Using Deep Learning for Detection and Correction Speeches Based on Human Emotions

Mateusz Półtorak, Janusz Pochmara
Poznan University of Technology
mateusz.poltorak@student.put.poznan.pl, janusz.pochmara@put.poznan.pl

Abstract

The main problem in speech data recognition is acoustic variability. It's depends on [1]: First: there is variation in what is said by the speaker. Second: there is variation due to differences between speakers. Third: there is influence of noise conditions.Fourth: emotions that affect the form of expression and its quality. Many projects are based on steps from First to Third. We propose using Fourth step as main goal of investigation. Before the appearance of convolutional neural networks [2], most of the emotion speech recognition models were based on the extraction of features (for example: energy-based distributions of speech [3]. This process was followed by simple machine learning classifier like SVM or dense neural network. Our work doesn't include feature extraction. It is based only on image decoder. We propose a model that is independent of the First to Third problems. Speaker's emotions are encoded in the word he uses, but the tone and intonation of his voice are even more important. Shapes created by different words, but spoken with the same feelings are quite similar. Hence, our model is resistant to the occurrence of untrained words. We want to normalize emotions stored in speaker's voice to improve the quality of algorithms for detection and correction of signals in terms of correct pronunciation.The main idea is to convert graph of the spectrum of frequencies and amplitudes of speech as the main information carrier. Deep convolutional neural network is responsible for decoding information stored in the spectrogram [4]. In our research we will focus on the convertion accuracy. We will also focus on problems with reduce of the system complexity. In this paper, we propose an evaluation system for classification and correction of speech.

References

1. Publication of National Institute on Deafness Other Communication Disorders. Statistics on Voice, Speech, and Language. https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language.
2. Krishna Asawa and Priyanka Manchanda. Recognition of Emotions using Energy Based Bimodal Information Fusion and Correlation. International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, N° 7.
3. Yann LeCun and Yoshua Bengio. Convolutional Networks for Images, Speech and Time Series. The Handbook of Brain Theory and Neural Networks, MIT Press, 1995, .