I. Introduction

The recognition of lip motion, which is also one of the visual speech recognition, is a technique to interpret visual data. The task focuses on the mouth area and aims to recognize lip motion so that it can be classified in class [1]. The recognition of Indonesian vowel phonemes on lip motion is needed to build the foundation of research on word recognition and even sentences in Indonesian Language on lip motion. Another benefit provided by the introduction of lip motion is making transcripts from a video without audio [2] and helping the deaf person in order to communicate with others. Besides communication, another benefit is the creation of therapeutic tool technology for the learning process to speak for a deaf person [3].

Research related to lip reading is mostly done for English, but very few for the Indonesian language. Lip reading research for the Indonesian language started in 2012 using Artificial Neural Networks (ANN). The study on the recognition of Indonesian vowel phoneme on lip motion was conducted by implementing lip localization followed by lip segmentation [3]. Moreover, the classification process carried out by Artificial Neural Networks (ANN) and achieved an accuracy of 75.9%. However, the performance of the system needs to be developed further so that the accuracy value may increase.

Several studies of computer vision use Convolutional Neural Networks (CNN) as the standard classifier. Since Krizhevsky et al. [4] with the AlexNet won the ILSVRC ImageNet Large Scale Visual Recognition Competition in 2012, CNN has become very popular for research in the field of computer vision such as object detection and classification of visual data into several classes. One of the advantages of CNN is that the entire system is trained in end-to-end. Because it is trained end-to-end, CNN also performs feature extraction so that the process of feature extraction on visual data is no longer needed before the model trains the data. For lip motion recognition problems, CNN has a significant contribution to this problem. A new CNN architecture, LipNet, was created and introduced for the recognition of lip motion for English [5].

The topic to be discussed in this paper focuses on the recognition of Indonesian vowel phonemes (/a/, /i/, /u/, /e/, and /o/) on lip motion. We introduce a model for the topic using 3D CNN. The primary goal of this paper is to perform data classification by the proposed model that can recognize five vowel phonemes. In the Indonesian language, it is difficult to form a word or sentence without involving the role of vowel phonemes. The existence of vowel phonemes is needed to produce sounds so that words and sentences in the Indonesian language can be formed and has a meaning. The use of 3D CNN is expected to capture the spatial features from the data accurately since the previous study that use ANN as the classifier cannot capture the spatial features of an image data. Spatial features refer to the arrangement of pixels and the relationship between them in image data. Despite the primary goal of this research, we hope this research can be used for the basis for visual speech recognition research in the Indonesian language in the form of words or sentences.

Furthermore, studies related to the methods used are in section II. The system design is explained in section III. In section IV, the result of the proposed method is presented and discussed. The conclusion and future works are finally highlighted in section V.

II. Related Works

A. Lip Reading

Lip motion recognition is one of the visual speech recognition methods that only require visual data. This method gives the system an ability to understand what is being said without audio data input. Research on the recognition of lip motion has been carried out by researchers in various parts of the world and in several languages [6], [7]. The research is still being developed by adding more complex data, hyperparameter tuning models, or combining novel methods with the new method so that lip recognition will be better [8], [9].

The recognition of lip motion for the Indonesian language started in 2012 using artificial neural networks or ANN. Faridah et al. [3] achieved 75.9% accuracy for recognizing Indonesian vowel phonemes on lip motion using ANN. Then in 2015, the Hidden Markov Model (HMM) method used for recognizing Indonesian syllable phonemes [10]. The study achieved an accuracy of 63% in the palatal phoneme and 78% in the bilabial-palatal phoneme.

Research on the recognition of lip motion for some common everyday words in Indonesian was conducted by Nasuha et al. [11] in 2017. They used the horizontal-vertical image projection method and the combination of different frame methods to recognize common Indonesian words. The classification process uses Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). In 2018, Nasuha et al. [12] conducted a study related to the recognition of lip motion for the five visual phonemes namely /ma/, /me/, /mi/, /pu/, and /sa/ using CNN and managed to produce an average accuracy of 93.93% and can achieve the highest accuracy of 96.44%.

B. Convolutional Neural Networks

When facing video analytics problems, it is desirable to capture the motion information encoded in multiple contiguous frames. In 3D CNN, the 3D convolutions to compute features from both spatial and temporal dimensions where 2D convolutions are applied only on the 2D feature maps to compute features from the spatial dimensions only [13]. The feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer, thereby capturing motion information. The computational process of 3D CNN is illustrated in Fig. 1.

In 2013, Ji et al. [14] using 3D CNN for human action recognition instead of 2D CNN. The developed model generates multiple channels of information from the input frames, and the final feature representation combines information from all channels. Then Molchanov et al. [15] and Camgoz et al.