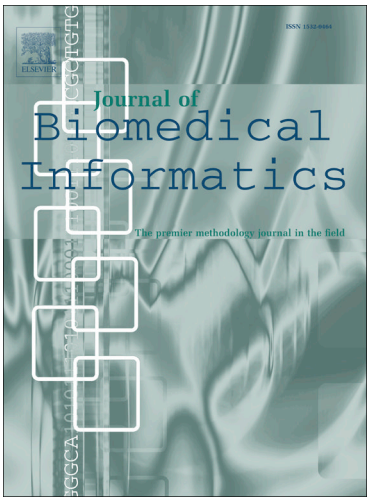

Automated Depression Analysis Using Convolutional Neural Networks from Speech

Lang He, Cui Cao

PII: S1532-0464(18)30090-X
DOI: https://doi.org/10.1016/j.jbi.2018.05.007
Reference: YJBIN 2977

To appear in: *Journal of Biomedical Informatics*

Received Date: 6 November 2017
Revised Date: 25 April 2018
Accepted Date: 12 May 2018

Please cite this article as: He, L., Cao, C., Automated Depression Analysis Using Convolutional Neural Networks from Speech, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.05.007

Automated Depression Analysis Using Convolutional Neural Networks from Speech

Lang He[a], Cui Cao[b]

[a]NPU-VUB joint AVSP Research Lab, School of Computer Science, Northwestern Polytechnical University (NPU), Xi'an, China
[b]Moscow Institute of Arts, Weinan Normal University, Weinan, China

Abstract

To help clinicians to efficiently diagnose the severity of a person's depression, the affective computing community and the artificial intelligence field have shown a growing interest in designing automated systems. The speech features have useful information for the diagnosis of depression. However, manually designing and domain knowledge are still important for the selection of the feature, which makes the process labor consuming and subjective. In recent years, deep-learned features based on neural networks have shown superior performance to hand-crafted features in various areas. In this paper, to overcome the difficulties mentioned above, we propose a combination of hand-crafted and deep-learned features which can effectively measure the severity of depression from speech. In the proposed method, Deep Convolutional Neural Networks (DCNN) are firstly built to learn deep-learned features from spectrograms and raw speech waveforms. Then we manually extract the state-of-the-art texture descriptors named median robust extended local binary patterns (MRELBP) from spectrograms. To capture the complementary information within the hand-crafted features and deep-learned features, we propose joint fine-tuning layers to combine the raw and spectrogram DCNN to boost the depression recognition performance. Moreover, to address the problems with small samples, a data augmentation method was proposed. Experiments conducted on AVEC2013 and AVEC2014 depression databases show that our approach is robust and effective for the diagnosis of depression when compared to state-of-the-art audio-based methods.

Keywords: Depression, Automatic diagnosis, Median Robust extended Local Binary Patterns(MRELBP), Speech processing

1. Introduction

Depression and anxiety disorders are highly prevalent worldwide, which have placed undue burden on individuals, families, and society. Studies suggest that effective treatments for depression can be aided by the detection of the problems at its early stages. According to the World Health Organization (WHO), depression will become the fourth most mental disorder by 2020 [1].

Depression is often difficult to diagnose because it manifests itself in different ways. The assessment methodologies for its diagnosis rely on subjective patient self-report or clinical judgments of symptom severity [2], [3]. The Hamilton Rating Scale for Depression (HAMD) [4] is currently the standard for depression severity estimation. It is worth noting that, evaluations by clinicians vary depending on their expertise and the used diagnosis methods, such as Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [5], the Quick Inventory of Depressive Symptoms-Self Report (QIDS) [6], the Beck Depression Inventory (BDI) [7], the 10-item Montgomery–Asberg Depression Rating Scale (MADRS) [8], the 9–item Patient Health Questionnaire (PHQ–9) [9], and the PHQ–8 [10].

In recent years, some machine learning methods have been proposed utilizing audio cues for depression analysis [11–16]. Meanwhile, there is a wealth of research, which suggests that voice patterns have a close relationship with emotion and stress [17–19]. In [20], the author suggested that the analysis of voice patterns can be divided into three primary categories, including prosodics, the vocal tract, and the glottal source. Although hand-crafted features have been proven to obtain better perfor-