

Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation

Endang Setyati, Surya Sumpeno, Mauridhi Hery Purnomo, *Member, IAENG*,
Koji Mikami, Masanori Kakimoto, Kunio Kondo

Abstract— In a communication using texts input, a phoneme must be mapped to viseme (visual phoneme). Phoneme is the smallest element of a language that can differentiate a meaning. Viseme is derived from a group of phonemes having similar visual appearances, the equivalent unit in the visual domain that models a speech recognition system audio-visually. This paper proposes a classification of visemes in Indonesian Language and establishes phoneme-viseme mapping for Indonesian Language. In Indonesian words, there are lots of absorbed words from other languages. They come from vernacular or foreign languages, such as Arabic and English. From 49 phonemes, 12 Indonesian visemes have been produced, including silent. Viseme classes are defined through linguistic knowledge and grouped by the same visual appearance, and then validated through a survey. The approach used in Indonesian phoneme-to-viseme mapping is based on linguistic data. This is the first research in Indonesian phoneme-to-viseme mapping, which is expected to grow into a reference for further development in human language technology.

Index Terms— phoneme-viseme mapping, Indonesian Language, linguistic approach

I. INTRODUCTION

Audio-visual synthesis becomes more interesting to research, creative industries and e-learning environment, because the fusion of the auditory and the visual representation is the new research theme in human-computer interaction [1]. Many researchers have been exploring, experimenting and implementing the relationship between speech and facial expressions which correspond to articulation. [2] has shown that the perception of speech depends not only on acoustic cues, but also on visual cues such as lips movements and shapes of mouth.

Lips movements when speaking give a visual direction about the things spoken. [3] shows that the use of a video, which contains lips movements as well as sound, can increase

phonemes recognition more significantly than the use of sound only.

Every sound of a language, if proven to be able to differentiate a meaning, can be considered as a phoneme. Phoneme is the smallest unit of sound which becomes a basis for building a human speech. Phoneme-to-viseme mapping is essential to the visual recognition of speech and the synthesis of talking heads. In this paper, we propose a phoneme-to-viseme mapping for Indonesian language using linguistic approach.

Many literatures state that visemes have multiple interpretations. There are two plausible definitions [4] which state that (1) Visemes can be considered as terms of articulatory gestures, such as lips closing together, teeth and tongue exposure; and (2) Visemes are derived from groups of phonemes having similar visual appearance. The second definition [4, 5, 6, 7, 8, 9, 10] is the most widely used, despite a lack of evidence that it is better than the first definition [6].

Phonemes and visemes have high correlation [4], and visemes can be derived using mapping of phonemes to viseme. The mapping has to be a many-to-one map, because many phonemes can not be distinguished using only visual cues. In this paper, we propose the linguistic approach then validate the mapping result through a survey. Classes of visemes are defined through linguistic knowledge and the intuition of which phonemes might appear the same visually.

Up to now there has been no research that discusses Indonesian phoneme-to-viseme mapping, whereas mappings in other languages such as Arabic [11], German [1], English [12], have been investigated by many. Therefore, as an initial step of the development of the system, we try to create an Indonesian phoneme-to-viseme mapping that will help to create the appropriate lips movements of a spoken speech. This paper builds mouth animation using blend shapes based on video recording.

Blend shape models help analyze face images and video. Traditionally the regions are defined manually. The original idea [13, 14] is rapidly extended to a segmented face where the regions are blended individually. [15] has designed an algorithm that fits a blend shape model onto a single image and the result is an estimation of the geometry and texture of the person's face.

[16, 17] describe a keyframe animation system with a painting interface to assign blending weights. The system gives the animator freedom to assign the blending weights at the granularity of a vertex [18], render the detail texture maps of a single blend shape using geometry parameter, and

Manuscript received January 19th, 2015; revised May 6th, 2015. This work was supported by *Ditendik - DIKTI Kemendikbud RI* (Indonesian Higher Education General Director, Ministry of Education and Culture, RI) for its financial support under the scholarship program to improve the quality of international publications (sandwich doctoral scholarship program abroad) to Tokyo University of Technology, Japan.

Endang Setyati is with the Informatics Department, Sekolah Tinggi Teknik Surabaya, Surabaya, Indonesia (e-mail: endang@stts.edu).

Surya Sumpeno and Mauridhi Hery Purnomo are with the Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, 60111 (e-mail: {surya, hery}@ee.its.ac.id})

Koji Mikami, Masanori Kakimoto, Kunio Kondo are with the School of Media Science, Tokyo University of Technology, Hachioji, Tokyo, Japan (e-mail: {kondo, kakimotoms, mikami@stf.teu.ac.jp}).

add it to the previous rendering. The result is an animation which preserves the original blend shape textures better and maintains the high frequency content constant throughout the animation.

Section 2 of this paper is a brief introduction to the Indonesian language followed by the list of its phonemes. Section 3 is devoted to the adopted research methodology and the use of the model in keyframing. In section 4, we interpret the results and establish the phoneme-viseme mapping of the Indonesian language. We conclude it with a discussion of our results and ideas for future research.

II. INDONESIAN LANGUAGE

A. Characteristic of Indonesian Language

The Indonesian language (*Bahasa Indonesia*), is a unity language formed from hundreds of languages spoken in the Indonesian archipelago [19, 20]. Indonesian is the official language used by almost more than 250 million people in 34 provinces of the Republic of Indonesia.

Indonesia is the fourth most populous nation in the world due to its large population, making Indonesian language one of the most widely spoken languages in the world [19, 20]. Most Indonesians, aside from speaking the national language, are often fluent in vernacular as their mother languages, including Javanese and Balinese, which are commonly used at home and within the local community. Most formal education, and nearly all national media and other forms of communication, are conducted in Indonesian language.

Standard and formal Indonesian language is used in books, newspapers, television, and radio news broadcasts. The standard Indonesian language is continuously developed. Although the earliest records in Malay inscriptions are syllable-based written in Arabic script, modern Indonesian is phonetic-based written in Roman script [21]. It uses 26 letters as in the English/Dutch alphabet, as seen in Table I. Letters “q”, “v”, “x”, and “z” are used in loanwords from Europe and India, almost never used at the end of a Indonesian words.

TABLE I
INDONESIAN ALPHABET

No	Letters	Spelling	No	Letters	Spelling
1	A a	ah	14	N n	en
2	B b	bé	15	O o	oh
3	C c	ché	16	P p	pé
4	D d	dé	17	Q q	ki
5	E e	é	18	R r	air
6	F f	ef	19	S s	es
7	G g	gé	20	T t	té
8	H h	ha	21	U u	oo
9	I i	ee	22	V v	fé
10	J j	jé	23	W w	wé
11	K k	ka	24	X x	iks
12	L l	el	25	Y y	yé
13	M m	em	26	Z z	set

Indonesian Language was originally written using the system known as *van Ophuijsen* spelling. In 1947, the spelling was changed into *Soewandi Spelling* (so named at

the time of Minister of Education, Soewandi). This spelling changed the formerly spelled “oe” into “u”. All other changes were part of the Perfected Spelling System (*Ejaan Yang Disempurnakan*), an officially-mandated spelling reform in 1972, “tj” into “c”, “dj” into “j”, “j” into “y”, “nj” into “ny”, “sj” into “sy”, and “ch” into “kh”. Some of the old spellings (which were derived from Dutch orthography) do survive in proper names; for example, the name of a former president of Indonesia is still written *Soekarno*.

B. Description of Phoneme in Indonesian Language

1) Basic Definition of Phoneme

Before determining the exact number of Indonesian phonemes, the meaning of phoneme shall be first formularized. Phoneme, the smallest units of language sounds, is functional, it means that the unit has a function to distinguish the meaning. Every language sound has an equal chance to become a phoneme, but not all language sounds must become a phoneme. Usually the number of phonemes in a language is fewer than that of language sounds.

To recognize and determine functional language sounds (phonemes), a minimum pair contrast is used. A minimum pair is the smallest pair of a language structure which has a meaning in a language, or a pair of single words which are ideally similar, except for a different sound. For example, “dari” (from) and “tari” (dance), whose phonemes are /d/ dan /t/. Note that the slash marks are conventionally used to indicate a phoneme.

2) Pronunciation of Phoneme

In Indonesian language, the sounds of letters /f/, /v/ and /p/ in a word are not essentially three different phonemes, for example, *provinsi* (province). “Provinsi” (province), when pronounced as “provinsi” or “profinsi” or “propinsi”, means the same thing.

Phonemes can not stand alone because they have no connotation. An example is phonemes /l/ and /r/. If these phonemes are stand-alone, surely we will not be able to get any meaning. It will be different if the two phonemes are combined with other phonemes such as /m/, /a/, and /h/, now they can form the meanings of “malah” (even) and “marah” (angry). For the Japanese, the words “malah” and “marah” might be the same, because the Japanese does not have /l/ phoneme. Another example: “mari” (let) and “lari” (run), if one element of the first letter is replaced by another element, it will cause a big effect that changes the meaning.

C. Comparison between Phonemes in Indonesian and Other Languages

The Indonesian phoneme set is defined based on Indonesian grammar described in [22]. Six primary vowels in Indonesian language [23] are represented as six phonemes: /a/, /e/, /ə/, /i/, /o/, and /u/. They are similar to vowels in English, i.e., /a/ (like “a” in “father”), /e/ (like “e” in “bed”), /ə/ (a schwa sound, like “e” in “learn”), /i/ (like “ee” in “meet”), /o/ (like “o” in “stop”), /u/ (like “oo” in “soon”), and three diphthongs, /ai/, /au/, and /oi/. There are vowels in Indonesian pronunciation similar to vowels in other language pronunciation. For example, vowel /a/ in German (like “a” in “mann”), vowels in Japanese, i.e., /a/ (like “a” in “aka”), /e/ (like “e” in “eki”), /i/ (like “i” in “ika”), /o/ (like “o” in

“oto”), /u/ (like “u” in “ushi”), vowels in Arabic, i.e., /a/ (like “a” in “bacaba”), /i/ (like “i” in “bicibi”), /u/ (like “u” in “bucubu”). There are no /a:/, /i:/, and /u:/ in Indonesian language. Interestingly, Japanese language in nature does not have the sound “r”, “l”, “th”, “v”. It cannot differentiate between “r” and “l”. Japanese do not move their lips very much when talking. A lot is done by the tongue movement. The English word “weekend” becomes a six-syllable word when pronounced in Japanese – “u-ii-ku-e-n-do” (oo-ee-koo-en-doh). When pronounced in Indonesian language it becomes a two-syllable word – “wik-en” (week-end) [24].

Based on [4], there is a difference in the number of phonemes used in phoneme-to-viseme mapping, even in the same language. An example is the number of phonemes in English that have been studied [4]. The first map [7] groups 43 phonemes in English for what is described as “usual viewing conditions”. The second map [8] is composed by 42 phonemes. This map can be considered as a mixture of linguistic and data driven approach. The third map [9] groups 52 phonemes using a data driven approach. It performs bottom-up clustering using models created from phonetically labelled visual frames. [10] uses 45 phonemes and creates a map using the linguistic approach.

D. Indonesian Phoneme Set

An International Phonetic Association organizes the letters of the International Phonetic Alphabet (IPA) into two categories: vowels and consonants [25].

A vowel is a sound produced by the unrestricted flow of air in the vocal chords. Vowels are important because nearly every word has at least one. There are five letters that represent Indonesian vowels, but there are ten distinct sounds associated with them.

However, in Indonesian language, we just use 10 of 28 vowels [25], including the allophones. Sounds produced by human speaking devices are allophones, which are merely variants of the same phoneme. Therefore, we can summarize that the IPA vowels chart [25] becomes the IPA Indonesian vowels chart [26].

INDONESIAN VOWELS

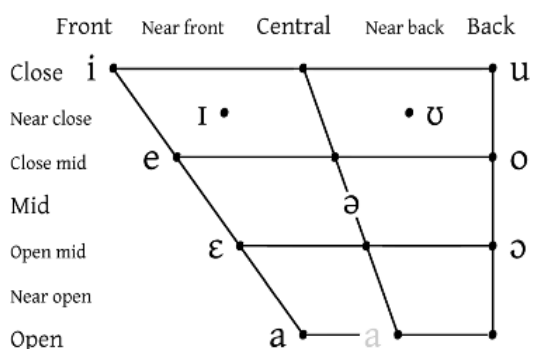


Fig. 1. IPA Indonesian Vowels Chart [26]

The IPA Indonesian vowels chart is shown in Fig. 1. Vowels are grouped in pairs of rounded and unrounded vowel sounds. These pairs also arranged from front on the left to back on the right, and from maximal closure at top to

minimal closure at bottom. Six ‘primary’ vowels in Indonesian language are represented by six phonemes: /a/, /e/, /ə/, /i/, /o/, and /u/. The other four are the allophones: “e” (/e/), “i” (/i/), “o” (/o/), and “u” (/u/).

A consonant is a sound produced by the restricted flow of air in the vocal chords. Consonants are arranged singly or in pairs of voiceless and voiced sounds, and then grouped in columns from front (bilabial) sounds on the left to back (glottal) sounds on the right. In official publications by the IPA, consonants arranged in rows from plosives, affricates, fricatives, nasal, trill, lateral to semi vowel.

TABLE II
ARTICULATORY PATTERN OF INDOONESIAN CONSONANTS [23]

Manners of Articulation	Places of Articulation					
	Bilabial	Labiodental	Dental/Alveolar	Palatal	Velar	Uvular
Plosives	p, b		t, d, dh, dl, dz, th		k, g	q, gh
Affricates				c, j		
Fricatives		f, v, ph	s, z, ts, ps	sy, sh	ks, kh	x
Nasals	m		n	ny	ng, ngg	
Trill			r			
Lateral			l			
Semi-vowels	w			y		

The articulatory pattern of Indonesian consonants used in this study is shown in Table II. The consonant table is arranged in rows that designate manners of articulation, meaning how the consonant is produced, and columns that designate places of articulation, meaning where in the vocal tract the consonant is produced.

Indonesian language also contains absorbed words from local dialects such as Javanese, Balinese, and foreign languages such as Arabic and English. Some examples of the loanwords are phonemes /dh/ in “dhuafa”, /dl/ in “ramadlan”, /dz/ in “muadzlin”, /gh/ in “maghrib”, /sh/ in “sholat” from Arabic. Loanwords [27] is the original form of words in one language used in another language, for example, English lexical items “therapy”, “tsunami”, “basement”, “final”, “stop”, “urgent”, and “stopwatch” which are frequently used in Indonesian Language.

The number of Indonesian phonemes used in this research is 49, including vowels (V) and consonants (C), monophthong (single letter) and diphtong (double letters). The vowel phonemes used in this research consist of 10 single vowels, consisting of /a/, /ə/, /e/, /ε/, /i/, /I/, /o/, /ɔ/, /u/, /U/ and 3 double vowels, consisting of /ai/, /au/, /oi/. In English, these vowels often represent just a single sound. This is not the case in Indonesian because each vowel must be clearly pronounced when speaking.

A consonant is a sound produced by the restricted flow of air in the vocal chords. There are 21 single consonants, consisting of /b/, /c/, /d/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /p/, /q/, /r/, /s/, /t/, /v/, /w/, /x/, /y/, /z/ and 15 double consonants (compounds), consisting of /dh/, /dl/, /dz/, /gh/, /kh/, /ks/,

/ph/, /ps/, /sh/, /sy/, /th/, /ts/, /ng/, /ny/, /ngg/. A compound is two or more adjacent consonants in a word. In Indonesian language, there are 15 distinct compounds.

E. Visemes

Visemes are represented by mouth shapes. Viseme is an equivalent unit in a visual domain which models a speech recognition system audio-visually [28]. It has many interpretations in literatures and there are some which disagree with how to define visemes. Actually, two practical approaches to define logically are: (1) Viseme is assumed to be an articulation movement such as closing the lips together, moving the jaw; and (2) Viseme is derived from a group of phonemes having similar visual appearances [5, 6, 7, 8, 9, 10]. Using the letter, visemes and phonemes are correlated through phoneme-to-viseme mapping. It has to be a many-to-one mapping, because many phonemes can not be distinguished using visual signals [4].

Several phonemes can correspond to a single viseme, since we can not visually differentiate between pronouncing “ba” and “pa”, both sounds are bilabial plosives and they have the same lip-based realization. A number of classifications exists for many languages such as French [29], English [30], German [1], modern classical Arabic [11], but there is no classification for Indonesian Language.

In 1960, Woodward and Barber proposed a phonemic articulatory gestures into a hierarchy of visual contrasts produced by a speaker. In 1968, Fischer introduced a viseme to represent the smallest lips movement and articulation corresponding to a sound [29, 11]. Until now, there has been no single classification or standard viseme set to a specific language such as phonemes [11].

As stated by [4] and all map properties from [5, 6, 7, 8, 9, 10] in contrast, vowel visemes in English are quite different from one researcher to another. The number of vowel visemes varies from 4 to 7, no specific cross-map patterns are present within maps. Some similarities are clearly present, particularly between [7] and [8] maps. In these two maps, 5 consonant classes are identical. Across all maps, the consonant classes show similar class separation.

Japanese language is composed of some very basic sounds that are constructed to make more sounds. Japanese has 5 basic visemes [24], it allows us to construct every mouth pose required in a Japanese speech, where tongue is used more often than in any other languages and requires less lips movements to make the language’s basic phonemes.

[1, 4, 11, 30] indicate that a number of visemes in other languages may have a number of different phonemes for each viseme, but they still phoneme-to-viseme mapping to the pattern of many-to-one map, an example is Arabic language which consists of 10 visemes from 31 phonemes, not including silent, whereas German language consists of 15 visemes from 42 phonemes.

III. RESEARCH METHODOLOGY

The method proposed in this research is aimed to produce the visualization of a 3D animated face which can be applied to the mouth shapes of the Indonesian viseme model. The construction of Indonesian viseme model is the first in the

research of making Indonesian Language Talking Head System and is expected to be a reference for future research.

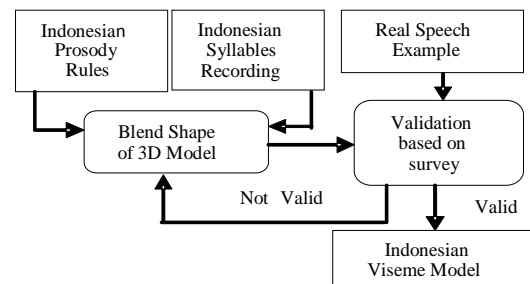


Fig.2. Indonesian Viseme Model Construction

The processes of Indonesian viseme model construction is to build mouth animation using blend shapes, as seen in Fig. 2 are:

- A. Construction of 3D Model using two input data: Indonesian prosody rules and the recording of Indonesian syllables read by a number of speakers.
- B. After the literature comprehension and speakers recording have been done, the 3D Model is constructed using makehuman program, which is read by blender program. This process is applied to the mouth shapes of the Indonesian visemes, including silent, translated in Tx,Ty,Tz blend shape parameters according to the expected viseme models.
- C. The next process is validation based on a survey using the input data of real speech examples from a number of speakers. If the constructed Indonesian viseme model is not valid, the process is repeated. If it is valid, the Indonesian viseme model output can be obtained. This last process will be explained in the next section.

A. Input Data

1) Definition of Indonesian Prosody Rules

Traditionally, phonetics is a study of speech sounds, i.e. the sounds produced by human vocal organs only, in the context of spoken language. However, it is widely accepted today that phonetics is emphasized more than just studying the properties of vowels and consonants that make up a spoken sentence. Prosody comprises all properties of speech that cannot be understood directly from a linear sequence of segments. The linguistic functions of prosody are: to mark off domains in time (e.g. paragraphs, sentences, phrases), to classify the information presented in a domain (e.g. as statement, question), and to highlight accentuation of certain constituents within these domains. [31]

The smallest domain that can be marked off is a syllable. In this paper, a syllable that ends with a vowel is called open syllable, and when it ends with a consonant, it is called closed syllable. Open syllables are pronounced longer than closed syllables. Prosody literally means accompaniment. This suggests that the segmental structure defines the verbal contents of the words, while prosody provides the melody and rhythm. Prosody is often divided into two broad categories of phenomena: temporal and melodic structures. The temporal structure of a language is a set of regularities that determines the duration of speech sounds and pauses in

utterances spoken. Melodic structure is defined as a set of rules that characterizes pitch variation over the course of utterances spoken in a given language, no two languages have the same melodic properties. [32]

2) The Classification of Indonesian Syllables

According to Indonesian Dictionary, syllable is a structure from one or a sequence of phonemes as part of a word. Each syllable is marked with a vowel (including diphthongs). In separating a word into syllables, we use the following guidelines: (a) if in the middle of the word, there are two successive vowels, the separation is placed before the second vowel, such as “*ba-ik*” (well); (b) if there is a consonant between two vowels, the separation is placed before the consonants, such as “*i-bu*” (mother); (c) for all compounds consisting of two letters, namely: “*ng*”, “*ny*”, “*sy*”, “*kh*” which represent a consonant, the separation is done before or after a series of letters, such as “*ang-ka*” (digit), “*nya-nyi*” (sing), “*a-khir*” (end); (d) if there are two consecutive consonants, the separation is made between the consonants, such as “*sam-ping*” (side); (e) if there are three or more consonants, the separation was placed between the first consonant (including “*-ng*”) and the second consonant, such as: “*ben-trok*” (clash), “*bang-krut*” (bankruptcy); and (f) if the word has a prefix, insert, or suffix, the basic word is separated following the rules above, such as: “*mem-prak-tek-kan*” (practice), “*ge-me-tar-an*” (trembling).

For the classification of Indonesian syllables where vowel and consonant are abbreviated as V and C respectively, there are eleven patterns: (1) V-pattern: “*a-nak*” (child); (2) VC-pattern: “*an-da*” (you); (3) CV-pattern: “*li-ma*” (five); (4) CVC-pattern: “*pin-tu*” (door); (5) CCV-pattern: “*in-fra*” (infra); (6) CCVC-pattern: “*trak-tor*” (tractor); (7) VCC-pattern: “*ons*” (ounce); (8) CVCC-pattern: “*teks-tur*” (texture); (9) CCVCC-pattern: “*trans-fer*” (transfer); (10) CCCV-pattern: “*stra-te-gi*” (strategy); (11) CCCVC-pattern: “*struk-tur*” (structure).

3) Concept of Indonesian syllables reading

The speech recording is made by ten speakers consisting of five male students and five female students of Indonesian nationality. They pronounced Indonesian syllables in the patterns of V, CV, CCV, VC, and VCC. Three types are used to analyze the effect of mouth shapes when the vowels are placed in front and the consonants at the back, and vice versa. Single and double consonants are considered to have the same type, because it just shows the phoneme to be taken from the set of Indonesian phonemes.

Type 1 : silent V silent

Type 2 : silent VC and VCC silent

Type 3 : silent CV and CCV silent

In the first type, we put a vowel before and after silent. For the second type, a vowel is placed before a consonant. In the third type, a vowel is placed after a consonant. All syllables are read by the speakers, containing 13 vowels (including diphthongs) for the first type, 10 vowels (excluding diphthongs) multiplied by 36 consonants (including compound phonemes) for the second type. In this second type, only 10 types of vowel are used because Indonesian diphthongs never appear in front of syllables, but always at the end. In the third type, all syllables contain 36

consonants (including compound) multiplied by 13 vowels. In total, the speakers read 13, 360, and 468 syllables from each type respectively. The reading order of the syllables is started and closed with silent.

B. Recording of Indonesian Syllables Reading

1) Video Recording

The recording took place in a laboratory using a digital camera with high resolution, in a controlled environment. The 10 speakers were seated facing the video camera mounted on a tripod one meter in front. The vertical position of the camera was adjusted to the level of the speaker's face. The experiment was done in the morning to give optimal lighting while minimizing the shadows of the mouth.

The Indonesian Syllables Recording to speakers is shown in Fig. 3.

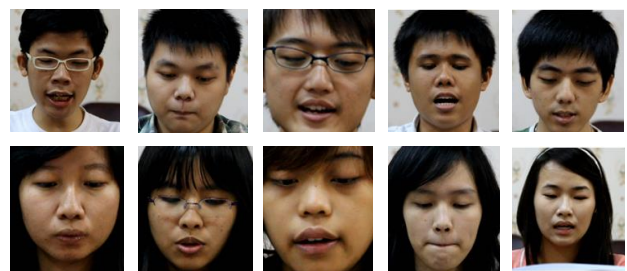


Fig. 3. Indonesian Syllables Recording

The syllables reading was taken at thirteen scenes for the first and the third types, while the second type in ten scenes. Each recording would stop if the speaker had read all the consonants (single and double) with the same vowel. Thus, each speaker pronounced 841 different syllables from all types, 36 scenes for each type. The phase of articulation of these phonemes is built from an open and a closed mouth. The goal is to obtain the perfect shape of the mouth.

2) Choice of Geometric Parameters from Blend Shape Animation

A blend shape animation is the method of choice for keyframe facial animation, namely a set of blend shapes (key facial expressions) which are used to define a linear space of facial expressions [33]. However, blend shapes need to be segmented into smaller regions. In this paper, we propose physically-motivated segmentation that learns the controls and parameters directly from the set of blend shapes. We also provide a rendering algorithm to enhance the visual realism of a blend shape model. [16]

Human face always holds a particular interest for the computer graphics community. Facial animation requires a deformable model of the face to express the wide range of facial configurations related to speech or emotions. A blend shape model directly considers every facial expression as a linear combination of a few selected facial expressions. By varying the weights of the linear combination, a full range of facial expressions can be made with little computation.

To show a significant range of highly detailed expression, digital animators have to create large libraries of blend shapes. By manipulating a smaller area, the user is guaranteed that the modification will impact only a specific part of the face, e.g. the mouth. The segmentation should


reflect the idiosyncracies of the face being modeled and provide editing and different level of details. A prototypical example is the segmentation of a face into an upper region and a lower region: the upper region is used for expressing emotions, while the lower region expresses speech. Blend shape models have helped computer vision community analyze face images and video [33].

Lips are very elastic, they can move toward different directions, constructing the shape of a cone while speaking a phoneme. Lips are very expressive in speaking articulation, so the modeling of lips and lips movements is the main requirement for a high quality facial animation system [34, 35, 36]. [37] separates lips from face model and lips special model to increase the realism of facial animation.

Lips model is controlled geometrically, externally, and internally upon 9 control points (CP), as seen in Fig. 4. External geometry contains all red lips area. Internal geometry contains the mucous membrane inside the lips. This geometry is very important for maintaining the realism when mouth is opened.

The control points of an animated facial model is adapted to makehuman program, it is divided into 10 CP around the face and 9 CP around the lips. The positions and the term names at control points on the face of 3D facial model can be found in Table III.

TABLE III
CONTROL POINTS (CP)

3D Facial Model with 19 CP	CP	CP in Mouth Area	CP	CP in Face Area
	CP-1	Pmouth_R	CP-10	Pbrow_R
	CP-2	Pmouth_L	CP-11	Pbrow_L
	CP-3	Pmouth_mid	CP-12	Pbrows
	CP-4	Puplip_R	CP-13	Puplid_R
	CP-5	Puplip_L	CP-14	Puplid_L
	CP-6	Puplip_mid	CP-15	Plolid_R
	CP-7	Plolid_R	CP-16	Plolid_L
	CP-8	Plolid_L	CP-17	Pcheek_R
	CP-9	Plolid_mid	CP-18	Pcheek_L
			CP-19	Pnose

C. Construction of 3D Model

1) Blend Shape Face Model

A blend shape face model is defined as a convex linear combination of n basis vectors, each vector is one of the blend shapes [33]. Each blend shape is a face model that includes geometry and texture. All blend shape meshes for a given model share the same topology. The texture at a particular point of the blend shape model is similar to a linear combination of the blend shape textures with the same blending weights as those used for the geometry.

The coordinates of a vertex V of the blend shape model can be written as Equation (1):

$$V = \sum_{i=1}^n \alpha_i V_i \quad (1)$$

where scalars α_i are the blending weights, V_i the locations of the vertices in the blend shape i , and n the number of blend shapes. These weights must satisfy the convex constraint and be equalized to 1 for rotational and translational invariance, as seen in Equation (2).

$$\alpha_i \geq 0, \text{ for all } i \text{ and } \sum_{i=1}^n \alpha_i = 1 \quad (2)$$

This transformation is applied to a group of points which construct an object. If the object changes, the locations of the points will also change from their initial locations to new locations. 3D vertex transformation is basically the same as vertex transformation on 2D objects, consisting of x-axis (length) and y-axis (width). In 3D transformation, z-axis is added, which gives the impression of depth in the sight of human eyes. The depth can be considered as the distance from the eyes of the viewer to the object.

2) Building Mouth Animation using Blend Shapes based on Video Recording

The steps of building mouth animation using blend shapes based on video recording are as follows:

- 1) Library in Fig. 3 consists of video recordings used as reference for natural mouth shapes of phonemes articulations (see Fig. 3).
- 2) 3D mesh model of neutral mouth shape (silent) will be used as the initial of blend shape.
- 3) Create mouth shapes for each phoneme by deforming vertices, manually or by deriving deformations from facial motion capture device. The input of this process is the neutral shape from step 2. Manual deformations should use video recording from step 1 as a reference.
- 4) Mouth shapes from step 3 can be combined with certain proportions to create combinations of shapes.
- 5) For combining shapes in step 4, control points can be utilized for convenience. Control points can be dragged to adjust mouth shapes.

In step 3 we get variations of mouth shapes that will be used as morph targets for each phoneme. From these shapes, other shapes can be built by combining them with certain proportions (step 4 through 5).

Custom shapes for control points can be used as more friendly user interface for animators, e.g. lower jaw control point can be visualised as jawlike arc is shown in Fig. 4.

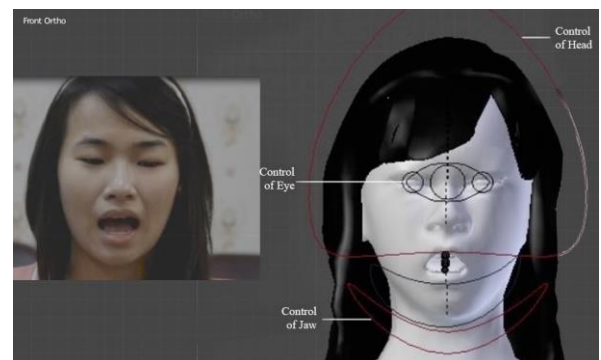


Fig. 4. Custom Shapes for Control Points can be Used for Animators

Mathematically, the result of step 3 is a set of vertices translation from neutral shape to each shape of morph targets. The set of the vertices translation is similar to the result of automatic face-capturing device. For creating more complex mouth shapes, some of the mouth shapes in step 3 can be combined with certain proportions. Combinations of two or more mouth shapes can be done by taking the

average of the certain weights to each translation. This step is possible because mouth shapes are a set of vertices translation. Building mouth animation using blend shapes based on video recording from Fig. 4, is shown in Fig. 5.

Fig. 5 is showing the results of rendering of the blend shape 3D Indonesian viseme model using makehuman in Fig. 5, (a) is part of texture of makehuman, (b) is one of example 3D of facial model for animator, (c) is result of rendering from 3D of facial model and (d) is red lips area from 3D of facial model with 9 CP in mouth area (as seen in Table III).

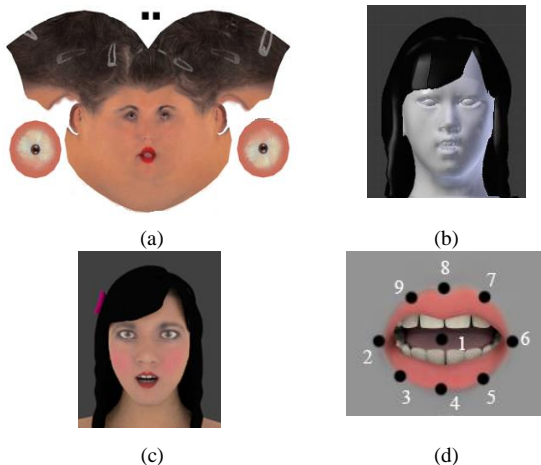


Fig. 5. Rendering of Blend Shape 3D Indonesian Viseme Model Using makehuman: (a) Texture of makehuman, (b) 3D of Facial Model, (c) Result of Rendering from (b), (d) Red Lips Area from (c) with 9 CP in Mouth Area

IV. EXPERIMENTAL RESULTS

A. Indonesian Phoneme-Viseme Mapping

According to the listed Indonesian Phoneme Set and IPA Indonesian vowels and consonants, we now need to map the the phoneme units to a viseme set of symbols which will represent the visual sequence according to the spoken utterance from the real speech example. Furthermore, we display the steps we did in order to create an appropriate mapping of the phoneme and viseme segments of the three-dimensional synthesis based on linguistic approach.

Indonesian phoneme set based on consonants and vowels classification is shown in Table IV. It is obtained from the previous discussion on Section 3. The discussion starts with the type of consonant phonemes. The first class of viseme includes three phonemes /p/, /b/, and /m/, similar to English synthesis systems. We combine the three phonemes, because bilabial plosives and bilabial nasal have similar place of articulation and all lips stay close for final position, although visually the sound of /m/ is different from the other plosives.

In the second class of viseme in Table IV, phoneme /ph/, included as part of a loanword in Indonesian, is read /f/. It is combined with two labiodental fricatives /f/ and /v/. While, phoneme /w/ is also included in this class, because /w/, bilabial semivowel, has a slightly open mouth shape in the middle, similar to three phonemes /f/, /v/ and /ph/.

The alveolar plosives /t/, /d/, /dh/, /dl/, /dz/, /th/, nasal /n/ and lateral /l/ are put together in the third class of visemes, although they differentiate manners of articulation, but since

this difference is produced within the mouth and in front of the tongue, it is not visually distinguishable. While the fourth class of viseme contains only the phonemes /r/, because /r/ is alveolar trill, where the front of the tongue as an articulator is in the palate and the mouth is slightly open.

The fifth class of viseme contains twelve phonemes, /c/, /j/, /s/, /z/, /ts/, /ps/, /ks/, /sh/, /sy/, /x/, /y/, /ny/ which have different places and manners of articulation. Interestingly, many phonemes are mapped by one viseme. In Indonesian language, phonemes /c/, /j/, /z/, /x/, /y/, /ny/ never appear at the end of the word. Furthermore, phonemes /ts/, /ps/, /ks/, /sh/, /sy/ at the end of the word are very rarely used in Indonesian language because they are part of the loanword. However, if we must create the mouth shapes of these phonemes, they can be replaced by other phonemes which have similar mouth shapes to that of phoneme /s/ because the articulators are dominated by the front and the middle of tongue of fricative articulation manners.

TABLE IV
INDONESIAN PHONEME SET BASED ON CONSONANTS AND VOWELS
CLASSIFICATION







Type of Phoneme	Phoneme Class	Number of Phoneme	Set of Phoneme	Viseme Class
Consonants	1	3	/b/, /m/, /p/	V1
	2	4	/f/, /v/, /w/, /ph/	V2
	3	8	/d/, /dh/, /dl/, /dz/, /l/, /n/, /t/, /th/	V3
	4	1	/r/	V4
	5	12	/c/, /j/, /s/, /z/, /ts/, /ps/, /ks/, /sh/, /sy/, /x/, /y/, /ny/	V5
	6	8	/g/, /gh/, /h/, /k/, /kh/, /q/, /ng/, /ngg/	V6
Vowels	7	1	/a/	V7
	8	3	/i/, /I/, /oi/	V8
	9	4	/ə/, /e/, /ɛ/, /ai/	V9
	10	3	/o/, /ɔ/, /au/	V10
	11	2	/u/, /U/	V11
-	-	-	silent	V12
Total		49		

The sixth class of viseme or the last type of consonants phoneme, /g/, /k/, /kh/, /ng/, /ngg/, /gh/, /q/; and /h/, have different places of articulation, namely velar, uvular and glottal (as seen in Table II) respectively. We put them into one viseme class, although phonemes /q/ and /ngg/ are never placed at the end of the word. Nevertheless, the mouth shape of these phonemes are visually distinguishable.

A vowel is a sound produced by the unrestricted flow of water in the vocal chords. Vowels are important, because nearly every word has at least one vowel. In this paper, we combine the phonemes according to the height of the tongue's body and its front-back position. The quality of a vowel is generally determined by three things, namely: (1) shape of lips open-rounded, (2) top-down tongue, and (3) forward-reverse movement of tongue.

As stated in the previous sections, from the five letters of vowels, consisting of "a", "e", "i", "o", "u" in Indonesian language, there are six primary vowels represented by six phonemes /a/, /e/, /ə/, /i/, /o/, /u/. Consequently, we divide thirteen phonemes including diphthongs in the classes of vowels phoneme in accordance with the five vowel letters group based on the similarity of the pronunciation.

TABLE V
AN EXAMPLE OF SURVEY RESPONSE

Viseme Class	Mouth Shapes of 3D Animation	n th Phoneme	Phoneme (how to read)	3D Visualization of Mouth Shapes			
				1 Strongly Disagree	2 Disagree	3 Agree	4 Strongly Agree
V1		1	p (əp)				√
		2	b (əb)			√	
		3	m (əm)			√	
V2		4	f (éf)				√
		5	v (év = éf)		√		
		6	w (éw)			√	
		7	ph (éf)		√		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
V9		41	ə (əh)				√
		42	e (eh)				√
		43	ε (éh)			√	
		44	ai (pantai=panté)		√		
V10		45	o (oh)				√
		46	ɔ (Oh)			√	
		47	au (puO)				√
V11		48	u (uh)				√
		49	U (Uh=uh)			√	
V12		50	Silent (neutral)				√

B. Validation Based on Survey

1) Survey Instrument

In this research, we made a questionnaire with scaling method as the major statistical tool in model testing. Scaling is the process of measuring with respect to quantitative attributes or traits. To ensure the scales content validity, the measurement item was developed based on relevant study.

In the questionnaire, there are 12 questions, each question represents one viseme class. The twelve questions represent the visualization of viseme to mouth shape of the 3D animation models. We use 3D visualization of mouth shapes containing 50 items that should be filled based on the Likert scale. These fifty items represent 49 phonemes and 1 silent which the respondents must match with the form of visualization. Each item was measured by 4-point bipolar Likert scale with end points: (1) strongly disagree, (2) disagree, (3) agree and (4) strongly agree on an assessment of Indonesian Phoneme-Viseme Mapping. An example of the survey response is presented in Table V.

An important distinction must be made between a Likert scale and a Likert item. Likert scale is the sum of the responses on several Likert items, while Likert item is the statement which the respondents were asked to evaluate according to a certain criterion. The level of agreement or disagreement was measured generally. A good Likert scale is considered symmetrical because there are equal numbers of positive and negative options for a statement. In this questionnaire, the middle option of "neither agree nor disagree" when the respondent is unsure is not available.

2) Data Collection

First, we sent the questionnaire files via email to the master's and doctorate students at two famous universities in Surabaya, Indonesia, and also to several graduates of these two universities. In the files, we provided preliminary

information about our research, then we asked them to look at it carefully and to fill in the questions with the truth.

TABLE VI
THE PROFILE OF RESPONDENTS

Gender	Male: 77 (77%), Female: 23 (23%)
Age (Years)	Mean: 30.61, SD: 2.22
Educational status	Student: 70 (70%), Not Student: 30 (30%)
Status of Student	Undergraduate: 27 (38.57%), Master: 32 (45.71%), Ph.D: 11 (15.71)
Status of Not Student	Under graduate: 9 (30%), Graduate: 21 (70%)
Job Career	Lecturer/Teacher: 54 (54%), Field of IT: 11 (11%), Others: 35 (35%)

120 questionnaires files were sent and 100 were returned. All files have been filled out completely and answered with consistency, so no one was dropped. The effective return rate is 83,33%. All of them were valid. The demography of the participants are presented in Table VI.

TABLE VII
SCORE OF RESPONSES FROM RESPONDENTS

Viseme Class	Number of Phonemes	Number of Responses				Total
		(1)	(2)	(3)	(4)	
V1	3	27	46	56	171	300
V2	4	32	87	105	176	400
V3	8	62	127	240	371	800
V4	1	2	17	20	61	100
V5	12	64	243	374	519	1200
V6	8	41	135	225	399	800
V7	1	1	5	13	81	100
V8	3	32	71	68	129	300
V9	4	34	75	114	177	400
V10	3	6	20	66	208	300
V11	2	5	1	25	169	200
V12	1	1	9	19	71	100
Total	50	307	836	1325	2532	5000

(1).Strongly Disagree, (2).Disagree, (3).Agree, (4).Strongly Agree

The data collection of the questionnaire is shown in Table VII. Data scores of this table is given by the respondent. Columns (1) to (4) in Table VII shown number of responses from respondents, start from strongly disagree until strongly agree using Likert scale. Total column of each viseme class, meaning multiplication of the number of phonemes to the number of respondents. Total rows (5000), meaning multiplication of total phonemes (50, include silent) with number of respondents have filled questionnaire (100 respondents).

While, Table VIII is a summary of the data entry field from Table VII and consists of the percentage of responses from 100 respondents using Likert scale for 50 phonemes (including sillent) in 12 questions (viseme class).

TABLE VIII
PERCENTAGE OF RESPONSES FROM RESPONDENTS

Viseme Class	% Responses from Respondents				Total
	(1)	(2)	(3)	(4)	
V1	9	15.33	18.67	57	100
V2	8	21.75	26.25	44	100
V3	7.75	15.88	30.97	44.65	100
V4	2	17	20	61	100
V5	5.33	20.25	31.17	43.25	100
V6	5.13	16.88	28.13	49.88	100
V7	1	5	13	81	100
V8	10.67	23.67	22.67	43	100
V9	8.50	18.75	28.5	44.25	100
V10	2	6.67	22	69.33	100
V11	2.50	0.50	12.5	84.50	100
V12	1	9	19	71	100
Total	63,13	171,17	272,84	692,86	1200
AVG	5,26	14,26	22,74	57,74	100

(1).Strongly Disagree, (2).Disagree, (3).Agree, (4).Strongly Agree

3) Instrument Validity and Reliability

The properties of the instrument are assessed in term of internal consistency, validity and reliability. The values of reliability coefficients range from 0 to 1.0. A coefficient of 0 means no reliability and since all tests have some errors, the coefficients never reach 1.0. Generally, the reliability of a standardized test is above 0.8, which is said to be very reliable. If it is below 0.5, it is not considered a reliable test.

Validity refers to the accuracy of an assessment, whether or not it measures what it is supposed to measure. Even if a test is reliable, it may not provide a valid measure. The validity inferred from the assessments is essential.

Instrument Validity and Reliability is measured using equation (3), where r is the correlation of variables between x and y , x_k and y_k are variable between phonemes, $k = 1, 2, \dots, N$, and N is the number of phonemes (including silent).

$$r = \frac{N \sum_{k=1}^N x_k y_k - \left(\sum_{k=1}^N x_k \right) \left(\sum_{k=1}^N y_k \right)}{\sqrt{N \sum_{k=1}^N x_k^2 - \left(\sum_{k=1}^N x_k \right)^2} \sqrt{N \sum_{k=1}^N y_k^2 - \left(\sum_{k=1}^N y_k \right)^2}} \quad (3)$$

C. Data Analysis and Results

1) Validity

Validity refers to the degree to which a study accurately reflects the specific concept that the researcher attempts to

measure. Based on the data from the respondents, the results show that all entry data are valid. The results come from the calculation of 50 question fields, each consists of scoring 1, 2, 3 or 4. For 100 respondents, the number of the entry data is 50×100 or 5000. By calculating the correlation of r and comparing it to 0.576 (the value of the product moment using 5% level of significance), the result is 96.19%, which shows a high validity and strong classification.

2) Reliability

Reliability is the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials. Reliability refers to the extent to which assessments are consistent. We did not compare Cronbach's alpha to measure the reliability, because the data from the questionnaires consist of 1 to 4 scoring only, or strongly disagree until strongly agree. The result of the reliability test is 96.85%. This result shows that the data are very reliable.

3) Rating Index, Mean Rating and Mean Square Error

Rating Index, commonly known as the RI, is a quantity used to rank a particular item, but it lacks theoretical justification from a statistical standpoint. Table IX shows the data rating index of viseme based on the number of phonemes. However, RI can be used to measure the index of success in a research involving opinions of respondents, whereas the mean rating index (MRI) is the average of all RI measured.

TABLE IX
RATING INDEX OF VISEME FROM RESPONDENT

Viseme Class	Number of Phoneme	RI of Viseme for Each Phonemes	Mean Rating Index
V1	3	3.17; 3.07; 3.47	3,24
V2	4	3.15; 3.21; 2.96; 2.93	3,06
V3	8	3.27; 3.13; 3.02; 3.09; 2.96; 3.17; 3.24; 3.32	3,12
V4	1	3.40	3,40
V5	12	3.25; 3.26; 3.13; 3.33; 3.24; 3.33; 2.96; 2.97; 3.04; 2.95; 3.07; 2.95	3,12
V6	8	3.35; 3.30; 3.17; 3.26; 3.24; 3.20; 3.14; 3.16	3,23
V7	1	3.74	3,74
V8	3	3.15; 3.08; 2.71	2,98
V9	4	3.22; 3.14; 2.99; 2.99	3,09
V10	3	3.65; 3.62; 3.49	3,59
V11	2	3.83; 3.75	3,79
V12	1	3.60	3,60
		Average	3,33

Table IX shows that the average MRI is 3.33. The results of the average value indicate that the MRI range is between 3.00 and 4.00. Based on the Likert scale, this is a good result because the average value of MRI is between agree and strongly agree. Thus, the representation of the mouth shape of viseme with visualization of 3D models is similar.

Mean Squared Error (MSE) is used to measure the performance of our RI between mouth shapes and phoneme-to-viseme mapping. MSE is found based on the difference between the value of MRI observation and expectation. The lower the value of MSE is, the better the performance is.

RI, MRI and MSE can be measured by equation (4), (5), and (6) respectively. ω_j is the weight of Likert scale ($\omega_1=1$, $\omega_2=2$, $\omega_3=3$, $\omega_4=4$, which shows the rating, ranging from

strongly disagree to strongly agree); S_j is the number of respondents at each weight; $j = 1, 2, 3, 4$; $k = 1, 2, \dots, 50$; $z = 1, 2, \dots, 12$; M is the number of all respondents; $P(z)$ is the number of phonemes at the z^{th} viseme class, including silent; and $Q(k)$ is the number of phoneme sets including silent at a viseme class.

$$RI(z) = \frac{1}{M} \sum_{j=1}^4 \omega_j S_j, \text{ where } M = \sum_{j=1}^4 S_j \quad (4)$$

$$MRI(z, k) = \frac{1}{P(z)} \left[\sum_{z=1}^{12} RI(z) \right] \quad (5)$$

$$MSE = \frac{1}{P(z)Q(k)} \sum_{z=1}^{12} \sum_{k=1}^{50} [MRI_{obs}(z, k) - MRI_{exp}(z, k)]^2 \quad (6)$$

Table X shows the MSE for each viseme. This table calculates the rating index of viseme. This is very important, because we can answer these following questions: (1) What is the mean rating index of each viseme? and (2) Does the visualization of 3D facial animation of viseme represent the mapping of phoneme to viseme appropriately?

TABLE X
MEAN SQUARED ERROR OF RESPONSES FROM RESPONDENT

Viseme Class	MRI	Mean Rating Expected	Error	Abso lute Value	Square Error	MSE
V1	3,24	4	-0,76	0,76	0,5776	0,0116
V2	3,06	4	-0,94	0,94	0,8836	0,0177
V3	3,12	4	-0,88	0,88	0,7744	0,0155
V4	3,4	4	-0,60	0,60	0,3600	0,0072
V5	3,12	4	-0,88	0,88	0,7744	0,0155
V6	3,23	4	-0,77	0,77	0,5929	0,0119
V7	3,74	4	-0,26	0,26	0,0676	0,0014
V8	2,98	4	-1,02	1,02	1,0404	0,0208
V9	3,09	4	-0,91	0,91	0,8281	0,0166
V10	3,59	4	-0,41	0,41	0,1681	0,0034
V11	3,79	4	-0,21	0,21	0,0441	0,0009
V12	3,6	4	-0,40	0,40	0,1600	0,0032
Avg	3,33	4	-0,67	0,67	0,4489	0,01045

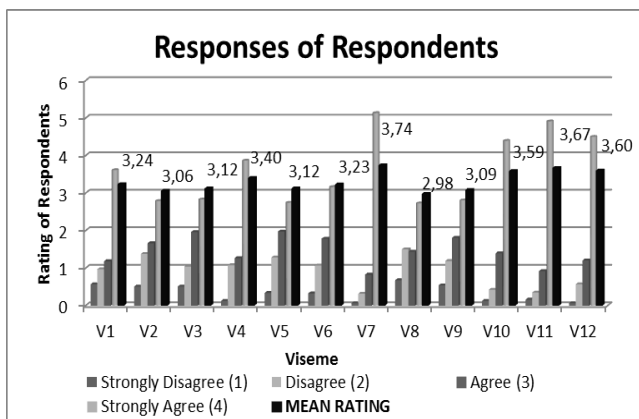


Fig. 6. Graph of Rating of Responses from Respondents

Fig. 6 is derived from Table VIII and Table IX. It shows a graph of the percentage of responses from respondents, showing the relationship between visemes and the scores given by the respondents. Most respondents gave good scores between agree and strongly agree for each viseme.

Fig. 6 also shows the mean rating of each viseme. High mean rating indicated that the mapping of phoneme to viseme for Indonesian language has been well represented.

The results of the average value of MSE is 0.01045, indicating that the responses from the respondents are qualified in accordance with the value of the condition, which is only about 1%, less than 5% ($MSE \leq 0.05$). Thus, it indicates that the error in the measurement is small.

V. CONCLUSION

Based on the nine geometrical parameters from blend shape animation of mouth shapes around the lips of an animated facial model, we establish a distinct phoneme-to-viseme mapping for Indonesian language, as can be seen in Table XI. It is possible to recognize a vowel and determine the consonant analysis. For our future research, we plan to use this outcome to explore other geometric parameters to develop a more refined class of Indonesian Viseme.

The result of phoneme-to-viseme mapping for Indonesian language is based on linguistic approach and validated through a survey, 12 Indonesian visemes have been produced including silent. These 12 are considered complete enough to represent all 49 Indonesian phonemes. The error measured is relatively small, the average of the mean rating index for all viseme is good, and the correlation value obtained is very high, close to 1, which shows a good result.













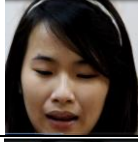



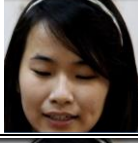







ACKNOWLEDGMENT

The author would like to appreciate the promoter of Tokyo University of Technology, Hachioji, Tokyo, Japan, who has accepted and given the author an opportunity to stay in Japan for 4 months in a quality improvement program of international publications in accordance with the Memorandum of Understanding between Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, and Tokyo University of Technology, Hachioji, Tokyo, Japan.

REFERENCES

- [1] Bianca Aschenberger and Christian Weiss, "Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis," IKP-Working Paper NF 11, Institut für Kommunikationen für Schung und Phonetik, Universität Bonn, 2005.
- [2] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, No. 5588. December 23, 1976, pp. 746-748.
- [3] M. M. Cohen and D.W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech," In *Magnat Thalmann, N. and Thalmann, D.*, editors, *Models and Techniques in Computer Animation*, pp. 139-156. Springer, Tokyo, 1993.
- [4] Luca Cappelletta and Naomi Harte, "Phoneme-to Viseme Mapping for Visual Speech Recognition," *Proceeding of the 2012 International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012)*, February 7, 2012.
- [5] G. Potamianos, *et al.*, "Recent advances in the automatic recognition of audio-visual speech," *Proceeding of the IEEE*, vol. 91, no. 9, 2003.
- [6] K. Saenko, "Articulatory Features for Robust Visual Speech Recognition," Master Thesis, Massachusetts Institute of Technology, 2004.
- [7] Bianca Aschenberger and Christian Weiss, "Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis," IKP-Working Paper NF 11, Institut für Kommunikationen für Schung und Phonetik, Universität Bonn, 2005.

TABLE XI
INDONESIAN PHONEME TO VISEME WITH MEAN RATING

Viseme Class	Phone me Set	Mean Rating	Mouth of Human Model	Mouth of 3D Animation	Viseme Class	Phone me Set	Mean Rating	Mouth of Human Model	Mouth of 3D Animation
V1	b, m, p	3,24			V7	a	3,74		
V2	f, v, w, ph	3,06			V8	i, I, oi	2,98		
V3	d, dh, dl, dz, l, n, t, th	3,12			V9	ə, e, ε, ai	3,09		
V4	r	3,4			V10	o, O, au	3,59		
V5	c, j, s, z, ts, ps, ks, sh, sy, x, y, ny	3,12			V11	u, U	3,79		
V6	g, gh, h, k, kh, q, ng, ngg	3,23			V12	silent	3,6		

- [8] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, No. 5588, December 23, 1976, pp. 746-748.
- [9] M. M. Cohen and D.W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech," In *Magnat Thalmann, N. and Thalmann, D.*, editors, *Models and Techniques in Computer Animation*, pp. 139-156. Springer, Tokyo, 1993.
- [10] Luca Cappelletta and Naomi Harte, "Phoneme-to Viseme Mapping for Visual Speech Recognition," *Proceeding of the 2012 International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012)*, February 7, 2012.
- [11] G. Potamianos, *et al.*, "Recent advances in the automatic recognition of audio-visual speech," *Proceeding of the IEEE*, vol. 91, no. 9, 2003.
- [12] K. Saenko, "Articulatory Features for Robust Visual Speech Recognition," Master Thesis, Massachusetts Institute of Technology, 2004.
- [13] Jeffers and Barley, "Speechreading (Lipreading)," Charles C Thomas Pub Ltd, 1971.
- [14] C. Neti, G. Potamianos, J. Luetin, I. Matthews, H. Glotin, D. Vergyri, S. Sison, A. Mashari, and J. Zhou, "Audio-visual Speech Recognition," Tech. Rep., October 12, 2000.
- [15] Hazen, Saenko, La, and Glass, "A segment-based audio-visual speech recognizer: data collection, development, and initial experiments," In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 235-242, State College, PA, USA. ACM, 2004.
- [16] Elif Bozkurt, Cigdem Eroglu Erdem, Engin Erzin, Tanju Erdem, Mehmet Ozkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation," *Proceeding in 3DTV of the 2007 IEEE International Conference*, pp. 1-4, 2007.
- [17] Pascal Damien, Nagi Wakim, and Marcel Egéa, "Phoneme-Viseme Mapping for Modern, Classical Arabic Language," *Proceeding of the The 2009 IEEE International Conference on Advances in Computational Tools for Engineering Applications (ACTEA 2009)*, Zouk Mosbeh, Lebanon, July 14-17, 2009, pp. 547-552.
- [18] Salah Werda, Walid Mahdi and Abdelmajid Ben Hamadou, "Lip Localization and Viseme Classification for Visual Speech Recognition," *International Journal of Computing and Information Sciences*, Vol. 5, No. 1, April 2007, On-Line, pp. 62-75.
- [19] F.I. Parke, "Computer Generated Animation of Faces", *Proceedings ACM annual conference.*, August 1972.
- [20] F.I. Parke, "A Parametric Model for Human Faces," PhD Thesis, University of Utah, Salt Lake City, Utah, December 1974. UTEC-CSC-75-047.
- [21] T. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," In *SIGGRAPH 99 Conference Proceedings*. ACM SIGGRAPH, August 1999.
- [22] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, "Synthesizing Realistic Facial Expressions from Photographs," In *SIGGRAPH 98 Conference Proceedings*, pages 75-84. ACM SIGGRAPH, July 1998.
- [23] F. Pighin, R. Szeliski, and D.H. Salesin, "Resynthesizing Facial Animation through 3D Model-based Tracking," In *Proceedings, International Conference on Computer Vision*, 1999.
- [24] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun and Frédéric Pighin, "Learning Controls for Blend Shape Based Realistic Facial Animation," *Eurographics/SIGGRAPH Symposium on Computer Animation*, D. Breen, M. Lin (Editors), 2003.
- [25] James Neil Sneddon, "The Indonesian Language: Its History and Role in Modern Society," UNSW Press, 2004, pp. 14
- [26] James Neil Sneddon, "The Indonesian Language: Its History and Role in Modern Society," UNSW Press, 2003, pp. 70
- [27] George Quinn, "Bahasa Indonesia: The Indonesian Language", Australian National University.
<http://www.hawaii.edu/sealit/Downloads/>.
- [28] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, "Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)," Balai Pustaka, Jakarta, Indonesia, 2003.
- [29] Sakriani Sakti, Eka Kelana, Hammam Riza, Shinsuke Sakai, Konstantin Markov, Satoshi Nakamura, "Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-

STAR Project,” In *Proceeding TCAST*, Hyderabad, India, January 2008, pp. 19-24.

- [30] Alvin Sebastian Hoo, MA, “Computer Animation,” <http://www.ordix.com/pfolio/research/>, published in 2002.
- [31] International Phonetic Alphabet IPA. Available <http://www.langsci.ucl.ac.uk/ipa/fullchart.html>
- [32] <http://aanugraha.wordpress.com/2008/03/13/ipa-indonesian-vowels-chart/> published at March 13, 2008
- [33] Anna Marietta da Silva, “The English Borrowings and the Indonesian-English Codeswitching in Two Collections of Blog Short-Stories,” ISSN 1411-2639 (Print), ISSN 2302-6294 (Online) OPEN ACCESS, 2013. DOI: [10.9744/k@ta.15.2.9-18](https://doi.org/10.9744/k@ta.15.2.9-18), pp. 9-17.
- [34] Arifin, Muljono, Surya Sumpeno, and Mochamad Hariadi, “Towards Building Indonesian Viseme: A Clustering-Based Approach,” *Proceeding of the 2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, Yogyakarta, Indonesia, 3-4 Dec 2013, pp. 57-61.
- [35] K. Nielsen, “Segmental Differences In The Visual Contribution Tospeech Intelligibility,” In *Proceeding ISCA-INTERSPEECH 2004 (ICSLP)*, October 4-8, 2004, pp. 1-4.
- [36] C. Benoit, T. Lallouache, T. Mohamadi, A. Tseva et C. Abry, “Nineteen (±two) French Visemes for Visual Speech Synthesis,” In *The ESCA Workshop on Speech Synthesis Autrans (SSW1)*, France, September 25-28, 1990, pp. 253-256.
- [37] Chen, T., & Rao, R. R, “Audio-visual integration in multimodal communication,” *Proceedings of the IEEE, USA*, 1998, pp. 837-852.
- [38] Vincent J. Van Heuven, “Introducing prosodic phonetics,” In *V.J. van Heuven & C. Odé (eds) Phonetic studies of Indonesian prosody*. Semaian, 9. Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië, Leiden University, 1-26, 1994.
- [39] Jialin Zhong, Wu Chou, Eric Petajan, “Acoustic Driven Viseme Identification for Face Animation,” *Proceeding of the IEEE First Workshop on Multimedia Signal Processing*, June 23-25, 1997, pp.7-12.
- [40] Jui-Chen Wu, Yung-Sheng Chen, and I-Cheng Chang, “An Automatic Approach to Facial Feature Extraction for 3-D Face Modeling,” *IAENG International Journal of Computer Science*, vol. 33, no. 2, pp1-7, 2007.
- [41] Surya Sumpeno, Mochamad Hariadi, and Mauridhi Hery Purnomo, “Facial Emotional Expressions of Life-like Character Based on Text Classifier and Fuzzy Logic,” *IAENG International Journal of Computer Science*, vol. 38, no. 2, pp122-133, 2011.
- [42] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun and Frédéric Pighin, “Learning Controls for Blend Shape Based Realistic Facial Animation,” *Eurographics/SIGGRAPH Symposium on Computer Animation*, D. Breen, M. Lin (Editors), 2003.
- [43] King, S. A., Parent, R. E., and Olsafsky, B, “An Anatomically-based 3D Parametric Lip Model to Support Facial Animation and Synchronized Speech,” In *Proceedings of Deform 2000*, pages 7-19.



Endang Setyati earned her bachelor’s degree in Mathematics from Institut Teknologi Bandung, Indonesia, in 1992 and master’s degree in IT from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2000. She started her doctorate study at the Graduate School of Electrical Engineering, ITS, in 2011. She is faculty member at the IT Department, Sekolah Tinggi Teknik Surabaya. Her research interests are mathematical logic, digital image processing and human

computer interaction. She is an IAENG and IEEE student member.



Surya Sumpeno earned his bachelor’s degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember, Surabaya, in 1996, and MSc degree from the Graduate School of Information Science, Tohoku University, Japan in 2007. He earned doctor degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember, Surabaya, in 2011. His research interests include natural language processing, human computer interaction and artificial

intelligence. He is an IAENG and IEEE member.



Mauridhi Hery Purnomo earned his bachelor degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1985, then his M.Eng., and Ph.D degrees from Osaka City University, Osaka, Japan in 1995 and 1997 respectively. He joined ITS in 1985 and has been a Professor since 2004. His current interests include intelligent system applications, electric power systems operation, control and management. He is an IAENG and IEEE Member.



Koji Mikami earned his Bachelor, Master and Ph.D degrees from Keio University, Tokyo, Japan between 1991 and 2008. He joined School of Media Science, Tokyo University of Technology, Tokyo University of Technology, Hachioji, Tokyo, Japan in 2007 and has been an Associate Professor since 2012. His interests include 3D Computer Animation, Planning and Producing of Game and Motion Picture, Contents Production Support and Management System. He is an ACM SIGGRAPH Member.



Masanori Kakimoto earned his bachelor and Ph.D degrees from The University of Tokyo, Tokyo, Japan in 1982 and 2005 respectively. He joined School of Media Science, Tokyo University of Technology, Hachioji, Tokyo, Japan and has been a Professor since 2012. His current interests include Computer Graphics, Visual Simulation, and Visualization. He is an ACM SIGGRAPH Member.



Kunio KONDO is a Professor at the School of Media Science, Tokyo University of Technology. He earned his Bachelor degree from Nagoya Institute of Technology in 1978 and Dr. Eng from the University of Tokyo in 1988. He is the former Associate Professor at the Department of Information and Computer Sciences, Saitama University in 1989-2007, faculty member at Tokyo Polytechnic University in 1988-1989, technical staff at Nagoya University in 1973-1988, part-time teacher at Tokyo University in 1991-2007, Aichi Prefectural University of Fine Arts and Music in 1989-1999, and Kyushu Institute of Design in 2002-2010. His research interests are computer graphics, animation, game, and interactive modelling. He won the IPSJ Anniversary Best Paper Award in 1985, JSGS Research Award in 1985, and JSGS Best Paper Award in 2011. He is the President of The Institute of Image Electronics Engineers of Japan, former President of The Society for Art and Science, former Vice President of Japan Society of Graphic Science, and Chair of SIG on Computer Graphics and CAD of Information Processing Society of Japan, Board member of Asia Digital Art and Design Association.