A Realistic Visual Speech Synthesis for Indonesian Using a Combination of Morphing Viseme and Syllable Concatenation Approach to Support Pronunciation Learning

https://doi.org/10.3991/ijet.v13i08.8084

Aripin(✉), Hanny Haryanto
Universitas Dian Nuswantoro, Semarang, Indonesia
arifin@dsn.dinus.ac.id

Surya Sumpeno
Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

Abstract—This research aims to build a realistic visual speech synthesis for Indonesian so that it can be used to learn Indonesian pronunciation. In this research, We used the combination of morphing viseme and syllable concatenation method. The morphing viseme method is a process of deformation from one viseme to another, so that the animation of the mouth shape looks smoother. This method is used to create the transition of animation between viseme. The Syllable Concatenation method is used to assemble viseme based on certain syllable patterns. We built a syllable-based voice database as a basis for synchronization between syllables, speech and viseme models. The method proposed in this research consists of several stages, namely the formation of Indonesian viseme models, designing facial animation character, development of speech database, a synchronization process and subjective testing of the resulting application. Subjective tests were conducted on 30 respondents who assessed the suitability and natural movement of the mouth when uttering the Indonesian texts. The MOS (Mean Opinion Score) method is used to calculate the average of respondents' scores. The MOS calculation results for the criteria of Synchronization and naturalness are 4,283 and 4,107 on the scale of 1 to 5. This result shows that the level of Synchronization and naturalness of the synthesis of visual speech is more realistic. Therefore, the system can display the visualization of phoneme pronunciation to support learning Indonesian pronunciation.

Keywords—morphing viseme, realistic, syllable concatenation, visual speech synthesis for Indonesian

1 Introduction

The computer technology is growing very fast, the need for flexibility and futuristic capability, especially in the field of human computer interaction is also increasing

[1]. It also occurs in character animation, games, biomechanical analysis, and others. In the field of animation, the demand for more flexible and realistic animation is higher. Animation must be able to display characters that are very similar to the character in the real world. The face animation is one of the focus of animation development. To produce a realistic face animation takes a long time for the animator due to the complexity of the facial animation. Facial animation also required a different level of accuracy compared with other parts of the human body [2].

Recently, natural and realistic facial animation is one of the most challenging research areas [3]. The effort to produce natural and realistic face animation is to add transition animations between pronunciations or apply dynamic viseme to the visual speech synthesis [4]. Some applications can be developed for this field are face character animation for animation film production, sound therapy, human computer interaction system design. Examples of the applications are Face Plus 3D Facial Animation, CAPT (Computer Aided Pronunciation Technology), and others. The talking-head animation is an important part of the face character animation. In general, talking-head animation displays visualization of phoneme pronunciation (viseme) synchronized with phoneme pronunciation sound.

Viseme is a visualization of phoneme pronunciation [5]. Implementation of viseme is often found in the production of animation films, especially talking-head animation. There are two ways to define a viseme, namely by a linguistic and data driven approach [6]. Defining viseme classes based on linguistic knowledge can be done manually by identifying the viseme class based on visual similarity when speaking phonemes. While the data driven approach is used to define viseme classes based on the results of the machine learning process. The definition of viseme classes is closely related to the language used so that each language has different definitions of viseme classes.

Research on the definition of Indonesian viseme is still rare. We have conducted research on the definition of Indonesian static and dynamic viseme in previous research [4][7]. We used a machine learning process approach to define the class structure of Indonesian static and dynamic viseme. An application developed from the results of this research is the animation of pronunciation for supporting the learning media of Indonesian pronunciation and production of 3D animation film specially on the scenes of dialogue.

Today, the need for use of Indonesian language is increasingly widespread. The number of foreigners who visit Indonesia for vacation, work, or study increases very rapidly. They need knowledge of Indonesian pronunciations in order to communicate with Indonesians. Some data related to the number of foreigners in Indonesia, namely : (1) Head of the statistics central agency (Indonesia : Badan Pusat Statistik) said that in August 2017 the number of foreign tourist arrivals to Indonesia increased 25% [8]; (2) The number of foreigners who work in Indonesia in 2016 reached 74,183 people [9]; (3) The Indonesian government through the '*Darmasiswa*' program launched the Indonesian language learning program for foreign speakers (Indonesian: Bahasa Indonesia untuk Penutur Asing, abbreviated BIPA) followed by 110 countries from the five continents of Asia, Africa, Australia, Europe and America. The problems that often occur in learning Indonesian by foreigners are pronunciation, accents, grammar, and vocabulary [10], for example is a problem of pronunciation [11]. Indonesian learning objectives will not be achieved, if this problem is not addressed immediately.

The pronunciation of the word in Indonesian of each vowel is not necessarily the same in every word. For example, different pronunciation of speech 'E'. Incorrect pronunciation will cause different meaning. In the word '*teras*' which has the meaning of 'home page' and ‘official’. Example of this problem is a small part of the error in the pronunciation found. There are many other words that are wrong in pronunciation. This condition is so serious that it requires a medium that can be used to assist in learning Indonesian pronunciation.

Visual speech synthesis is the process of combining sound and viseme to form speech visualizations that can visualize the phonemic pronunciation of the mouth accompanied by a speech-like voice. Speech visualization is based on the phoneme sequence of text used as input. In general, the synthesis of visual speech consists of several processes, namely the text to phoneme conversion, the text to speech conversion, the synchronization process between speech, viseme models and phoneme.

This research aims to build a realistic visual speech synthesis for Indonesian using a combination of morphing viseme and syllable concatenation approach. Application generated from this research can be used to support the learning media of Indonesian pronunciation especially foreigners.

2 Related Works

Research from Arifin et. al. on Indonesian language visual speech synthesis using the Syllable Concatenation method [12]. Development of Indonesian Text-to-Audiovisual system based on syllable-based voice database, which used viseme arrangement in synchronization process using the syllable concatenation method. This method is used to arrange the visemes with reference to a particular syllable form in Indonesian. The system generated by this method shows that the visualization of animation the phoneme pronunciation becomes smoother than using a phoneme-based method.

The morphing viseme method is also used in the synthesis of visual speech for Indonesian [13]. The method is used to create the transition animation from one viseme to another. Therefore, the visualization of the phoneme pronunciation animation resulting becomes smoother.

Research on dynamic viseme models are used to perform the visual speech synthesis for English [14]. Visual speech articulator motion is used to produce dynamic visual speech gestures. Dynamic viseme is applied to animate the pronunciation by assembling viseme units. The results showed that the display of dynamic pronunciation animation is more accurate and natural.

In [12] and [13] have resulted in smoother but less realistic pronunciation visualizations like pronunciation visualization by humans. While in [14], the resulting of pronunciation animation have the realistic and more accurate, but this only applies to English viseme. Therefore, we perform the visual speech synthesis for Indonesian by combining two methods of syllable concatenation and morphing viseme so that it can produce the visualization of dynamic pronunciation animation.

3 Indonesian Language

Indonesian is the Malay language which is used as the official language of the Republic of Indonesia and the language of Indonesian unity [15]. Indonesian language was inaugurated the use after the Indonesian Independence Proclamation, exactly the following day, simultaneously with the coming into force of the constitution. In 1972, the President of the Republic of Indonesia inaugurated the use of the enhanced of Indonesian Spellings (Indonesian: *Ejaan Yang Disempurnakan* / abbreviated: EYD) and the Decree of the Minister of Education and Culture concerning the EYD General Guidelines and the establishment of official terms applicable throughout the territory of Indonesia.

Since 1945, Indonesian has functioned as a national language. International scope, Indonesian is one of the important languages in the world. This is evidenced by one of the fact that some websites on the Internet provide translation for the Indonesian language as one of their features or services, such as Google, Wordpress, and Facebook.

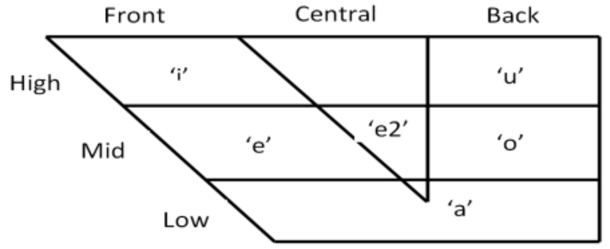


Fig. 1. Articulation Pattern of Indonesian Vowels, Image Taken From [16]

Indonesian consists of written language and spoken language. A phoneme is a term in spoken language, meaning the smallest unit of language that can distinguish meaning [17]. In general, phonemes are divided into vowels (abbreviated V) and consonants (abbreviated C). The consonant is the sound of the language produced by moving the air out with obstacles (the air out is obstructed by movement or articulator position change). Meanwhile, the vowel is the sound of the language produced by moving the air out without obstacles. Figure 1 shows articulation patterns of Indonesian vowels. Meanwhile, Table 1 shows the Indonesian consonant's articulation pattern. The Indonesian phonemes consist of 33 phonemic symbols, namely 10 vowels (included diphthongs :'au', 'ai', 'oi'), 22 consonants and 1 silent phoneme [17] as shown in Tabel 2.

In the Great Indonesian Dictionary (Indonesian: *Kamus Besar Bahasa Indonesia*), syllables are structures consisting of one or phoneme sequences that are part of the word. Each syllable is marked with a vowel and several consonants. A diphthong can be categorized as a vowel. In general, Indonesian recognizes several syllabic patterns, namely V, VC, CV, CVC, CCV, CCVC, VCC, CVCC, CCVCC, CCCV, CCCVC [18].

Table 1. Articulation Pattern of Indonesian Consonants

	Bilabial	**Labiodental**	**Dental**	**Palatal**	**Velar**	**Glottal**
Plosives	‘p’, ‘b’		‘t’, ‘d’		‘k’, ‘g’	
Affricates				‘c’, ‘j’		
Fricatives		‘f’	‘s’, ‘z’	‘sy’	‘kh’	‘h’
Nasal	‘m’		‘n’	‘ny’	‘ng’	
Trill			‘r’			
Lateral			‘l’			
Semivowel	‘w’			‘y’		

Table 2. Indonesian Phoneme Set

Consonants	‘b’, ‘p’, ‘m’, ‘f’, ‘d’, ‘t’, ‘n’, ‘l’, ‘g’, ‘k’, ‘h’, ‘j’, ‘z’, ‘c’, ‘s’, ‘r’, ‘w’, ‘y’, ‘v’, ‘sy’, ‘ng’, ‘kh’, ‘ny’
Vowels	‘a’, ‘e’, ‘E’, ‘i’, ‘o’, ‘u’, ‘au’, ‘ai’, ‘oi’
Neutral	Silent

4 Proposed Method

In general, several of the proposed research steps consist of data acquisition, the formation of Indonesian viseme classes structure and viseme models, the establishment of syllable-based speech databases, text to syllables conversion, synchronization process between speech, syllable and viseme models, and testing the result of visual speech synthesis through a survey by respondents. Overall steps represented in Figure 2. Several of the proposed research steps consist of data acquisition. In this research, the discussion focused on the synchronization process, especially stringing the viseme by combining two methods of morphing viseme and syllable concatenation.

4.1 Data Acquisition

The first step, we created a video that records the scene of people who were saying Indonesian sentences. The number of spoken Indonesian sentences is 1029 sentences. The sentences are used in the recording process to cover all phonemes and syllables in the Indonesian language (phonetically balanced sentence corpus). The sentences are derived from various literatures such as novels, books, newspapers and the Internet.

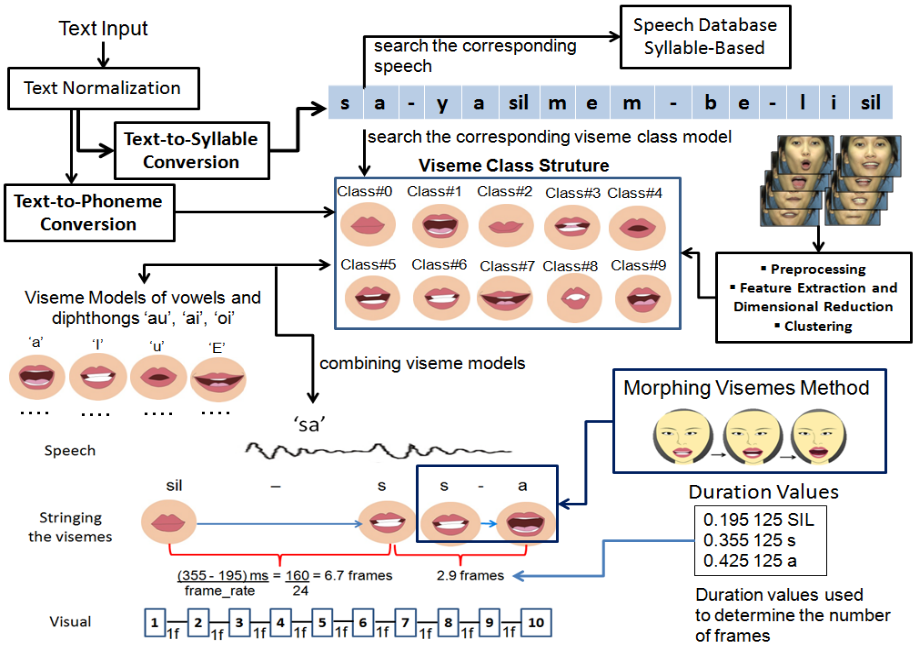

Text Input
Text Normalization
Text-to-Syllable Conversion
Text-to-Phoneme Conversion
search the corresponding speech
Speech Database Syllable-Based
s a - y a sil m e m - b e - l i sil
search the corresponding viseme class model
Viseme Class Struture
Class#0 Class#1 Class#2 Class#3 Class#4
Class#5 Class#6 Class#7 Class#8 Class#9
▪ Preprocessing
▪ Feature Extraction and Dimensional Reduction
▪ Clustering
Viseme Models of vowels and diphthongs 'au', 'ai', 'oi'
'a' 'I' 'u' 'E'
combining viseme models
'sa'
Morphing Visemes Method
Speech
sil – s
s - a
Stringing the visemes
Duration Values
0.195 125 SIL
0.355 125 s
0.425 125 a
(355 - 195) ms / frame_rate = 160 / 24 = 6.7 frames
2.9 frames
Duration values used to determine the number of frames
Visual
1 1f 2 1f 3 1f 4 1f 5 1f 6 1f 7 1f 8 1f 9 1f 10

Fig. 2. System Overview

4.2 Development of Indonesian Viseme Models

The development of Indonesian viseme models is based on the results of the clustering process on the 2D visual speech database. We have created a video that contains the mouth movements of person speaking the sentences in Indonesian. The video is extracted into frames so that it gets about 10,000 2D image frames. Next, we select the unique frames that represent a particular viseme by taking into account the context of the phonemic sound. Therefore, it is worth noting the sequence of phonemes before and after the pronunciation certain phoneme as seen in Table 3. Some preprocessing steps are changing the image color format, cropping the image in the mouth area and resizing the entire image to have the same size.

The feature extraction and dimensional reduction of the visual speech image dataset are performed using the LDA Subspace method which is a combination of PCA and LDA methods [19][20]. The use of PCA method aims to reduce the dimensions by performing a linear transformation from a high-dimensional space to a low dimensional [20]. The LDA method is used to maximize the dissemination of data to different classes is represented by the S_B (between-class scatter) matrix and minimize the spread of data in the same classes represented by the S_W (within class scatter) matrix. The results of the dimensional reduction process by the Subspace LDA method is the LDA projection matrix.

TABLE 3. Example of choosing a unique frame

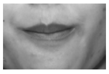
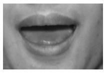
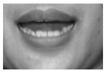
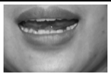
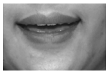
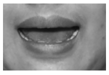
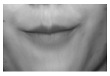
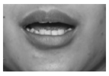
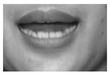
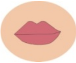
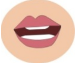
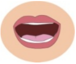
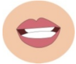
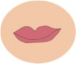
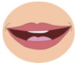
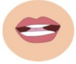
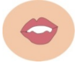
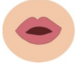
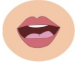
Words Examples	**Phoneme**		
	Before	***Spoken***	***After***
Bantu	 'b'	 'a'	 'n'
olEh	 'l'	 'E'	 'h'
Penulis	 'p'	 'e'	 'n'

Table 4. The Indonesian Viseme Class Structure

Viseme Classes	**Associated Phoneme**	**Viseme**	**Viseme Model**	**Viseme Classes**	**Associated Phoneme**	**Viseme**	**Viseme Model**
Class#0	Silent	-		Class#5	'k', 'g', 'kh'	'k'	
Class#1	'a', 'h'	'a'		Class#6	'c', 'j', 's', 'i', 'z', 'sy', 'ny'	'c'	
Class#2	'p', 'b', 'm'	'b'		Class#7	'E', 'y', 'oi', 'ai'	'E'	
Class#3	'd', 't', 'n', 'l', 'r'	'd'		Class#8	'f', 'v'	'f'	
Class#4	'o', 'au', 'u', 'w'	'u'		Class#9	'ng', 'e'	'ng'	

We cluster the LDA projection matrix to obtain viseme classes. The clustering process uses the K-Means method and measurement of similarities between data using a Euclidean Distance calculation. In the clustering process, different k values were tested repeatedly to obtain the k values that could produce the most optimal clusters [21]. The optimal cluster is measured by clustering quality by calculating the SSE (sum of squared error) and the ratio value of BCV (between-class variation) and WCV (within class variation) [22]. The smaller the SSE value indicates the better the cluster quality. Meanwhile, the greater the ratio value of BCV and WCV, the better the cluster quality. The most optimal cluster quality results are used as the basis for establishing the Indonesian viseme class structure. This result refers to previous studies [7] as seen in Table 4.

4.3 Indonesian Text Processing

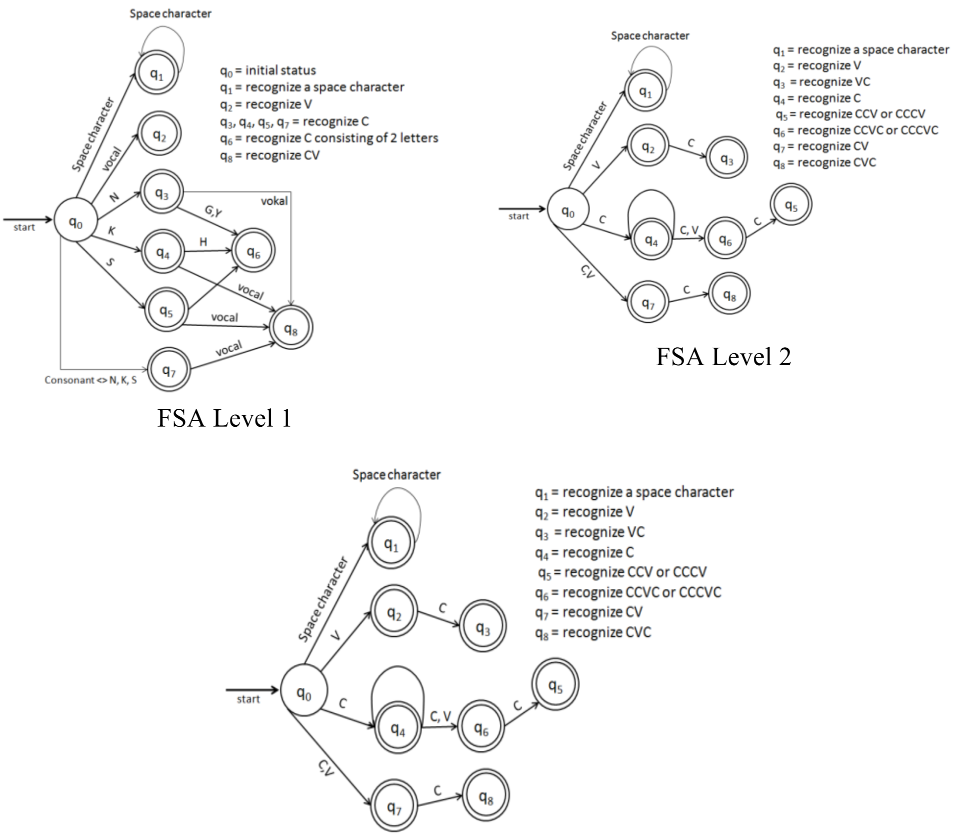
The initial processing step is required to convert text so that it can be used in the process of converting text to phonemes. The initial processing step consists of case folding and normalization. The case folding process changes all capital letters into lower case. Meanwhile, the normalization process is to remove special characters such as commas, exclamation points and others, and change the numbers into a series of letters [6].

Text to Syllables Conversion. The conversion process from text to syllables is done by applying the following rules. Each syllable is generally built by consonant, vowels and semi-consonant sounds. Each syllable must at least consist of a vowel sound or a combination of vowels and consonants. The sound of a vowel in a syllable is the peak of filtering or sonority, whereas the sound of the consonant acts as a syllable valley. In a syllable, there is only a syllable peak and this peak is indicated by a vowel sound. The syllable valley marked with consonant sounds can be more than one in number. The number of syllables in a word can be calculated by looking at the number of vowels present in the word.

If there is a word containing three vowel sounds, it can be determined that the word consists of three syllables. For example, the word "tElEr" is a word consisting of two tribes, namely "tE" and "lEr". Each syllable contains a vowel sound, the sound of "E". In the decomposition of the words of the tribes there are several things that must be considered, among others: 1) If a consonant is flanked by two vowels, then the consonant follows the vowel behind it; 2) The prefix and suffix should be written apart from the basic word; 3) If two consonants are enclosed in two vowels, then the two vowels must be separated.

Patterns of syllables are taken by formulating every syllable in the word. The vowel sound and the consonant sound will be the pattern formulation of each syllable. In the Indonesian language are found words that each syllable can only be a vowel sound, a vowel with a consonant sound, and a vowel with two consonant sounds and others. Based on this provision used a method that can convert Indonesian text into certain syllable patterns, namely Finite State Automata (FSA) [24][25]. This method is one of the machines in a simple class language.

The FSA method used in this research consists of 3 levels. First-level FSAs are used to recognize syllable patterns V, C and VC. Second level FSAs can recognize syllables with V, VC, CV, CVC, CCV, CCVC, CCCV and CCCVC patterns. The third level FSA algorithm can recognize syllables with VCC, CVCC, and CCVCC patterns. FSA algorithm is presented in Figure 3.



FSA Level 1

FSA Level 2

FSA Level 3

Fig. 3. FSA Algorithm to Determine Syllables

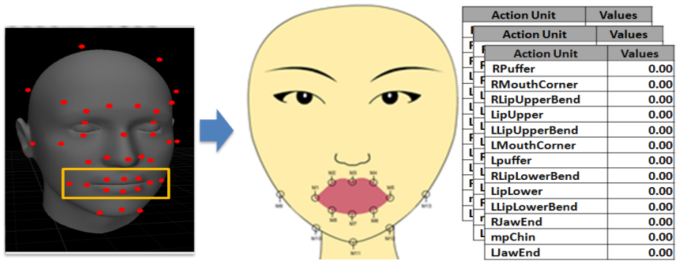
Text to Phonemes Conversion. Convert text to phoneme is a process whereby a text input in the form of text or sentence is broken down into pieces of words and phonemes. This is done to make it easier to classify phonemes into the viseme class. In the process of breaking the text as an example, if a sentence "sampai jumpa" is split into "sampai" "jumpa", while the phoneme becomes "/ s // a // m // p // a // i // // J // u // m // p // a / ".

There are several conditions in which the word input there are diphthongs in it. So if there is a diphthong in a word such as ‘au’, ‘oi’, ‘ai’, ‘kh’, ‘sy’, ‘ny’, ‘ng’ it should be given the condition to not be read as monophthong. For example, the word "sampai" when pronounced with monophthong phoneme becomes / s // a // m // p // a / / i /, then to avoid it is made a condition in the conversion of diphthong phonemes as in the viseme class so that the phoneme / S // a // m // p // ai / becomes "cabbE".

In general, the process of converting text into phonemes is started with text entered by keyboard, then split into text based on spaces to convert text into words. Next, check the characters, if the characters contain special symbols, then deleted. Checking is done again if there is a diphthong, if 'ya' then two characters will be merged, for example, the reading of character ‘n’ followed by ‘g’ will be changed to ‘ng’. And so on until all characters are read.

4.4 Face Animation Modelling

The animation model used in the developing visual speech synthesis in this research is a 2D animation with a focus on mouth motion. Therefore, in the area of the mouth mounted markers with formation refers to the FACS (Facial Action Coding System) [26] as seen in Figure 4. FACS is a codification developed by Eckman and Friesen in which each facial expression triggers the activity of certain facial muscles that, if traced. With reference to FACS, the number and position of facial feature points that serve as controls can be obtained. This control can control the movement on the face so that the animated face characters become more alive. To produce different facial expressions is done by combining Action Unit (AU) defined by FACS.



Action Unit	Values
RPuffer	0.00
RMouthCorner	0.00
RLipUpperBend	0.00
LipUpper	0.00
LLipUpperBend	0.00
LMouthCorner	0.00
Lpuffer	0.00
RLipLowerBend	0.00
LipLower	0.00
LLipLowerBend	0.00
RJawEnd	0.00
mpChin	0.00
LJawEnd	0.00

Fig. 4. Face Animation Model Based On FACS

In this research, we used 13 AUs selected according to the mouth anatomy referring to FACS. Each model of the viseme class has different AU values. Each AU of each visual grade model is used as a basis for linking one viseme to the next viseme as seen in Figure 5. This method causes the motion of the mouth from one viseme to the next to occur continuously on a regular basis resulting in smoother motion.

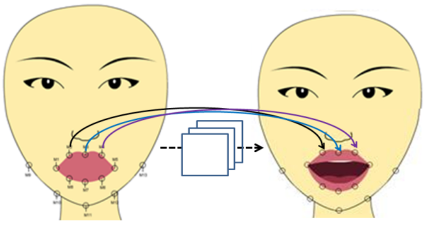


Fig. 5. AU Relates The One Viseme with The Other Viseme

4.5 Development of Speech Database

One of the important stages in the development of this synthesis is to build a voice database, both phoneme-based voice databases, and syllable-based voice databases. The sounds in this research were taken or recorded using Bear Mountain Audio (BMA) microphones connected to the microphone port on the laptop. Voting is assisted with audacity software version 2.0.6. After that, do the sound settings with right click on the speaker icon, recording device, choose an active microphone with a microphone level 100 and microphone boost + 30 dB for receiving sound and sound issued more clearly. Once the microphone settings then complete start recording sound with audacity software.

During the recording process, we select people who are able to pronounce the Indonesian sentences with several criteria, namely using standard Indonesian accent and pronouncing with normal intonation and duration. We have tested 5 people and then chose 2 people that is a man and a woman who meet the above criteria. The recording of the 2 people shows similarity both in accent, intonation and duration. This means that the recording result of the 2 persons may be considered representative of the standard Indonesian pronunciation as a whole.

4.6 Synchronization Process

The synchronization process is a process to synchronize between voice, phoneme and viseme model so as to produce a series of phoneme pronunciation visualization based on the input of Indonesian text. The synchronization process works by connecting each phoneme and sound so that it will be interconnected and when executed will work in sequence. The duration value of each phoneme voice is one of the factors determining the number of frames of each phoneme in the animation process. The determination of the number of frames of each phoneme will have an effect on the precision of visualization transition of one phoneme pronunciation to another. Since the determination of the number of frames per phoneme is based on the duration of the sound of the phoneme determination, the determination of the number of frames of each phoneme will affect the correspondence between the visualization of phoneme pronunciation and the sound of the phoneme pronunciation.

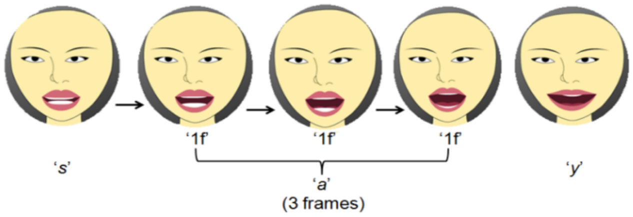


Fig. 6. Implementation of Frame Number per Phoneme of '*a*' on The Word '*s**a**ya*'

In this research, we use equation (1) to calculate the number of frames per phoneme. The duration value of each phoneme is the difference between the duration value at the end and the beginning of the pronunciation of a particular phoneme. While the duration value of each phoneme can be obtained from the speech database of each phoneme pronunciation.

$$FeP = \frac{(End_{d\_value} - Beginning_{d\_value})}{frame\_rate} \tag{1}$$

Where FeP is the number of frames per phoneme, End_{d_value} is the duration value at the end of a particular phoneme pronunciation, $Beginning_{d\_value}$ is the value of the initial duration of a particular phoneme pronunciation and $frame\_rate$ is the frame rate used in the animation development. For example, the duration values at the end and the beginning of phoneme pronunciation 'a' are 425 ms and 355 ms. The frame rate is 24 fps, the frame number is 2.9 frames (rounded to 3 frames). Implementation of the number of frames per phoneme in animation is illustrated as shown in Figure 6. The duration value of each phoneme in a syllable is combined with other phonemes resulting in a visualization of the syllable pronunciation corresponding to the pronunciation of the syllable.

4.7 Stringing Viseme using A Combination of Morphing Viseme and Syllable Concatenation Method

The Morphing visemes method is a technique used to manage visually change from one viseme to another [23]. The morphing method requires the correspondence map specification with $C_o : I_0 \Rightarrow I_1$ relating to viseme I_0 and visemeI_1. Where, C_o is the initial correspondence, C_1 is the objective correspondence, I_0 is the initial viseme and I_1 is the objective viseme as defined in equation (2) [13].

$$C_0(P_0) = \{d_x^{0\to1}(P_0), d_y^{0\to1}(P_0)\} \tag{2}$$

Pixels in viseme I_0 in position $P_0 = (x, y)$ adjust to pixel in viseme I_1 in position $\left(x + d_x^{0\to1}(x,y), y + d_y^{0\to1}(x,y)\right)$. The result of the morphing viseme process from the one viseme to the next viseme as shown in Figure 7. In the morphing viseme process, transition animation of mouth movement occurs gradually so that the transition animation between viseme look smoother.

The Syllable Concatenation approach is a method used to assemble viseme based on certain syllable patterns. With this approach, the static viseme models of the results of previous research are used as a reference to the basic viseme form that is strung together to form certain syllabic patterns based on the inputted Indonesian text. The static viseme models that make up a particular syllable pattern are based on the concept of articulation and co-articulation visualization. With this concept, visualization of articulation will be greatly influenced by the accompanying co-articulation visualization. Thus the visualization of certain phonemic articulations may vary depending on the co-articulation that accompanies the phoneme's articulation. Implementing the marker marking on each of the viseme models as described above can better realize this concept.

The use of the Syllable Concatenation approach is intended to make the visualization of a phoneme pronunciation more realistic than using a phoneme-based viseme sequence as shown in Figure 8. In the visualization of syllable-based pronunciation, visualization of a phoneme adjusts the form with the accompanying co-articulation. This is in contrast to the phoneme-based pronunciation visualization. Implementation of this concept can realize a dynamic visualization, based on the accompanying co-articulation visualization.

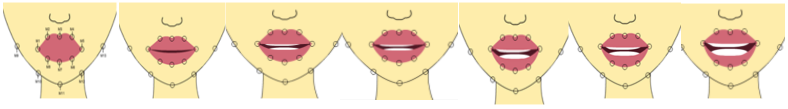


Fig. 7. The Result of Morphing Viseme Process from Viseme ‘sil’ to ‘s’

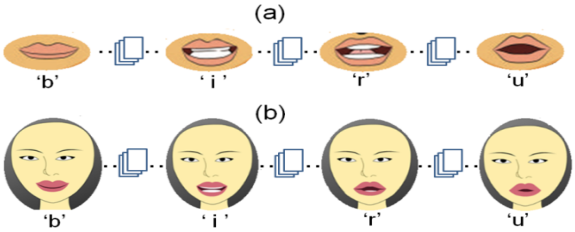
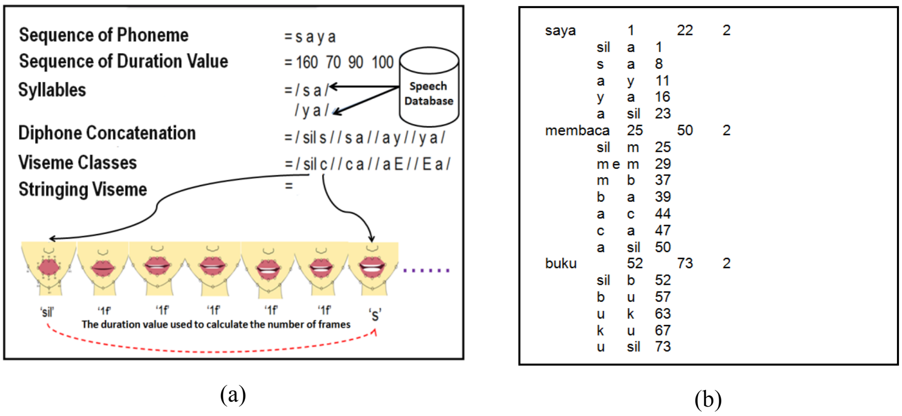


Fig. 8. Visualization of pronunciation based on : (a) phoneme, (b) syllable

4.8 Audio-visual Generation Process

The generation process is a system that can generate audio-visual (voice and viseme) simultaneously. Figure 9(a) shows the audio-visual generation process of an inputted Indonesian text. Information on the phonemic sequence and duration can be obtained from the output of the Indonesian Text-To-Speech (TTS) system which is an existing subsystem. The order of phonemes can also be obtained from the process of converting text to phoneme which is one of the models of this system. The information duration of each phoneme can also be obtained from the sound database for each phoneme. The order of phonemes is used as the basis of the generation process into audio-visual by coupling the selected sound segments of the sound database.

The phoneme sequence information obtained and the next assembled to form a particular syllable pattern. The resulting syllable pattern is used for a sequence selected sound segment from the sound database. The resulting syllable pattern is also used as a basis for assembling the visves selected from viseme models. Each syllable form is parsed into a series of diphone text. Each diphone text consists of two phonemes: the initial phoneme and the final phoneme. For example the syllable *'sa-ya'* is parsed into diphone *'sil-s'*, *'s-a'*, *'a-y'*, *'y-a'* and *'a-sil'* text. Initial phonemes and final phonemes of a diphone text are used to select appropriate viseme models as basic viseme.



(a)

```
saya          1         22      2
        sil   a    1
        s     a    8
        a     y    11
        y     a    16
        a     sil  23
membaca       25        50      2
        sil   m    25
        m e   m    29
        m     b    37
        b     a    39
        a     c    44
        c     a    47
        a     sil  50
buku          52        73      2
        sil   b    52
        b     u    57
        u     k    63
        k     u    67
        u     sil  73
```

(b)

Fig. 9. The Audio-Visual Generation Process (a), The Audio Generation Process (b)

The basic viseme form of the initial phoneme is then animated to form the basic viseme of the final phoneme of each diphone text. The process of preparing the viseme form of the initial phoneme becomes the next phoneme using the viseme morphing method. Duration information is used to calculate the number of frames required in the visual form animation process from the initial phoneme to the next phoneme. The audio-visual generation process is done simultaneously based on the Indonesian text inputted so that all input text is converted into the audio-visual appropriate. The audio generation process for text input *'saya membaca buku'* as shown in Figure 9(b).

Generally, the audio-visual generation process can be explained in the description below. We cut the Indonesian text into the words. Every word is cutting process based on syllable form. Subsequently, the process of separating the text using the diphone method for each word, for example, the word *'saya'* is separated by diphone method to *'sil-s', ' s-a ', ' a-y ', ' y-a ', ' a -sil '*. Each phoneme of each syllable is used to select the appropriate viseme model. The duration value of each phoneme is used to make the animation process of a sentence, word or word separation with diphone method. The number of frames required to animate the pronunciation of a sentence can be determined, for example, in the frame to how many words begin to be pronounced and visualize, how many frames are required to animate a particular phoneme pronunciation and so on.

5 Experimental Results

We conducted experiments by including 10 Indonesian texts as seen in Table 5. Next, the visual speech system will visualize the pronunciation animation of each text. The assessment of the visual speech system was done by 30 respondents. The respondents can interact directly with the system and can directly provide an assessment. Respondents are foreign students studying at the Dian Nuswantoro

University and people who understand about the phonology of Indonesian language, including lecturers from the Department of Language and Literature at the Dian Nuswantoro University Semarang.

Table 5. Indonesian Texts are Used In Experiment

No	Indonesian Texts
1	hari minggu kami sekeluarga akan pergi memancing
2	nelayan itu menjual ikan tangkapannya di pasar
3	pasar itu ramai sekali setiap hari libur
4	liburan ini keluargaku pergi berkemah ke gunung
5	anak kecil itu rajin belajar sehingga juara di kelas
6	aku mendapat kiriman kado ulang tahun dari ibu
7	bibi sedang mencuci baju di kamar mandi
8	seragam sekolahnya belum kering dari kemarin dijemur
9	ayah membaca koran setiap pagi sebelum berangkat kerja
10	ibu membuatkan susu buat adik setiap pagi dan malam

The quality of the resulting system is measured based on two criteria, namely the synchronization level between speech, syllable and viseme models, and the naturalness level of pronunciation visualization. Naturalness is used to measure the level of naturalness of the mouth shape during pronunciation from one phoneme to the next. When pronunciation of a phoneme, the mouth shape of the phoneme pronunciation will soon change to prepare for the next phoneme pronunciation. It results a realistic pronunciation visualization like the visualization of pronunciation by humans. Each respondent provides a visualization assessment of the pronunciation for each text for the conformity level criteria with a scale value of 1 to 5, where scale 1 means 'very out of sync', scale 2 means 'out of sync', scale 3 means 'quite sync', scale 4 means 'sync' and scale 5 means 'very sync'. Meanwhile, the assessment for the naturalness criterion with a scale of 1 to 5, where scale 1 means 'the motion of the mouth is very unrepresentative and the movement of the mouth is very rough', scale 2 means 'unrepresentative mouth motion and rough mouth movement changes', Scale 3 means 'relatively representative mouth movements and moderately gentle motion movements', scale 4 means 'representing mouth motion and subtle gesture motion', scale 5 means 'very representative mouth movements and very fine motion movements' change. The average respondent scores were calculated using Mean Opinion Score (MOS) formulated as equation (3).

$$MOS = \sum_{i=1}^{\square} \frac{x(i).k}{N} \qquad (3)$$

Where $x(i)$ is the sample value of i, k is the weight and N is the respondent number.

In this experiment, we compared the synchronization level between visual speech synthesis for Indonesian using the morphing viseme method [13] and visual speech synthesis for Indonesian using the combination of the syllable concatenation and morphing viseme methods as seen in Figure 10.

While the comparison result of the naturalness level of visual speech synthesis for Indonesian using the morphing viseme method [13] and visual speech synthesis for Indonesian using the combination of the syllable concatenation and morphing viseme methods as seen in Figure 11.

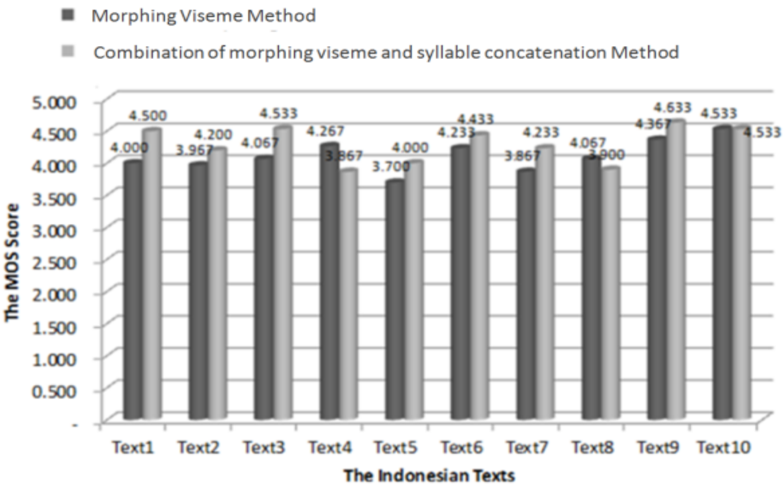


Fig. 10. Comparison of Synchronization Level Between Indonesian Visual Speech Synthesis Using Morphing Viseme Method and A Combination of Morphing Viseme and Syllable Concatenation Method

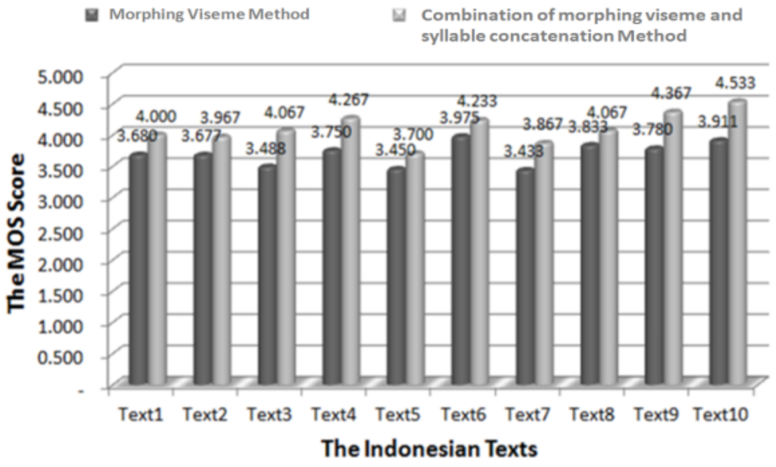


Fig. 11. Comparison of Naturalness Level Between Indonesian Visual Speech Synthesis Using Morphing Viseme Method and Combination of Morphing Viseme and Syllable Concatenation Method

6 Conclusion and Future Works

Based on some experiments and tests conducted, we can conclude that the combination of syllable concatenation and morphing viseme methods in this research can make the animation of pronunciation smoother. Therefore, the visual speech synthesis generated from this research becomes more realistic than using only the viseme morphing method as in previous research. This can be indicated by the assessment graph by the respondents for the criteria of conformity and naturalness as shown in Figure 10 and Figure 11. The MOS calculation results for the criteria of Synchronization and naturalness are 4,283 and 4,107 on the scale of 1 to 5. This result shows that the level of Synchronization and naturalness of the synthesis of visual speech is more realistic. Applications generated from this research can be used to help learn Indonesian pronunciation especially for foreigners.

In the future, the development of the synthesis visual speech for Indonesian will involve facial expressions based on the meaning of Indonesian text input. Semantic analysis of Indonesian text can produce emotional expressions such as sadness, anger, joy, disgust, surprise, and fear. The meaning of the text will also affect the intonation of the resulting sound so that the synthesis of visual speech can produce animated speech accompanied by facial expressions and voice intonation. Therefore, the synthesis of visual speech becomes more realistic and natural.

7 Acknowledgment

The author would like to appreciate Competitive Grant Research Program Directorate General of Higher Education, Ministry of Education and Culture of the Republic of Indonesia, who has helped become a funder so that this research can be completed.

8 References

[1] Wai Chee Yau, “Video Analysis of Mouth Movement Using Motion Templates for Computer-based Lip-Reading”, RMIT University, Australia, March 2008.

[2] A. Lofqvist and V.L. Gracco, “Interarticulator programming in vcv sequences: Lip and tongue movements”, Journal of the Acoustical Society of America, Volume 105(3):pp.1864-1876, 1999. https://doi.org/10.1121/1.426723

[3] Salil Deena, Shaobo Hou and Aphrodite Galata, “Visual Speech Synthesis by Modelling Co-articulation Dynamic using a Non-Parametric Switching State-Space Model”, School of Computer Science, University of Manchester, UK, 2010.

[4] Arifin, Surya Sumpeno, Muljono, Mochamad Hariadi, “A Model of Indonesian Dynamic Visemes From Facial Motion Capture Database Using A Clustering-Based Approach”, IAENG International Journal of Computer Science, pp. 41-51, Vol. 44, No. 1, 2017.

[5] I. Mazonaviciute, R. Bausys, “Translingual Visemes Mapping for Lithuanian Speech Animation”, Department of Graphical Systems, Vilnius Gediminas Technical University, ISSN 1392-1215, pp. 95-98, 2011.

[6] Luca Capellena, Naomi Harie, ”Viseme Definitions Comparison for Visual-Only Speech Recognition”, 19th European Signal Processing Conf., Barcelona, Spain, 2011.

[7] Arifin, Muljono, Surya Sumpeno, Mochamad Harriadi, *“Towards Building Indonesian Viseme: A Clustering-Based Approach”*, 2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM), pp: 57-61, 3-4 December 2013. https://doi.org/10.1109/CyberneticsCom.2013.6865781

[8] Suhariyanto, Head of Bureau of Statistics Central of Indonesia at Tempo.co site, “*Jumlah Wisatawan Asing Nasik 20% Sampai Agustus 2017*”, Publication Date : 2 October 2017”, URL: https://bisnis.tempo.co/read/1021520/jumlah-wisatawan-asing-naik-25-persen-sampai-agustus-2017, accessed on November 30, 2017.

[9] Nurhayati Ningsih, Head of Sub Directorate of Analysis and Permits of Foreign Workers of the Indonesian Ministry of Manpower at detik.com site, “*Ada 74.000 Tenaga Kerja Asing di Republik Indonesia*”, Publication Date : 17 July 2017, URL: https://finance.detik.com/berita-ekonomi-bisnis/3562880/ada-74000-tenaga-kerja-asing-di-ri-paling-banyak-dari-china, accessed on November 30, 2017.

[10] Rifca Farih Azizah, Widodo HS, Ida Lestari, “*Pengembangan BIPA Program CLS (Critical Language Scholarship)*”, Universitas Negeri Malang, pp. 1-13, 2012.

[11] Ifalani, S.Pd., “*Pembelajaran Pelafalan Bahasa Indonesia Dalam Pemerolehan Bahasa Kedua Untuk Penutur Asing Melalui* Distance Learning”, URL: http://ifalani-fhasyariflan.blogspot.co.id/2012/02/pembelajaran-pelafalan-bahasa-indonesia.html, Publication Date : 2 Februari 2012, accessed on November 30, 2017

[12] Arifin, Surya Sumpeno, Mochamad Hariadi, Arry Maulana Syarif, “Development of Indonesian Text-to-Audiovisual Synthesis System Using Syllable Concatenation Approach to Support Indonesian Learning”, International Journal: Emerging Technologies in Learning (iJET), vol. 12 no. 02, ISSN: 1863-0383, 2017.

[13] Arifin, Surya Sumpeno, Mochamad Hariadi, Hanny Haryanto,*“A Text-to-Audiovisual Synthesizer for Indonesian by Morphing Viseme”*, International Review on Computers and Software, (IRECOS), Vol. 10 No. 11. ISSN: 1828-6003, e-ISSN: 1828-6011, November 2015.

[14] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Iain Matthews, “*Dynamic Unit of Visual Speech*”, ACM SIGGRAPH Symposium on Computer Animation, 2012.

[15] *Undang-Undang Dasar (UUD) Republik Indonesia Tahun 1945, Pasal 36.*

[16] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, “Melgeneralizedcepstral analysis — A unified approach to speech spectral estimation,” Proc. ICSLP’94, pp.1043– 1046, Sep. 1994.

[17] Soenjono Dardjowidjojo, “*Bahasa Indonesia sebagai BahasaInternasional?, Menabur Benih Menuai Kasih”*, Yayasan Obor Indonesia, pp. 65-81, Jakarta, 2004.

[18] Subaryani D.H. Soedirdjo, Hasballah Zakaria, Richard Mengko, “Indonesian Text-to-Speech Syllable Concatenation for PC-based Low Vision Aid”, 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, 17-19 July 2011.

[19] Pande Made Mahendri Pramadewi et. al. “*Pengembangan Aplikasi Text to Speech Untuk Bahasa Bali*”, JANAPATI Vol. 2 No. 3, ISSN 2089-8673, Desember 2013.

[20] Aamir Khan, Hasan Farooq, “PCA-LDA Feature Extractor for Pattern Recognition”, IJCSI International Journal of Computer Science Issues, Vol 8, ISSN : 1694-0814, pp. 267-270, 2011.

[21] K.A. Abdul Nazeer, M.P. Sebastian, “Improving the Accuracy and Efficiency of k-means Clustering Algorithm”, Proceedings of the World Congress on Engineering, July 1 – 3, London, U.K., Vol I, ISBN : 978-988-17012-5-1, 2009.

[22] Daniel T. Larose, "Discovering Knowledge in Data", A John Wiley & Sons, Inc. Publication, USA, pp. 153–157, 2005.

[23] T. Ezzat, G. Geiger, and T. Poggio, "Trainable video realistic speech animation", In proceedings of the 29th ACM SIGGRAPH, pages pp.388-398, 2002.

[24] Vladimir Baltic, "Applications of The Finite State Automata For Counting Restricted Permutations And Variations", Yugoslav Journal of Operations Research, Number 2, pp. 183-198, 2012, https://doi.org/10.2298/YJOR120211023B

[25] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, R.C. Carrasco, "Probabilistic Finite-State Machines", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27, Issue 7, pp. 1013-1025, July 2005, https://doi.org/10.1109/TPAMI.2005.147

[26] Lailatul Husniah, Hardianto Wibowo, Eko Mulyanto Yuniarno, Facial Rigging Untuk Karakter 3D Berbasis Facial Action Coding System (FACS)", Journal of Animation & Game Studies, Vol. 1, No. 1, pp. 17-30, April 2015.

4 Authors

Aripin earned his bachelor degree in Information Systems from Dian Nuswantoro University, Semarang in 1997 and received an M.Kom degree in 2004 from the Informatics Engineering Department of Universitas Dian Nuswantoro Semarang, Indonesia (email: arifin@dsn.dinus.ac.id). He earned a doctor degree in Electrical Engineering from Institut Sepuluh Nopember Surabaya (ITS), Surabaya in 2017. He works as a lecturer in Informatics Engineering Department of Dian Nuswantoro University, Semarang. His research interests include natural language processing and human-computer interaction.

Hanny Haryanto earned his bachelor's degree in Informatics Engineering from Dian Nuswantoro University, Semarang, in 2007, and Magister Teknik (M.T.) degree from Institut Teknologi Sepuluh Nopember, Surabaya in 2009. His research interests include adaptive game, immersive environment, game based learning and artificial intelligence.

Surya Sumpeno is with the Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia (email: surya@ee.its.ac.id). He earned his bachelor degree in Electrical Engineering from ITS, Surabaya, Indonesia in 1996, and an M.Sc degree from the Graduate School of Information Science, Tohoku University, Japan in 2007. He earned a doctor degree in Electrical Engineering from ITS, Surabaya in 2011. His research interests include natural language processing, human computer interaction, and artificial intelligence. He is IAENG, IEEE, and ACM SIGCHI (Special Interest Group on Computer-Human Interaction) member.

Article submitted 08 December 2017. Resubmitted 20 March and 04 April 2018. Final accceptance 11 june 2018. Final version published as submitted by the authors.