So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Review of Depression Recognition Based on Speech

Xiaoyong LU[a*], Daimin SHI[b]
[a]*School of Psychology, Northwest Normal University, Lanzhou,China*
[b]*College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou,China*
*luxy@nwnu.edu.cn

Abstract: In recent years, the research on depression recognition has been widely concerned through machine learning or deep learning field. This paper reviews the relevant research on depression recognition using speech as biological indicators. The current challenges in this field are elaborated to further lead the future research direction.

Keywords: Phonetic features, depression, review

1. Introduction

Depression not only affects the physical and mental health of individuals, but also leads to high rates of disability and death (Pan et al., 2018). In recent years, it is common that objective indicators are used to assist in the diagnosis of depression. Speech is a sensitive indicator, slight physiological and cognitive changes may lead to significant changes in hearing (Schneider et al., 2012). At the same time, changes in respiratory muscles will affect the glottis. Both rhythm and glottic features have been proved to be affected by depression level (Moore et al., 2008). Therefore, the weakening of the rhythm, monotonous and slow pronunciation speed indicate depression (Hall et al., 1995).

This review is mainly divided into the following parts: First, makes a brief review on speech-based depression study at home and abroad; Secondly, describes the performance of rhythmic and spectral features of depression; Thirdly, description of two different research methods; Finally, some challenges provided for future research directions.

2. Research Status at Home and Abroad

In order to inspire future research directions, this section attempts to make a brief review of the relevant literature on depression diagnosis based on speech at home and abroad.

Compared with abroad, China's speech-based depression research started late. With the support of the National 973 Project, the Pervasive Computing Laboratory of Lanzhou University collects voice samples and analyzes them to identify depression. The Chinese Institute of Scientific Psychology uses machine learning approaches to establish an effective automatic recognition model of depression. The diagnosis of depression based on speech signals abroad is mainly the Audio Video Emotion Challenge Competition, which has been held for five consecutive years.

3. Acoustic Features of Depression Speech

Based on the results of psychological and prosody research, the most intuitive manifestation of the speaker's depression in speech is the prosody and spectral features.

3.1 Prosody Features

Prosody features are dynamic expressions of speech. Munt et al. and other studies from the perspective of rhythmic features found that the fundamental frequency of individuals with depression will decrease. Depression can be analyzed by prosody features, such as reduced loudness, pitch change, and loudness change (France et al., 2000), as well as the fundamental frequency and prosody change of pronunciation (Mundt et al., 2007).

3.2 Spectral Features

Spectral features usually reflect the short-term features of speech signals are related to changes in muscle tone and control related to vocalization. The researchers believe that spectral features can characterize the change in the speaker's mental state. Low et.al. pointed out that spectral features are useful for the identification of adolescent depression. Hönig et al. pointed out that MFCCs are a type of useful feature in depression recognition.

4. Approaches for Speech-Based Depression Recognition

4.1 Traditional Machine Learning Approaches

The approaches of traditional machine learning are to manually extract depression-related low-level descriptors (LLD). Low et al. also proposed a Teager-based Energy operator features, found that increasing the Teager energy operator features can improve the performance of depressed men by 31.35%. After the feature selection, the features are classified by machine learning algorithms. Commonly used classifiers are Gaussian mixture models (GMM) and support vector machines (SVM).

A variety of speech-based automatic prediction approaches for depression have been studied. Moore et al. and Ooi et al. studied the formation of a classification system from the combination of prosody, sound quality, frequency spectrum and glottal features. The above literature analysis shows that Teager energy and glottal features are more accurate than classifiers based on individual prosodic features. The results of these two experiments (Ooi et al., 2013) support the previous discussion that depression's effect on muscle tone and larynx control can lead to glottic flow. There are also several literatures that show the applicability of individual rhythm, sound quality, spectrum and glottal features. In literature (Stasak, Epps, Cummins, et al., 2016), GMM is used to classify a series of speech features and spectrum features. The Mel Cepstral Coefficients (MFCCs) have reached with 77% accuracy, the formant reaches 74% accuracy. Helfer et al. also showed the superior performance of GMM and SVM in the classification of depression severity.

4.2 Deep Learning Approaches

Recently, the recognition approaches of depression with deep learning have been proposed (Ringeva et al., 2015; Rejaibi, Komaty, et al., 2019). The deep learning approach is to build a multi-layer hidden neural network model and use a large number of speech samples to train the feature model to extract the most relevant features.

Several deep neural networks have been proposed. He et al. input low-level features manually extracted and high-level features extracted by DCNN together into the DCNN network, then through a joint fine-tuning layer to predict depression, get great results on a larger data set. Jain et.al. compared with BLSTM, CNN and LSTM-RNN with attention mechanism, the capsule network proved to be the most effective architecture. A novel approach is proposed (Chlasta et al., 2019), after the DAICWOZ data set is expanded, the residual CNNs with different depths reach 77% accuracy. In order to evaluate the depression level using the "Patient Health Questionnaire 8" (PHQ-8) scale for evaluating depression levels, Yang et al. proposed DCNN. To our knowledge, their approach is superior to all existing approaches on the DAICWOZ dataset.

5. Research Challenge

Looking back at the previous research in the field of speech-based depression recognition, although speech signal research has made a series of progress in the field of depression recognition, there are still many challenge worth considering.

5.1 Sample Scale

The small scale of depressive speech samples of depressed individuals is not conducive to the generalization and training the model. Each hospital can establish a unified and standardized patient health system database for data sharing, thereby improving the generalization ability of the model.

5.2 Longitudinal Control Group

In future research, a longitudinal control group should be added to increase the breadth of the population of speech prediction. Therefore, the comparison between disease types can increase the specificity of speech recognition, thereby providing effective biological indicators for clinical diagnosis.

5.3 Feature Selection

Distinguishing between depressed and non-depressed individuals is based on different features of speech. Using the appropriate feature selection approach to select the most relevant features of depression is critical to the diagnosis of depression.

6. Conclusion

In this study, relevant researches on speech-based depression recognition were summarized. Thereby, using speech recognition technology to understand the role of speech signals as biological indicators in the diagnosis, recurrence prevention and efficacy evaluation of depression, lays the foundation for the clinical application.

References

Pan, W., Wang, J., Liu, T., Liu, X., Liu, M., & Hu, B., et al. (2018). Depression recognition based on speech analysis. Chinese ence Bulletin.

Schneider, D., Regenbogen, C., Kellermann, T., Finkelmeyer, A., Kohn, N., & Derntl, B., et al. (2012). Empathic behavioral and physiological responses to dynamic stimuli in depression. Psychiatry Research, 200(2-3).

Moore, E. I., Clements, M. A., Peifer, J. W., & Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Transactions on Biomedical Engineering, 55(1), 96-107.

Low, L. S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. IEEE transactions on bio-medical engineering, 58(3), 574-86.

Ooi K E B, Lech M, Allen N B. (2013). Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents. IEEE Transactions on Biomedical Engineering, 60(2):497-506.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, & Gordon Parker. (2012). From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. FLAIRS Conference.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2019). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering, 47(7), 829-837.

L. Yang, H. Sahli, X. Xia, E. Pei, M.C. Oveneke and D. Jiang (2017). "Hybrid depression classification and estimation from audio video and text information," In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 45-51.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. Journal of Neurolingus, 20(1), 50-64.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., & Gratch, J., et al. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. Image & Vision Computing, 32(10), 648-658.