Abstract

Visual Speech Recognition (VSR) is a technology that recognizes and interprets spoken language by analyzing facial and lip movements in video. Its primary goal is to decode language content using visual cues, which is particularly valuable when audio information is limited or absent. VSR has made significant progress, with the current mainstream approach focusing on extracting lip features and using deep learning to enable the model to understand a speaker's content from video alone. However, one might question whether visual language recognition is synonymous with lipreading. Can we extract additional information beyond lip movements to improve model performance? In this study, by developing the Lip-Face-Surrounding model to comprehensively extract information from videos. This model supports three input channels, utilizes 3DCNN for feature extraction, and applies a CTC layer to align the extracted features with the text sequence. 3D Convolutional Neural Networks (3DCNN) excel at extracting spatial and temporal features, making them well-suited for visual speech recognition tasks. By employing 3DCNN, the model captures dynamic changes across facial, lip, and surrounding cues within video sequences. The Connectionist Temporal Classification (CTC) layer effectively addresses alignment, allowing extracted features to align with the target text sequence without requiring predefined alignment, thus enhancing the model's capability to handle variable-length input. The findings show that direct information beyond the lips—such as eye corner movements, jaw movements, nostril movements, throat movements, and shoulder movements—are captured by the model and serve as discriminative features for visual speech recognition. This direct information is also applicable to handcrafted datasets. Additionally, indirect information such as the speaker's body language and interactions with the surrounding can impact model

performance. Sometimes, this information provides extra context that enhances performance, while other times, it introduces noise that affects model convergence. This model achieved promising results on the CN-CELEB and GRID datasets, with a 5% absolute performance improvement over the lip-only approach.