Abstract

People with articulation and phonological disorders need exercise to execute sounds of speech. Essentially, exercise starts with production of non-articulatory sounds in clinics or homes where a huge variety of the environment sounds exist; i.e., in noisy locations. Speech recognition systems considers environment sounds as background noises, which can lead to unsatisfactory speech recognition. This study aims to assess a system that supports aggregation of visual features to audio features during recognition of non-articulatory sounds in noisy environments. The methods Mel-Frequency Cepstrum Coefficients and Laplace transform were used to extract audio features, Convolutional Neural Network to extract video features, and Support Vector Machine to recognize audio and Long Short-Term Memory networks for video recognition. Report experimental results regarding the accuracy, recall and precision of the system on a set of 585 sounds was achieved. Overall, the results indicate that video information can complement audio recognition and assist non-articulatory sound recognition.
References


Index Terms

Computer Science

Artificial Intelligence

Keywords

Assistive technology; health information; speech recognition; machine learning; down syndrome.