Correlation between Speech Emotion Recognition Deep Learning Models and Noises

Authors : Leah Lee

Abstract : This paper examines the correlation between deep learning models and emotions with noises to see whether or not noises mask emotions. The deep learning models used are plain convolutional neural networks (CNN), auto-encoder, long short-term memory (LSTM), and Visual Geometry Group-16 (VGG-16). Emotion datasets used are Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), Toronto Emotional Speech Set (TESS), and Surrey Audio-Visual Expressed Emotion (SAVEE). To make it four times bigger, audio set files, stretch, and pitch augmentations are utilized. From the augmented datasets, five different features are extracted for inputs of the models. There are eight different emotions to be classified. Noise variations are white noise, dog barking, and cough sounds. The variation in the signal-to-noise ratio (SNR) is 0, 20, and 40. In summation, per a deep learning model, nine different sets with noise and SNR variations and just augmented audio files without any noises will be used in the experiment. To compare the results of the deep learning models, the accuracy and receiver operating characteristic (ROC) are checked.

Keywords : auto-encoder, convolutional neural networks, long short-term memory, speech emotion recognition, visual geometry group-16

Conference Title : ICSERDL 2024 : International Conference on Speech Emotion Recognition with Deep Learning

Conference Location : Copenhagen, Denmark

Conference Dates : July 18-19, 2024

Open Science Index, Computer and Information Engineering Vol:18, No:07, 2024 publications.waset.org/abstracts/170547.pdf