World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:06, 2022

Speech Emotion Recognition: A DNN and LSTM Comparison in Single and Multiple Feature Application

Authors : Thiago Spilborghs Bueno Meyer, Plinio Thomaz Aquino Junior

Abstract : Through speech, which privileges the functional and interactive nature of the text, it is possible to ascertain the spatiotemporal circumstances, the conditions of production and reception of the discourse, the explicit purposes such as informing, explaining, convincing, etc. These conditions allow bringing the interaction between humans closer to the human-robot interaction, making it natural and sensitive to information. However, it is not enough to understand what is said; it is necessary to recognize emotions for the desired interaction. The validity of the use of neural networks for feature selection and emotion recognition was verified. For this purpose, it is proposed the use of neural networks and comparison of models, such as recurrent neural networks and deep neural networks, in order to carry out the classification of emotions through speech signals to verify the quality of recognition. It is expected to enable the implementation of robots in a domestic environment, such as the HERA robot from the RoboFEI@Home team, which focuses on autonomous service robots for the domestic environment. Tests were performed using only the Mel-Frequency Cepstral Coefficients, as well as tests with several characteristics of Delta-MFCC, spectral contrast, and the Mel spectrogram. To carry out the training, validation and testing of the neural networks, the eNTERFACE'05 database was used, which has 42 speakers from 14 different nationalities speaking the English language. The data from the chosen database are videos that, for use in neural networks, were converted into audios. It was found as a result, a classification of 51,969% of correct answers when using the deep neural network, when the use of the recurrent neural network was verified, with the classification with accuracy equal to 44.09%. The results are more accurate when only the Mel-Frequency Cepstral Coefficients are used for the classification, using the classifier with the deep neural network, and in only one case, it is possible to observe a greater accuracy by the recurrent neural network, which occurs in the use of various features and setting 73 for batch size and 100 training epochs.

Keywords : emotion recognition, speech, deep learning, human-robot interaction, neural networks
Conference Title : ICSERDL 2022 : International Conference on Speech Emotion Recognition with Deep Learning
Conference Location : Barcelona, Spain
Conference Dates : June 09-10, 2022