Chapter 9
Advancements in Deep Learning for Automated Dubbing in Indian Languages

Sasithradevi A.
Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, India

Shoba S.
Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, India

Manikandan E.
Centre for Innovation and Product Development, Vellore Institute of Technology, Chennai, India

Chanthini Baskar
Vellore Institute of Technology, Chennai, India

ABSTRACT

After the proliferation of deep learning technologies in computer vision applications, natural language processing has used deep learning methods for its building steps like segmentation, classification, prediction, understanding, and recognition. Among different natural language processing domains, dubbing is one of the challenging tasks. Deep learning-based methodologies for dubbing will translate unknown language audio into meaningful words. This chapter provides a detailed study on the recent deep learning models in literature for dubbing. Deep learning models for dubbing can be categorized based on the feature representation as audio, visual, and multimodal features. More models are prevailing for English language, and a few techniques are available for Indian languages. In this chapter, the authors provide an end-to-end solution to predict the lip movements and translate them into natural language. This study also covers the recent enhancements in deep learning for natural language processing. Also, the future directions for the automated dubbing process domain are discussed.

DOI: 10.4018/978-1-6684-6001-6.ch009

INTRODUCTION

Dubbing is an intelligent phenomenon or process which is majorly used in film industry. Major companies are getting involved in this dubbing project where there is a need for multilingual. This process involves an audio-video translation which helps in adding new sound effects or audio during the production stage. Also, in translating the other language films (i.e. The dialogues) to the required language dubbing is being done. Not only for the translation, also in the original film adding sound tracks in synchronization with the situation dubbing helps. For example, in other language films translation, the subtitles are given in addition to the dubbing so that the viewers can get the full recipe of the motion picture. But there are some concerns to be addressed in the dubbing process. (Akman et al., 2022) The basic problem in the dubbing process is with the equipment quality used by the dubbing artist. It is suggested to use an efficient microphone which is a one-time investment helps in making profitable in the later stage. Another point is, the environment where the process is getting done. Use of proper noise cancellation system is required for avoiding this kind error. These basic problems can be resolved by using proper environmental setup.

The major issues come in the technical part where proper synchronization is required in audio-video i.e. the timing issue. In another way, the dubbing artist audio timing should get matched with the actor visual where the actor opens mouth for dialogue delivery etc. So the lip synchronization and also the facial expressions matching should be taken care in the dubbing. Next, the lip synchronization system and the importance of deep learning for the automated system are discussed Kunchukuttan, A. et.al. (2017), In short, the following challenges are present in the system.

- Context absence
- Spatial – temporal features extraction
- Difficulty in understanding more expression actors and its synchronization to be done by the artist
- Generalization among people
- Speaker accent such Guttural sounds
- Speakers sound level example: mumbling

So, the automated lip reading system has to be developed which should be able to address these concerns. The generalized flow of the automated lip synchronization model is depicted in Figure 1. The major components of this system are feature extraction and classification. Initially, the videos of speakers will be captured with the proper vision setup. Then the videos will be converted into image frames which represent the data to be decoded. The next step in this is very important where pre-processing will be done. Our objective here is to capture the proper lip movements which helps in identifying the right speech. At this point, the lip locations are our Region of Interest (ROI). So the lip locations are extracted from video images. Some basic transformation and manipulations will be applied which reduces the number of steps involved in the later stage. After this, the feature extraction which is the front-end of the system is being done. This process helps in obtaining the required features effectively from the redundant features. Then, the backend of the system comes which is nothing but a classification process. In the classification, the obtained images are being compared with the database where facial movements corresponding to the input speech is done. Finally, the decoded speech is encoded into in a form of spoken words or sentences.