ABSTRACT

Audio signal processing is a field that has been of special interest to researchers for over 50 years. This is a field with many practical applications, ranging from large systems like autonomous vehicles and robotics to personal mobile phone applications such as learning and healthcare apps. Speech recognition, also widely known as Speech-to-Text, is the starting point for these applications. We propose a fine-tuned version of the Facebook Wav2Vec2.0 Large XLSR-53 model for Vietnamese speech recognition. Wav2Vec2.0 is a retraining model that uses the most powerful network architecture available today – Transformers. This research presents a new speech recognition model that achieves a WER score of 14.06 on the Common Voice 8.0 Vietnamese dataset. Our model produces results close to the best Vietnamese language recognition model currently available, khanhld/wav2vec2-base-vietnamese-160h. The proposed model could be used to develop a variety of speech-enabled applications, such as voice assistants, dictation software, and automatic subtitling. Additionally, we have built a web User-Interface for this model, it is like a speech recognition application that we can use to test with your own voice directly on the browser.