CHAPTER 1 THE PROBLEM

This chapter discusses the background of this study; it includes the following subtopics: rationale, theoretical framework, conceptual framework/paradigm, statement of the problem, hypothesis, assumption, scope and delimitation, and importance of the study.

1.1 Rationale

Speechreading is a technique to understand speech by visually interpreting the movements of the lips, face and tongue when normal sound is not available. This technique is usually used by people who suffered from severe hearing loss to understand their interlocutor. However, in noise situation, normal people also often depend on this ability.

There is a condition in speechreading called coarticulation. This condition shows that speech sound is influenced by the preceding or following speech sound. In other word, lip form is not just depending on each letter of a word, but it is continuously influenced by the phonetic segments which build the word [1]. Coarticulation usually occurs in the pronunciation of Indonesian language which word is based on syllables. Since coarticulation influence the lip's form, it is very difficult to differentiate each form. It makes speechreading difficult, since many speech-sounds are partly recognized.

There are 11 patterns of Indonesian syllable, vowel-consonant, consonant-vowel, consonant-vowel-consonant, consonant-consonant-vowel, consonant-consonant-vowel-consonant, vowel-consonant-consonant, consonant-vowel-consonant-consonant, consonant-consonant-vowel-consonant-consonant, consonant-consonant-consonant-vowel, and consonant-consonant-consonant-vowel-consonant. This study only covers the first pattern because it is the simplest form of Indonesian syllable.

Many studies have been conducted to classify the lip form called viseme. However, these studies are limited to particular language and do not consider the conflicting muscle condition which occur when a person is saying words in a specific expression. This study attempts to classify viseme in Indonesian language while considers the involvement of emotion when the person is speaking.

1.2 Theoretical Framework

This research is about a computer facial animation technique to capture and determine the facial parameter from talking model. Facial animation can be used as an effective

communication channel for human-computer interface. Talking model is built and will animate using the deformation rule in each keyframe of animation to generate a facial animation.

One of the focuses of facial animation is speech animation. Speech animation focuses on mouth movement including jaw, lip, teeth and tongue during speech. Mouth movement during speech is continuous and relatively rapid, and the movement encompasses a number of visually distinct positions, which is called visual phoneme (viseme). A viseme represents the shape of the lips when articulating an auditory syllable. Viseme have similarity form to each other. This means in an utterance, not all visemes are necessary to be generated. In order to determine the generated viseme, classification of viseme is needed.

Classification is used to classify viseme facial parameter. Then, an animation talking head model is generated based on the result of viseme classification. In order to get the feature of viseme, a character's face needs to be accurately captured in geometry and texture. There are two techniques to capture the character's face, using parametric models and physic-based approach. Both parametric models and physic-based approaches require some parameters to be specified and modified over time. Based on these parameters, viseme are classified.

1.3 Conceptual Framework/Paradigm

The objective of this study is to classify viseme when a person is talking and showing emotion. Viseme used in this research is based-on syllable form. Then, the classified viseme is applied to head model. This process is done in order to generate realistic visual speech synchronization. To apply this, the adopted conceptual framework of this study is illustrated in Figure 1.1

First of all, direct parameterization is applied to get parameter of facial muscle from real human video. Parameters used in this study cover lower face area, because viseme is occurring in that area. This study limited the emotion into 3 expressions: neutral, sadness and happiness. This is related to the research about the emotion recognition in lip feature [2]. Based on the research, 3 expressions are recognized in the lip: happiness, sadness and surprise. Because the surprise are rarely comes from the utterance, this study limited the emotion into happiness and sadness in addition to neutral expression.

The next step is classifying viseme based on the parameters of face. Viseme is classified based on the pixel distance from each viseme to another. Distance of each viseme is calculated using Euclidean distance. This step is the contribution of this study.

Next step after classifying viseme is modeling the talking head and applying the face deformation to the talking head. These processes produced the video animation. Video animation is generated using interpolation between shape key's frames.

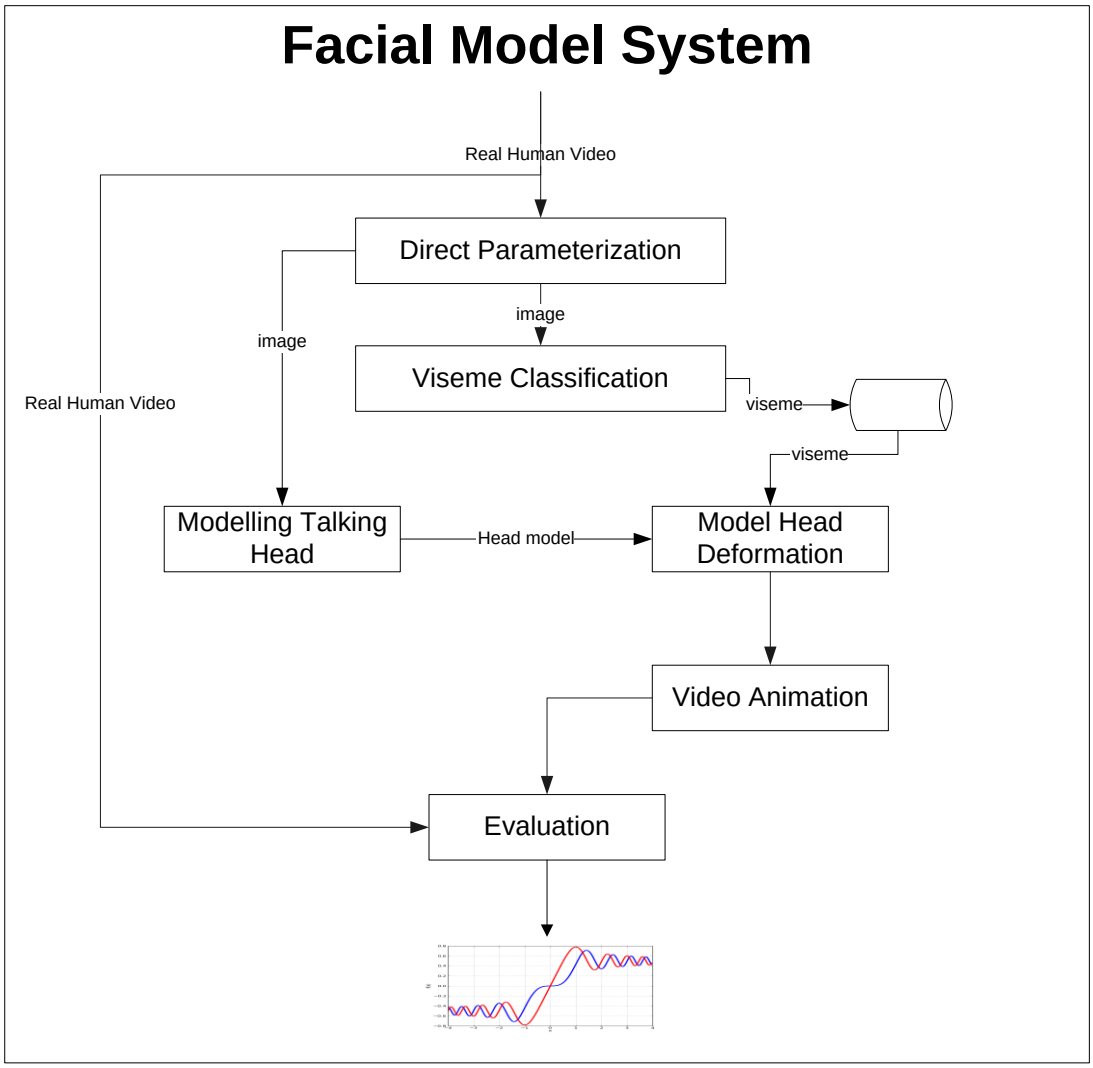

Figure 1-1: Conceptual Framework of facial model system

Evaluation is conducted by integrating the video animation and real human video. The process of evaluation is almost the same with the process to get the parameters of facial muscle. Frames in real model and the generated model which talk the same syllable are compared and pixel distances are calculated. The realistic of generated model are determined by that pixel distance.

1.4 Statement of the Problems

In facial animation, some studies to classify viseme were just applied to a limited particular language and they cannot be applied to Indonesia Language (Indonesian). Besides, viseme is classified without considering the muscle conflict condition. It means that there are inactive

muscles while other muscles are contracted when people talking in a specific expression. Natural movement of muscle deformation in video animation is needed to produce the facial expression talking head.

1.5 Hypothesis

Classifying phoneme-to-viseme mapping can simplify the generation process of video animation. Various combinations of expressive phonemes can be classified into several classes to make them simpler. Viseme classified in Indonesian language should be based on the coarticulation, because it gives more realistic movement.

1.6 Assumption

1. Real model show the true expression of emotion
2. Each expression in this study has maximum intensity for each model (human model and animation model)
3. Direct parameterization can be applied into muscle contraction area arbitrarily

1.7 Scope and Limitation

1. Phoneme to viseme mapping and co-articulation is only limited on consonant-vowel (CV) patterns
2. Only three expressions involved: sad, happy and neutral

1.8 Significant of the Study

1. Classifying viseme can reduce complexity when generating facial realistic model.
2. Facial realistic speech synchronization can improve the accuracy of speech reading application.
3. The proposed method can be used for the development of Talking-Head normally used in the field of telecommunications to help the communication of deaf people.