} - x_c)2^n$$

$$s(x) = \begin{cases} 1, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases}$$



- $LBP(x_C) = s(6-4)2^0 + s(1-4)2^1 + s(9-4)2^2 + s(5-4)2^3 + s(4-4)2^4 + s(7-4)2^5 + s(2-4)2^6 + s(3-4)2^7$

  $= 1 \times 2^0 + 1 \times 2^2 + 1 \times 2^3 + 1 \times 2^4 + 1 \times 2^5 = 1 + 2 + 8 + 16 + 32 = 59$

- Median Robust Extended Local Binary Patterns (MRELBP) applied on spectrograms of the audio to extract the textural features.
- It uses a median filter to maximize the robustness of the representation to noise.

Deep-learned features

- Two different models used to extract
  - Deep-learned audio features from frame-level raw waveforms
  - Deep-learned texture features from spectrogram images
- For the deep learned audio features
  - The frame-level raw waveforms were fed to the first CNN convolutional layer to learn a filter-bank representation which is equivalent to filter kernels in a time-frequency representation.
  - The output feature map will have the same as the spectrogram
  - Parameters of the convolutional layer (stride, filter length and number of filters) corresponds to the parameters of spectrogram (mel-size, window size, number of mel-bands)

Deep-learned features

- For the deep learned texture features
  - Segment of 6s and 20s were used for the extraction of vocal patterns
  - A data augmentation method used to tackle with the small samples in training data
    - Original images (1)
    - Flipped images (1)
    - Rotated images with six angles (-15, -10, -5, 5, 10, 15) and their flipped versions (12)
    - 14 times more data than the original images obtained
  - The input image size 128 x 128
  - Filter size 5 x 5 with stride size of 1
  - For pooling layer, window size 2 x 2 with stride 2
  - Euclidean loss was used as the loss function

$$E = \frac{1}{2N} \sum_{i=1}^{N} \|\hat{y}_i - y_i\|^2$$

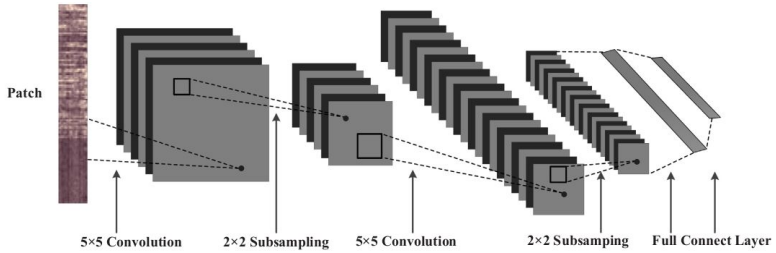


Fig. 2. The deep Convolutional Neural Network architecture.

Joint fine-tuning method

- To capture the complementary information within the two used models (Raw-DCNN and the Spectrogram-DCNN), joint fine-tuning Method used to boost the recognition performance.
- In the training process, the four DCNNs are trained separately, and then the joint fine-tuning is created using joint tuning layer
- The top layers retrained and other layers of the two trained networks were freezed.
- Euclidean loss function used.

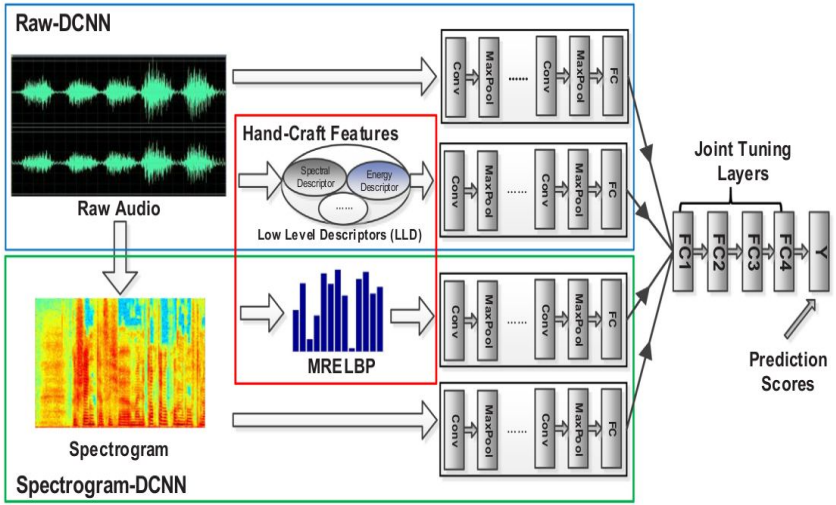


Fig. 1. Illustration of the proposed method for depression recognition using deep neural networks. The Raw-DCNN (Top) takes raw audio signals and low level descriptors (LLD) as input, while the Spectrogram-DCNN (Bottom) uses texture features as input. The red box in Fig. 1 is Hand-Crafted features. Other two arrows are Deep-Learned features. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from four DCNNs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Evaluation metrics

- The depression severity recognition performance is assessed with:
  - Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \tilde{y}_i|$$

  - Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}$$

- N denotes the number of data samples, y_i is the ground truth and y_i^ represents the predicted value of i-th sample

Results (Performance of single models)

Table 3
Performance of hand-crafted and deep-learned features on the development set and test set of AVEC2013.

Partition	Methods		RMSE	MAE
Dev.	Hand crafted model	LBP	9.3507	7.7314
		MRELBP	9.1673	7.5455
		LLD	9.3154	7.6502
	Deep learned model	Waveform	9.3896	7.8184
		Spectrogram	9.1129	7.5371
Test	Hand crafted model	LBP	10.9312	9.2443
		MRELBP	10.5611	8.6580
		LLD	10.6418	8.8935
	Deep learned model	Waveform	11.0983	9.4484
		Spectrogram	10.4561	8.4832

Table 4
Performance of hand-crafted and deep-learned features on the development set and test set of AVEC2014.

Partition	Methods		RMSE	MAE
Dev.	Hand crafted model	LBP	9.3478	7.5699
		MRELBP	9.1523	7.5026
		LLD	9.3000	7.5514
	Deep learned model	Waveform	9.3770	7.8813
		Spectrogram	9.1100	7.4969
Test	Hand crafted model	LBP	10.8211	8.7489
		MRELBP	10.4618	8.6420
		LLD	10.5648	8.6800
	Deep learned model	Waveform	10.9014	8.7810
		Spectrogram	10.4413	8.6014

- The deep-learned features obtained the better results on the test set for AVEC2013 and AVEC2014.
- Deep learned model can reduce some effort for finding suitable hand-crafted features for depression scale prediction.

Results (Overall performance by fusing the individual models and joint tuning)

- For fusing the individual models:
  - For AVEC2013:
    - RMSE: 10.2261
    - MAE: 8.2323
  - For AVEC2014:
    - RMSE: 10.1284
    - MAE: 8.2204
- Fusing the the hand-crafted and the deep-learned model showed better performance than the single model.
- For joint tuning method:
  - For AVEC2013:
    - RMSE: 10.0012
    - MAE: 8.2012
  - For AVEC2014:
    - RMSE: 9.9998
    - MAE: 8.1919
- The results implied that the proposed joint tuning method performance was improved when employing both the hand-crafted and deep-learned models.

Table 5
Overall performance on the development set and test set of AVEC2013.

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	9.1001	7.4456
	Hand & Deep (Joint Tuning)	9.0000	7.4210
Test	Hand & Deep (Ave.)	10.2261	8.2323
	Hand & Deep (Joint Tuning)	10.0012	8.2012

Table 6
Overall performance on the development set and test set of AVEC2014.

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	9.0089	7.4213
	Hand & Deep (Joint Tuning)	9.0001	7.4211
Test	Hand & Deep (Ave.)	10.1284	8.2204
	Hand & Deep (Joint Tuning)	9.9998	8.1919

Results (combined databases)

- RMSE : 9.8874
- MAE : 8.1901
- A potential reason for this is the new enlarged database has more data samples for training and the DCNN models can better predict the depression scores.

Table 7
Overall performance on the development set and test set (AVEC2013 + AVEC2014).

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	8.9971	7.4200
	Hand & Deep (Joint Tuning)	8.8920	7.4118
Test	Hand & Deep (Ave.)	10.0009	8.2323
	Hand & Deep (Joint Tuning)	9.8874	8.1901

Results (AVEC2014 Depression Challenge Results)

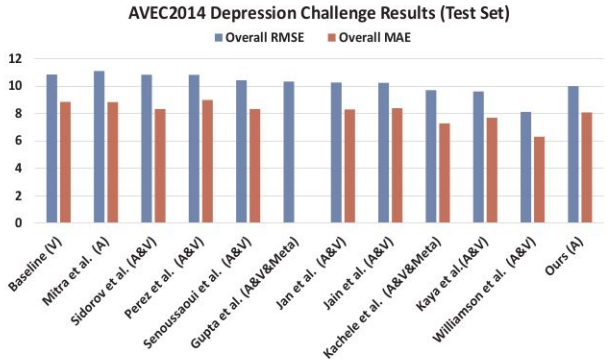


Fig. 4. AVEC2014 - Comparison with techniques of depression recognition using audio (A) and visual (V) features.

Table 9
AVEC2014 - comparison to state-of-the-art results. Note that the listed results use audio data only.

Partition	Methods	RMSE	MAE
Dev.	Baseline [15]	11.52	8.93
	Jain et al. [41]	11.51	9.75
	Jan et al. [40]	10.69	8.92
	Senoussaoui et al. [46]	10.09	7.41
	Parez et al. [43]	9.79	7.75
	Kachele et al. [45]	N/A	N/A
	Mitra et al. [48]	7.71	6.10
	Ours	***9.0001***	***7.4211***
Test	Baseline [15]	12.567	10.036
	Jain et al. [41]	10.25	8.40
	Jan et al. [40]	11.30	9.10
	Senoussaoui et al. [46]	12.71	9.82
	Parez et al. [43]	11.92	9.36
	Kachele et al. [45]	9.18	7.10
	Mitra et al. [48]	11.10	8.83
	Ours	***9.9998***	***8.1919***

- By using only audio features, the proposed methods provided comparable results to multi-modal approaches recognition.

References

[1] He, Lang, and Cui Cao. "Automated depression analysis using convolutional neural networks from speech." *Journal of biomedical informatics* 83 (2018): 103-111.

[2] Hafemann, Luiz G., Luiz S. Oliveira, and Paulo Cavalin. "Forest species recognition using deep convolutional neural networks." *2014 22Nd international conference on pattern recognition*. IEEE, 2014.

[3] Liu, Li, et al. "Median robust extended local binary pattern for texture classification." *IEEE Transactions on Image Processing* 25.3 (2016): 1368-1381.

Questions

1. Explain briefly the following terms:

- Convolutional layer
- Pooling layer
- Fully-connected layer
- Dropout method
- Activation functions

2. Explain briefly why deep learned model performs better than hand crafted model?