CHAPTER 1

INTRODUCTION

This chapter discusses the underlying background of the research , and the overview of previous works in Lip Reading and Deep Learning Model.

1.1 Rationale

Lip reading is a communication method to understand the speech by reading the speakers lips and tongue movement without hearing the sound. This method can help people who have hearing loss or to understand people who lost their vocal cords due to illness. Lip reading is also categorized in the speech recognition field. Precisely, it is known as visual speech recognition.

In fact, lip reading is not a new invention. It is noted that lip-reading has been used since the 1500s and probably before that time. The first known lip reading teacher was the Spanish Benedictine monk, named Pietro Ponce. After that time, the method began to spread to other countries. The first school which studied lip reading was opened in Leipzig around 1787. The founder was Samuel Heinicke, a German psychologist. Although it has been invented and practiced for a long time, lip reading is hard to learn by a human, because also it very much depends on the ability of each person. The overall accuracy for human lip-reading ability is around 53%[11]. The reader must guesstimate other words that have missed.

Currently, technological developments make a chance for a lip-reading method to improve. Lip reading has been researched in various languages, such as English, Korean, Japanese, Dutch, Arabic. However, it has not been widely studied in the Indonesian language[21].

The challenges often encountered in lip-reading or also recognized as visual speech recognition, are variances in inputs such as differences in skin color, face shape, and other facial features[2]. In order to simplify the problem, many systems are restricted to limited numbers of words or phrases and also speakers. Although many methods have been developed to detect lip contour accurately, it still has many problems to solve.

As described by Lin et.al in their research, facial features such as mustache can cause differences in detection of the Region of Interest (ROI) and lip contour positions. It reduces the system performance. Then they designed the criteria in determining ROI and lip contour, i.e., width, height, contour points, area, and ratio to recognize lip and English

vowels when speaking. Nonetheless, it has not reached the expected target. Speakers with thick-lip makeup effects got 79.92% accuracy.Meanwhile, speakers that have a mustache, decreased the result to 50.11%[18]. Nowadays, the availability of large data allows deep learning to develop and become the early success in lip reading. However, re-search on Indonesian lip-reading seemed to be not advancing, its latest research was still very limited on words and phonemes classification [1].

The other problem is about out of vocabulary (OOV). OOV is a word that cannot be recognized because there is no such word in the dictionary. Most speech recognition systems can only recognize words that belong to a fixed finite vocabulary. When encountering an OOV word, the recognizer will incorrectly recognize the OOV word with one or more similar vocabulary words. In addition, OOV words also affect the recognition performance of their surrounding words. That is the main reason why this model used syllable based not word-based. A syllable or sub-word based model gives the chance to construct a new word that does not exist in the dictionary. The combination of a syllable that already recorded can produce a new word. This research aims to create an Indonesian lip-reading model that can control OOV, which would address it relevant in the real world settings using deep learning.

1.2 Theoretical Framework

This research, Indonesian Lip Reading Model, proposes a model to translate a sentence spoken by a speaker with no sound video(silent video) in the Indonesian Language into text form. The sentence is formed from smaller parts called syllables. Started from a sentence, the system divided it into parts of syllables. Then, the system classified every syllable into a class of syllables that have learned previously by the system. If all of the syllables have been classified, the sentence will be re-formed.

With the use of a syllable-based system, new words can be formed by joining the syllables that have been learned by the system. So, it is not necessary for the system to learn all the words. Besides, learning all words is impossible to train. Considering that every year the vocabulary in Indonesian continues to grow for Indonesian vocabulary are also developed from foreign language absorption words and terms.

1.3 Conceptual Framework/Paradigm

The fundamental concept of the proposed approach is to read speaker lip's movement to listen the sentence that spoken. This research focuses on a system that learns syllables from some sentences and used it to predict words that are not included in the training process. Before a system can classify a syllable, a training process is needed to gain system

knowledge. The data input of the system is a video (sequence of images) following by its label. The video is containing a speaker to speak one sentence, where the speaker position must face the camera, aiming at lips so that the movement can be seen clearly and the system can learn it well. A label consists of the splitting point in the video that separates each syllable.

In addition, the important part of the video is the surrounding mouth, so there are many features which are not used in the lip-reading process. The use of unimportant features is a waste of memory and time, so cropping process around the mouth is needed. The cropped sequence of images and its label is used for the training process. The system will learn how each syllable is pronounced and save this knowledge. The knowledge base will be used in the testing process to translate a new video into text form, although there are words that have never been learned before by the system.

1.4 Statement of the Problem

Lip reading or visual speech recognition has been researched in various languages, such as English, Korean, Japanese, Dutch, Arabic. However, it has not comprehensively studied in the Indonesian language. There are 2(two) approaches in visual speech recognition: a visemic method and a holistic method. Visemic is the most commonly used for an automatic lip-reading method, which is based on visemes. Viseme stands for visual phoneme, which is facial images that describe a particular sound. In 'Indonesian Text to Audio Visual Speech with Animated Talking Head', Muljono et.al grouped Indonesian Visemes into 14 groups[20].

However, viseme is difficult to use for describing the word as a whole. This is because the shape of the viseme will change following the vowel sounds that exist before or after it. Different from the visemic approach, a holistic approach considers the lip movement for the whole world, instead just parts of some word. This provides a good alternative for visemic approach. However, it has a problem with the whole words that must be trained[11][22][16].

The other problem in visual speech recognition is a feature extraction method which used. Features extraction is an essential part of lip reading. Each feature will be a characteristic that describes each class. In order to distinguish one syllable from another, the characteristics obtained must be precise.

1.5 Objective and Hypotheses

This research will develop an Indonesian lip reading model with syllable based.

1. Deep learning implementation in feature extraction is excellent at deriving the correct features from images or videos. By using deep learning, the system will be able to extract various facial features. As a result, even though there are facial features such as a mustache or thin lips, accuracy will not decrease dramatically
2. Syllable-based model will be able to read auto vocabulary features. If some words do not exist in the dictionary, the model can still predict the words with a combination of syllable features from other words.

1.6 Assumption

The important part of lip reading or visual speech recognition is the surrounding mouth. As a consequent, every video must show the position of the mouth clearly; in other words, the speaker must face the camera when recording the video. The lighting when video record-ing must sufficient enough so that the lips movement can be clearly recorded. The system also automatically searched the mouth of the speaker, then cropped and saved it into a sequence of images with size 100x50px (width-height). As a result, all of the data change from video with extension MP4 into a sequence of images form in PNG with size 50x100x3 (3 from the RGB channel).

1.7 Scope and Delimitation

Every people has different intonation and also different speed of speaking. The result is the time needed to speak one syllable will be different for each person. In contrast, the system requires the same input of length (fixed size) for all syllable. To deal with this, normalization for each syllable was done. The average time needed for someone to speak one syllable would be sought. Then, all of the syllables would be normalized into the length following the average time.

1.8 Significance of the Study

This research shows the ability of deep learning in feature extraction. Although there are a variety of facial features on the speaker, the system accuracy is more stable. This study also shows that the use of syllables, raises the probability of new word recognition by combining syllables that has been studied by the previous system. Surely this will provide an opportunity to overcome the problem of Out of vocabulary (OOV)