

A Model of Indonesian Dynamic Visemes From Facial Motion Capture Database Using A Clustering-Based Approach

Arifin, Surya Sumpeno, *Member, IAENG*, Muljono, Mochamad Hariadi

Abstract—Realistic 3D facial animation is a challenging task in the entertainment industries. One of the efforts is to build a realistic lips animation. This research aims to build a model of Indonesian Dynamic visemes based on the results of the clustering process of the facial motion capture (MoCap) database. The Subspace LDA (Linear Discriminant Analysis) method is used to reduce the dimension. The Subspace LDA method is a combination of the PCA (Principal Component Analysis) and the LDA method. The clustering process is used to make up a natural grouping of data features which its dimensions are reduced into a number of groups. The quality of cluster results is measured by using Sum Square Error (SSE) and a ratio of Between-Class Variation (BCW) and Within-Class Variation (WCV). The measurement shows that the results of the clustering process achieving the best quality occurs at $k = 38$. In this research, it has been found out that the class structure of Indonesian dynamic visemes consists of 39 classes (38 classes from the clustering process and 1 class for neutral). For the future work, the results of this research can be used as a basis to build Indonesian visual speech synthesis smoother and as a reference to determine a structure of Indonesian dynamic visemes based on linguistic knowledge.

Index Terms—Clustering Process, Dimensional Reduction, Facial Motion Capture Database, Indonesian Dynamic Visemes

I. INTRODUCTION

MoCap is the process of recording and interpreting information about the movement and location of the subject from time to time into the digital model. The main function of the MoCap is to get motion data from specific points on the subject. There are various applications in MoCap, such as the production of animation, motion or industrial analysis, game development, filmmaking, medical and validation of computer vision [2]. Some of the applications refer to recording actions of human actors and use that information to animate digital character models in 2D or 3D computer animation. One of the products is a

realistic human lips animation.

Lips animation is closely related to visemes indicating certain phoneme articulation. A viseme is a visual representation of the phonetic speech [1], whereas dynamic visemes represent the coarticulation and prosody. Coarticulation (secondary articulation) is a symptom interplay between one sound with another sound when the primary articulation produces first sound, the speech organs make preparations to produce the next sound. For example, sound /b/ in the word 'buku' (book) with the sound /b/ in the word 'baca' (read), are pronounced differently, although the same point of articulation is bilabial. There are two types of coarticulations, which are anticipatory coarticulation, and preservative coarticulation. The first occurs when a feature or characteristic of a speech sound is anticipated (assumed) during the production of a preceding speech sound; The second occurs when the effects of a sound are seen during the production of sound(s) that follow(s).

The results of the research conducted by [3] show that the talking animation resulting from dynamic viseme more natural and reasonable as compared to the static viseme. Experimental results of the research [4] shows that the use of dynamic visemes can improve a naturalness of lips animation, as compared to the use of triphones. The forming of dynamic viseme classes depending on the language used. The number of dynamic viseme classes in each language is different. For example, English needs 150 dynamic viseme classes [5], while Chinese needs 40 dynamic viseme classes [3]. Therefore, it is important to define Indonesian dynamic visemes to sound off articulation. Up to now, there is no established Indonesian dynamic viseme standard defined.

A research conducted by [6] uses the dataset driven approach to automatic lip sync. The method is used in this research to build Indonesian dynamic visemes by configuring viseme classes based on the clustering process result to the mouth 3D coordinate data. It starts with feature extraction from each marker point in the mouth area. The next steps are the normalization of 3D position, segmentation of the 3D coordinates data and adding the mouth features into data features. The clustering process is performed based on the data features. The result of the clustering process is used as a basis of formation the dynamic viseme classes.

Manuscript received August 29th, 2016; revised November 25th, 2016. This work was supported by BPKLN (*Biro Perencanaan dan Kerjasama Luar Negeri*) - Excellence Scholarship Programme of The Ministry of Education, RI).

Arifin and Muljono are students doctoral program of Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. They work as a lecturer of Informatics Engineering Departement of Dian Nuswantoro University Semarang, Indonesia. (e-mail: arifin, muljono@dsn.dinus.ac.id).

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

II. RELATED WORKS

Research on the formation of an Indonesian static viseme class structure has been conducted by [7, 8, 9]. All of those researchers have similar goals, but the methods used are different. In [7], the formation of a static viseme class structure is based on the result of clustering process to the data set. The dataset is obtained from the result of a feature extraction process and dimensional reduction to a database of the 2D images of visual speech. An Indonesian static class viseme structure resulted from this method consists of 10 viseme classes, whereas in [8], grouping the static viseme classes is based on the linguistic knowledge. The result of the grouping is validated through a survey. An Indonesian static viseme class structure formed consisting of 12 viseme classes.

A research on the dynamic viseme models based on MoCap data has been conducted by [10]. In this research, a MoCap system measures 3D coordinates of individual markers glued on the speaker's face. Movements of markers are necessarily linked. The position of markers is affected by jaw gestures, by lip gestures like rounding and protrusion, and so on.

Another research on the dynamic viseme models is also conducted in the Chinese language by [5], which aims to a realize realistic visual speech synthesis. The research uses mouth feature parameters of Chinese static visemes, vowels, and consonants which are classified using the algorithm of the Fuzzy C-Means clustering. The data used in this research are 2D frames extracted from a video. To embody dynamic features of mouth motion, delta parameters are calculated, namely the differences between current and preceding frames. The combination of a static viseme with consonant - vocal and character of Chinese pronunciation results in 40 Chinese dynamic viseme.

Referring to several types of research above, we would like to build an Indonesian dynamic viseme class structure based on the facial MoCap database using a clustering-based approach. The database contains the data features of the mouth of a person talking in Indonesian. The MoCap

technology is used in the recording process. The results could be used as a reference for the development of an Indonesian talking system.

III. SYLLABLES IN INDONESIAN

Every word spoken in general is built by the sounds of language, either in the form of the sound of vowels, consonants, and semiconsonants. Consonants are the phonemes produced by moving air out with obstacles. In this case, one referred to the obstacles is the inhibition of the air out of the movement or change of position articulator. Vowels are the phonemes produced by moving air out without obstacles, whereas, phonemes are the smallest units of speech sounds that differentiate meaning. The Indonesian phoneme set consists of 33 phonemic symbols which comprise 9 vowels (including diphthongs), 23 consonants and 1 silent [15]. Table I shows the Indonesian phoneme set [17].

Every word consists of one or more segments. In phonology study, the segment is called a syllable. A syllable is a unit of organization for a sequence of speech sounds. Each syllable should at least be composed of a vowel or a combination of vowels and consonants. Vowels in a syllable are the peak of a filtering or sonority and the consonants act as a valley of syllables. In a syllable, there is only a peak marked with vowels. The valley of a syllable is marked by consonants the number of syllables can be more than one.

TABLE I
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

Each syllable in the Indonesian words consists of a vowel, a vowel with a consonant and a vowel with two consonants. Based on this rule, in the Indonesian, there are 11 patterns of syllables [19, 20]. A vowel is denoted by V, while a consonant denoted by C, and so the writing notation

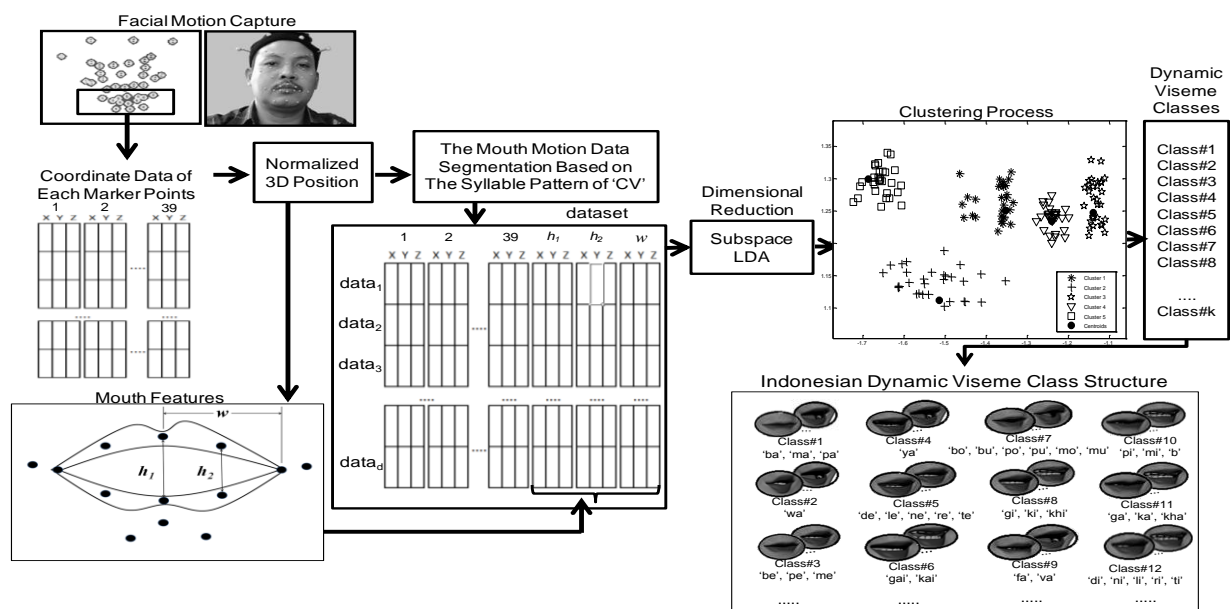


Fig. 1. An Overview of The Proposed Clustering Framework

of patterns of syllables as follows [18]:

- V as in a-nak
- VC as in da-un, ab-di
- VCC as in eks
- CV as in ba-pak, pa-di
- CVC as in kur-si, sa-yur
- CVCC as in teks-til, pers
- CCV as in sas-tra, psi-kis
- CCVC as in frik-si, kon-trak
- CCVCC as in kom-pleks, spons
- CCCV as in stra-ta, stro-bilus
- CCCVC as in struk-tur, skrip

In this research, we focus on the syllable pattern of 'CV' as a database of facial MoCap. Syllable pattern of 'CV' is one of the four types of patterns of general syllables (The original syllable patterns of Indonesian: V, VC, CV, and CVC). Furthermore, Indonesian still have several of additional syllable patterns, such as VCC, CVCC, CCV, CCVC, CCVCC, CCCV, and CCCVC [22]. The additional syllable pattern is a pattern of syllables adopted from foreign languages and regional languages.

In the syllable pattern of 'CV', the consonant is a speech sound is generated due to the articulation. Articulation is part of said tool which touched articulator (for example lower lip, tongue tip, leaf tongue, back of the tongue) to produce the sounds of language, whereas, a vowel is the sound of language produced without articulation. Therefore, the vocal is not prevalent called as articulation (named coarticulation). The mouth shape when pronouncing the articulation is strongly influenced by the accompanying coarticulation.

IV. PROPOSED METHODS

A. Overview

This research employs several steps as follows: transformation process from the marker points in the mouth area into the 3D coordinate data, normalization of 3D position for the 3D coordinate data, the dimensional reduction using subspace LDA. The Subspace LDA method is a combination of PCA and LDA method. Projecting the data to the direction having the biggest variance using PCA and projecting the data to a smaller space dimension wherein all patterns can be maximally separated and tightly grouped using LDA. The next step is a clustering process using K-Means method. The quality of clustering result is measured by using SSE and a ratio of BCV and WCV. The

clustering result is used to map into dynamic viseme classes so that the Indonesian dynamic viseme class structure can be formed. An overview of the proposed clustering framework as shown in Fig. 1.

B. Data Collection And Preprocessing

An OptiTrack Camera MoCap system is used to acquire realistic human facial motion data. The 6 cameras are used to track 33 retro-reflection markers on the face and 4 markers on the head at the rate of 60 frames/second (see Fig. 2). We recorded a person who was uttering 200 Indonesian sentences and at the same time, we used a video camera to record the subject's face. The sentences were recorded in this research has covered the whole pattern of the Indonesian syllables. The original data generated by facial MoCap are the C3D file format. From this file, there are 18.859 frames obtained.

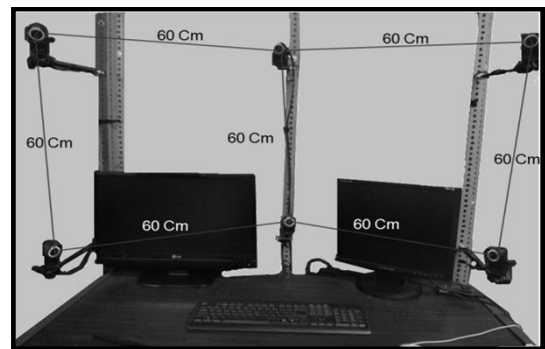


Fig. 2. The OptiTrack Camera Formation in Facial MoCap

We also recorded the speech sound data at the rate of 32 KHz and saved them in .wav format. Audacity software is used to analyze the pitch and loudness the speech data.

C. The Mouth 3D Coordinates Data

In this research, the observation of markers is focused on the mouth area related to syllable pronunciation (see Fig. 3(a)(b)). In [11], the 3D coordinates data of markers forming the feature vector were used to represent each frame in a mouth motion. Each marker in the mouth area can be obtained through the value of the translation. A Mokka (Kinematic and Kinetic Motion Analyzer) software is used to transform the file of the C3D format into the 3D coordinate data for each marker in the mouth area. From this process, 18.859 frames are obtained and each frame contains the 3D coordinate data of each marker in the mouth area.

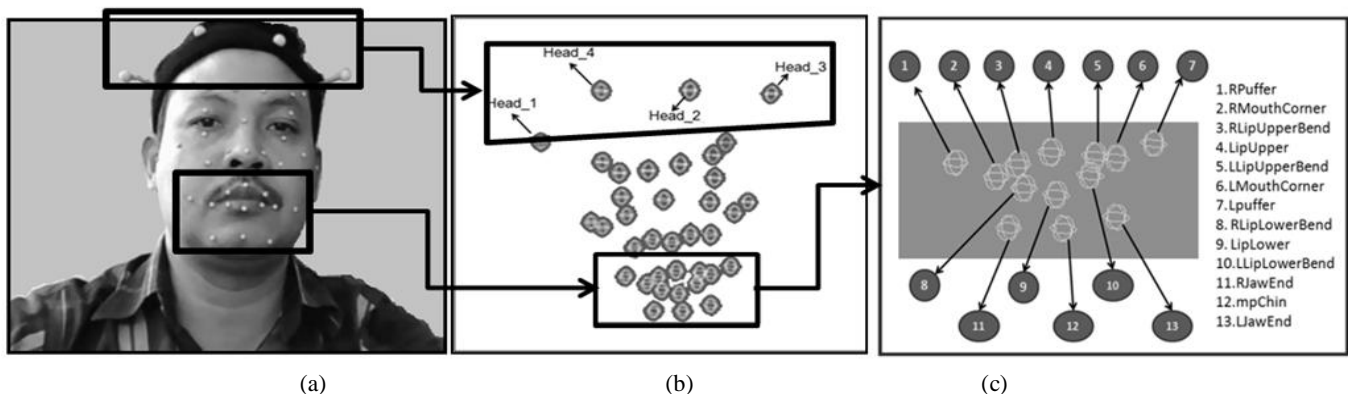


Fig. 3. Placement of Markers on an Actor's Face (a), Markers (b), Markers on Mouth Area (c)

The position of the 3D coordinates of the mouth movements will be changed according to the movement of the head. Therefore, the mouth 3D coordinate data need to be normalized.

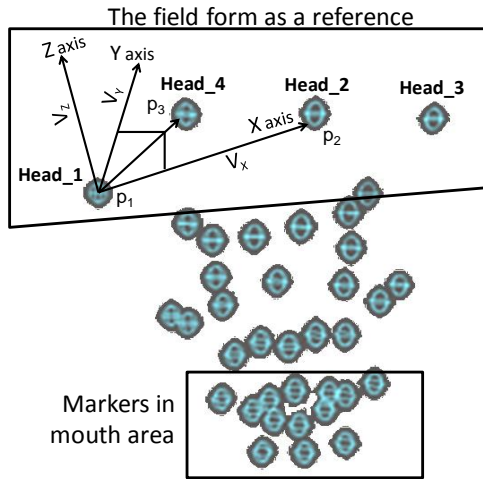


Fig. 4. Markers in Head and Mouth Area

D. Normalized 3D Position

The 3D coordinate data that has been generated are the coordinate data relative to the movement of the head. This means that coordinate data of each marker will be easily changed along with the movement of the head. Therefore, it requires a process of transformation to the local coordinate system. The process of this transformation requires a field that is used as a reference to the data coordinates of the markers. This field is composed of point markers that have relatively fixed to the movement of the head. We choose three-point markers, that is head_1, head_2, and head_4 (see Fig. 4). Each point is hereinafter referred to as p1, p2 and p3 thus forming a field as shown in Fig 4. The z-axis perpendicular to the field of $P_2P_1P_3$, Eq. (1)(2)(3) is used to calculate the coordinates of the axis of x, y and z namely V_x , V_y and V_z .

$$V_z = \frac{\vec{p_1p_2} \times \vec{p_1p_3}}{|\vec{p_1p_2} \times \vec{p_1p_3}|} = (Z_i, Z_j, Z_k) \quad (1)$$

$$V_x = \frac{\vec{p_1p_2}}{|\vec{p_1p_2}|} = (X_i, X_j, X_k) \quad (2)$$

$$V_y = V_z \times V_x = (Y_1, Y_2, Y_3) \quad (3)$$

Furthermore, the matrix M is formed as the Eq. (4).

$$M = \begin{bmatrix} X_1 & Y_1 & Z_1 & p_{11} \\ X_2 & Y_2 & Z_2 & p_{12} \\ X_3 & Y_3 & Z_3 & p_{13} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

$$Mi = \text{inv}(M) \quad (5)$$

where Mi is the matrix inverse of a matrix M . Finally, the 3D coordinates of every frame of each marker in mouth area are multiplied by the matrix Mi . The 3D coordinate data generated from this normalization process is used as the data features for the next steps.

E. The Mouth Motion Data Segmentation

The coordinate system is segmented based on the Indonesian syllable pattern. We choose to focus on the syllable pattern 'CV' (Consonant-Vowel). The frames are marked by writing numbers at the beginning and the end of each syllable pronunciation. A collection of frames that have been marked to be formed a movement trajectory. Next, the average value of the coordinates of the axis of x, y, z is calculated. The average values are used as data features of each syllable.

F. Mouth Features Extraction

The data features from the result of extraction of mouth features is added so that there will be more representatives of mouth movements. Fig. 5 describes the mouth features are used, namely the height and width of the mouth. We define the 3 features of mouth h_1 , h_2 , and w , where h_1 is a mouth height in the center of the mouth, which is calculated from the marker point differences of LipUpper and LipLower, h_2 is a mouth height in the edge of the mouth, which is calculated from the marker point differences of LipUpperBend and LipLowerBend, whereas, w is a mouth width which is calculated from a half of differences of RMouthCorner and LMouthCorner. The mouth features of each frame at the time t is : $V_t = (h_1, h_2, w)$.

To realize the dynamic features of mouth motion [11], we calculate the mouth feature on a frame at the beginning (V_{t_begin}) and at the end (V_{t_end}) of the syllable pronunciation.

Finally, the data features used in the clustering process consist of the 3d coordinate data resulting from the normalization process and the mouth features (V_{t_begin} , V_{t_end}).

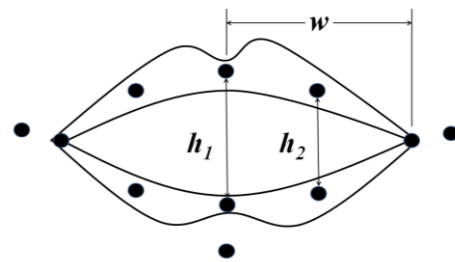


Fig. 5. Mouth Features Are Added To The Data Set, Namely h_1 , h_2 and w

G. Dimensional Reduction

The results of data features from the previous process is high-dimensional data so that the dimensional reduction is needed. There are two methods used to reduce the dimension of features, that is Principal Component Analysis (PCA) [12] and Linear Discriminant Analysis (LDA) [13]. The Subspace LDA method is a combination of PCA and LDA method which is used to reduce the optimal dimension and to find the discriminative subspace based on classified information [14]. The dimensional reduction is using PCA and LDA method simultaneously can produce accurate classes [15].

The basic principle of the dimensional reduction process using the PCA method is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while the representative features of the data are maintained [16]. Its features have an eigenvector value

greater than 0 and the rest is discarded [13]. In this research, eigenvector, and eigenvalue calculated from the covariance matrix (S_T) by using eig() function implemented in Matlab, while covariance matrix (S_T) calculated by using Eq. (6). Furthermore, the eigenvector value sorted based on eigenvalue from the largest to the smallest.

The Subspace LDA method consists of two steps, first, the dimensional reduction by PCA method performing a linear transformation set of the data features from a high-dimensional space into a lower-dimensional space, so that a classifying space is obtained [15]. Second, the LDA method is used to find linear projection, which maximizes the between-class covariance matrix and minimizes the within-class covariance matrix. Eq. (7) and (8) is used to calculate the within-class covariance matrix (S_W) and to calculate the between-class covariance matrix (S_B).

$$S_T = (A_j - \bar{A}_d)(A_j - \bar{A}_d)^T \quad (6)$$

where A_j is the j^{th} data, \bar{A}_d is the average of each row calculated by using equation $\frac{1}{k} \sum_{j=1}^k A_j$.

$$S_W = \sum_{i=1}^c \sum_{a_j \in A_d} (S_T) \quad (7)$$

$$S_B = \sum_{i=1}^c N_i (S_T) \quad (8)$$

where c is the number of class and N_i is the number of data in A_i class, while \bar{A}_i is the mean value of each class and A_j is *PCA_Projected* taken from each class.

H. Clustering Process

K-Means is an algorithm to classify a given set of data features into k number of disjoint clusters, where the value of k is fixed in advance [21]. Euclidean distance is generally used to determine the distance between data points and the centroids. The algorithm of K-Means clustering consists of several steps. The first step is determining the number of cluster k randomly. The next step is determining the value of centroids. At the beginning of an iteration, the value of the centroid is determined randomly. The next iteration, the value of the centroid is determined by calculating the average value of each cluster by using Eq. (9).

$$\bar{Y}_i = \frac{1}{N_i} \sum_{k=0}^{N_i} X_k \quad (9)$$

where \bar{Y}_{ij} is the centroid of the i^{th} cluster. N_i is the number of data in the i^{th} cluster, while X_k is the k^{th} data.

The next step is calculating the Euclidean Distance between centroids and each the data feature by using Eq. (10).

$$d_{(p_i, \bar{Y}_i)} = \sqrt{(p_1 - \bar{Y}_1)^2 + (p_2 - \bar{Y}_2)^2 \dots + (p_i - \bar{Y}_i)^2} \quad (10)$$

where $d_{(p_i, \bar{Y}_i)}$ is *Euclidean Distance*, while p_i is data points and \bar{Y}_i is centroid points.

The data clustering is performed based on the minimum Euclidean Distance. The steps are repeated until the location of centroids are fixed and membership of the cluster are fixed.

One of the methods to determine a well-defined cluster is by using the criterion function which measures clustering quality. There is a widely used method, namely the Sum of

Squared Error (SSE), which is calculated by using Eq. (11). The smaller SSE value is the better clustering quality, defined as Eq. (11).

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} d(p_i, \bar{Y}_i)^2 \quad (11)$$

where k is the number of clusters, p is the data points of a member of each cluster of C_i , $d(p_i, \bar{Y}_i)$ is the distance of each p data point to \bar{Y}_i for the i^{th} cluster.

The cluster quality can also be evaluated using between-class variation (BCV) and within-class variation (WCV). BCV is the mean of distance among centroids and WCV is the Sum of Squared Error [21]. A greater ratio value shows a better clustering quality. The ratio of BCV and WCV is formulated by using Eq. (12).

$$\frac{BCV}{WCV} = \frac{\frac{1}{n_k} \sum_{i=1}^k d(p_i, \bar{Y}_i)}{SSE} \quad (12)$$

where $\frac{1}{n_k} \sum_{i=1}^k d(p_i, \bar{Y}_i)$ is the mean of the distance between centroids.

V. EXPERIMENTAL RESULTS AND DISCUSSION

There are 18,859 frames resulted from the recording facial MoCap. Frames are segmented based on the pronunciation of syllables patterned 'CV'. From this process, there are 1,009 datasets used in this experiment. It is used to get the mouth 3D coordinates database. The mouth 3D coordinates database is normalized so that the data features representative of mouth movement is obtained. We used the Subspace LDA method to reduce the dimension of the data features. Next, the algorithm of K-Means is used to cluster the data features. Euclidean distance is used to measure the distance of the data neighborhood. In the process of clustering, given the value of different k to obtain the best quality clustering results. We use the ratio of BCV (between-class variation) and WCV (within-class variation) to determine the quality of the clustering results. A greater ratio value shows a better clustering quality.

In this research, we describe several of the experimental results. Firstly, comparison of the height and width of the mouth at the beginning and at the end of each the syllable pronunciation. The height of the mouth consists of the mouth high of the middle part (h_1) and the mouth high of the edge part (h_2).

The feature of h_1 is calculated based on the differences between the y-axis coordinates for the marker of LipUpper and LipLower. The feature of h_2 represents the differences between the y-axis coordinate for the marker points of LLipUpperBend and LLipLowerBend. While the width of the mouth of w represents the difference between the x-axis coordinate for the marker points of RMouthCorner and LMouthCorner divided by 2. Fig. 6 shows the differences of h_1 , h_2 and w at the beginning and at the end of the pronunciation of each syllable 'sa-ya-su-ka-ba-ju' (I like the shirt). The value of w becomes smaller when pronouncing the syllables that end with vowel 'u'. Whereas, the values of h_1 and h_2 becomes smaller when pronouncing the syllables that begin with the bilabial consonants ('b', 'm', 'p').

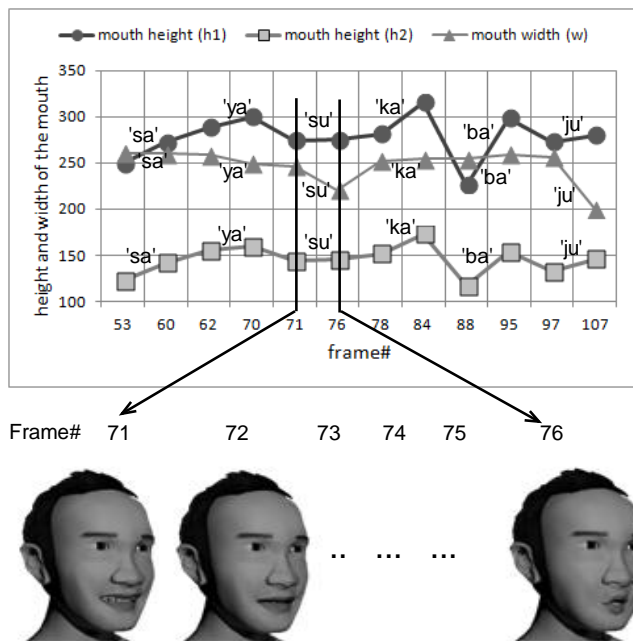


Fig. 6. Height and Width of The Mouth at The Beginning and The End of The Syllable Pronunciation "sa-ya-su-ka-ba-ju" (I like the shirt)

The next experiment is the clustering process to the data features of the syllables that begin the bilabial consonants ('b', 'p', 'm') followed by vowels ('a', 'i', 'u', 'o', 'e', 'E') or diphthongs ('ai', 'au', 'oi'). Clustering process result shows that the classes are formed based on vowels and diphthongs that follow. This means that the bilabial consonants ('b', 'p', 'm') will occupy the same class. Classes formation is strongly influenced by vowels and diphthongs that follow as shown in Fig. 7. The classes are formed from the clustering process in class#1 ('be', 'pe', 'me'), class#2 ('bo', 'po', 'mo', 'bu', 'pu', 'mu', 'bau', 'pau', 'mau'), class#3 ('bi', 'pi', 'mi'), class#4 ('bE', 'pE', 'mE', 'boi', 'poi', 'moi', 'pai', 'mai') and class#5 ('ba', 'pa', 'ma'). We determine the number of features of the projection matrix from the results of the dimensional reduction process using the LDA subspace method into the 2-features projection matrix so that it can be generated the graph as seen in Fig. 7.

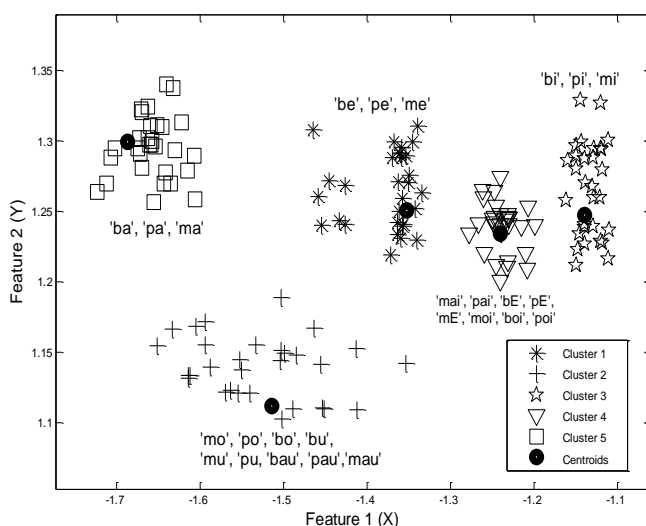


Fig. 7. The Result of Clustering Process The Features Data of Syllables Patterned 'CV' With Consonants ('p', 'b', 'm') That Followed Vowel ('a', 'i', 'u', 'o', 'e', 'E') and Diphthongs ('ai', 'oi', 'au')

In the next step, the clustering process is performed. Different k values are tested repeatedly in order to obtain the value of k with the best quality of clusters. The cluster quality is determined by calculating the ratio value of BCV and WCV. A greater ratio value shows a better clustering quality. Table II displays clustering process result from the different k value.

TABLE II
THE RESULTS OF THE RATIO VALUE CALCULATION OF BCV AND WCV

K value	Mean of Centroid Distance (BCV)	SSE (WCV)	$\frac{BCV}{WCV}$
k=20	0.9986	29.5723	0.0338
k=25	0.9873	16.5963	0.0595
k=30	1.0027	13.3606	0.0750
k=32	1.0733	6.7856	0.1582
k=34	0.9812	4.0012	0.2452
k=36	0.9754	3.4138	0.2857
k=38	1.0783	3.3400	0.3228
k=40	1.0728	3.3952	0.3160
k=42	0.9930	3.3901	0.2929
k=44	0.9930	3.4054	0.2916
k=46	0.9799	3.3738	0.2904
k=48	0.9513	3.3379	0.2850
k=50	0.9455	3.3844	0.2794
k=55	0.9027	3.3235	0.2716
k=60	1.3204	9.1866	0.1437

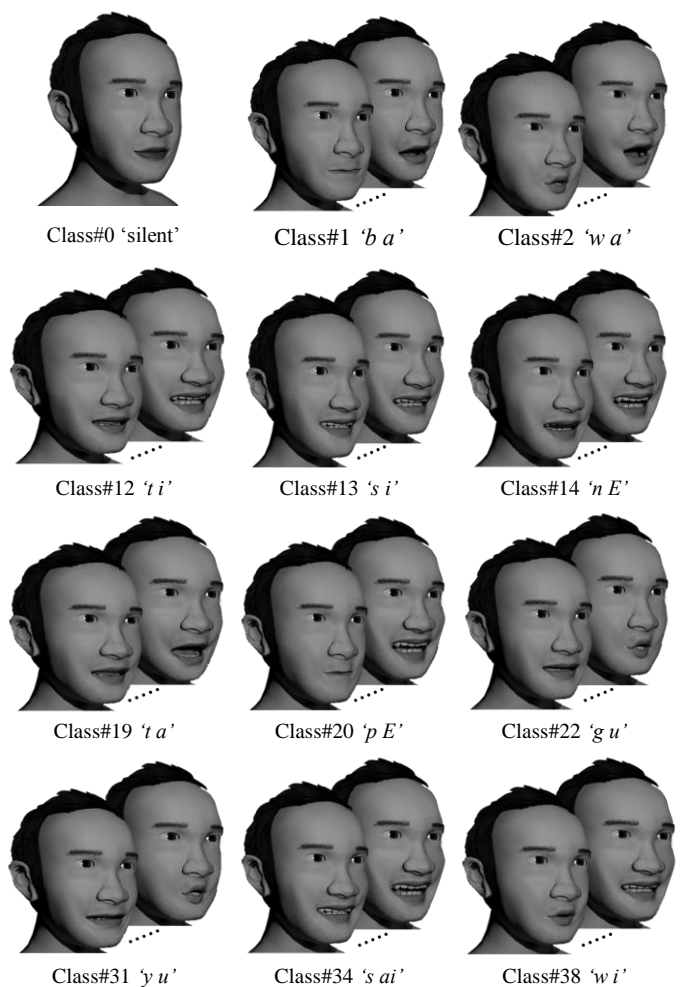
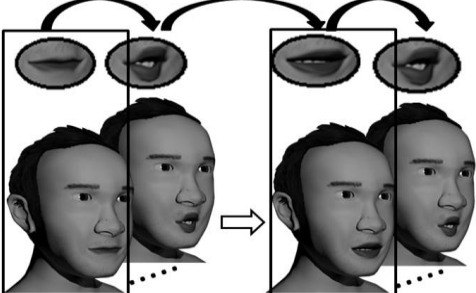
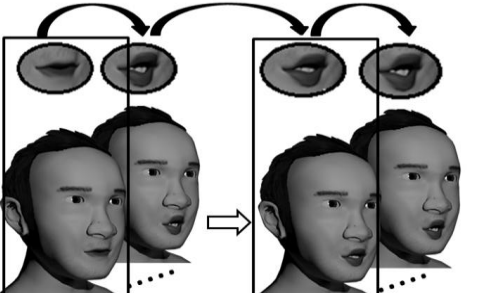
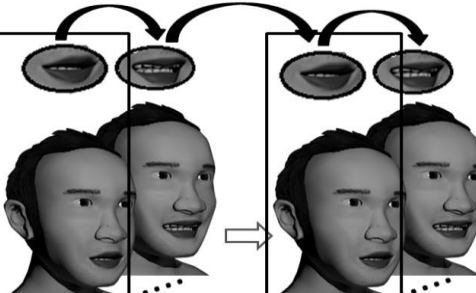
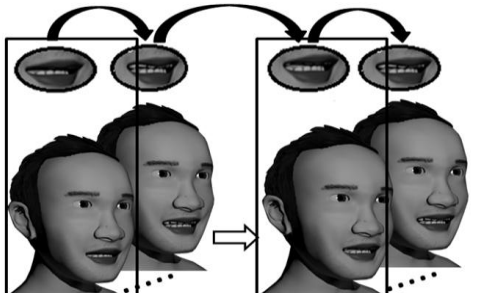
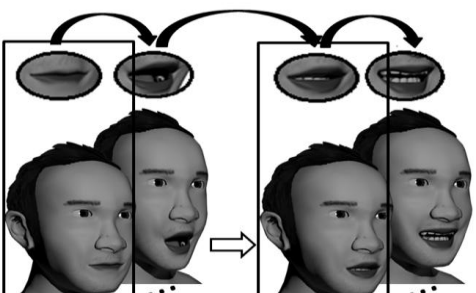
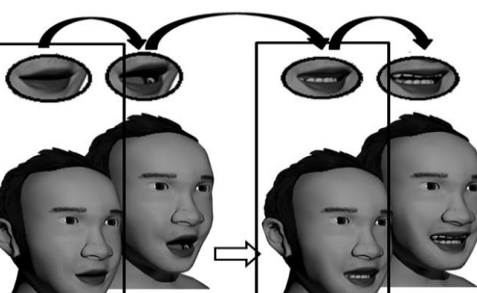
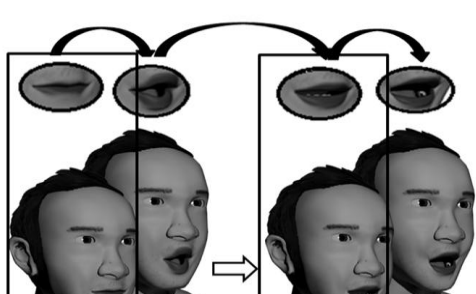
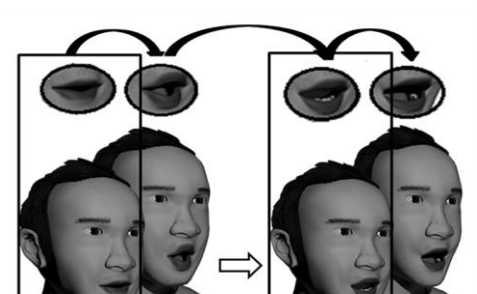


Fig. 8. Visual Representation at The Beginning g and The End of The Pronunciation of Some Dynamic Viseme Classes

TABLE III
A COMPARISON OF THE MOUTH SHAPES OF PRONUNCIATION OF A FEW OF INDONESIAN WORDS
USING STATIC AND DYNAMIC VISEMES

The Indonesian Words	Static Visemes	Dynamic Visemes
'bu-ku' (book)		
'gi-gi' (tooth)		
'ba-lai' (hall)		
'bo-la' (ball)		

In this research, data features of viseme 'silent' are not included in the clustering process. The viseme 'silent' is formed into a separate class independently such as in class the English dynamic viseme [3]. Based on the clustering

result, an Indonesian dynamic viseme class structure was formed as shown in Table IV. Visual representation of several classes of dynamic viseme illustrated by the models of dynamic viseme in Fig. 8.

The dynamic viseme classes that has been generated is evaluated through the application of the Indonesian text composed of the syllable patterns of 'CV'. This evaluation aims to check the correspondence between the mouth shape and the syllable pronunciation. Fig. 9 describes the results of visualization of the mouth shapes of the Indonesian text '*saya suka menari*' (I like to dance). The result of visualization shows that the mouth shapes each syllable look more realistic. Table III illustrates the comparison of mouth shapes of the syllable pronunciation with static and dynamic visemes. In dynamic visemes, the mouth shapes of the syllable pronunciation are strongly influenced by the phoneme that follows. It means the same phoneme will generate visualizations of different mouth shapes when followed by different phonemes.

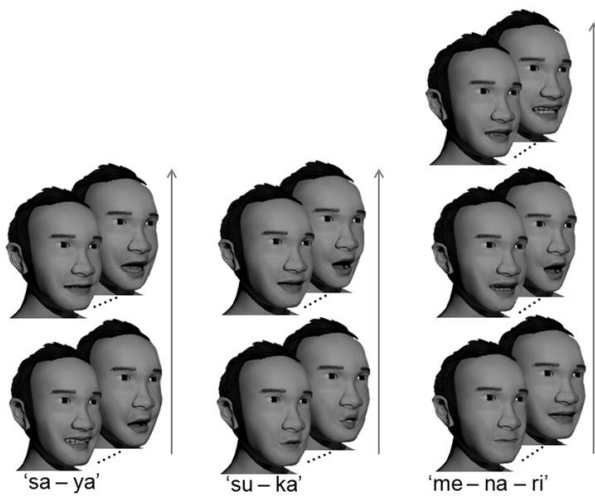


Fig. 9. Visual Representation of The Beginning and The END of The Pronunciation of The Text '*saya suka menari*' (I like to dance)

In Table III presented a comparison of mouth shapes of the syllable pronunciation of the word '*buku*' using static and dynamic visemes. This comparison shows the changes of the mouth shapes of the pronunciation of the word '*buku*' which is composed of a series of static visemes '*b-u-k-u*'. In the mapping of phoneme-viseme static rules, a phoneme is represented by a viseme. Therefore, the word '*buku*' is composed by a series of static visemes '*b-u-k-u*'. A static viseme represents a pronunciation of a certain phoneme which is constant, and it is not influenced by phonemes that precede and follow the phoneme. A static viseme for phoneme '*b*' in the word '*buku*' is the same as the static viseme for that in the word '*baca*', although the phoneme '*b*' is followed by a different coarticulation.

In the use of dynamic visemes, the mouth shape of the pronunciation of the syllable '*bu*' - '*ku*' is determined by the articulation and coarticulation phonemes. The mouth shape of the articulation phoneme is influenced by the coarticulation phoneme. Therefore, Dynamic viseme for the phoneme '*b*' in the word '*buku*' is not the same as dynamic viseme for the phoneme '*b*' in the word '*baca*'. It is caused by the phoneme '*b*' is followed by a different coarticulation.

Fig. 10 displays changes of the mouth height of pronunciation the word '*buku*'. The visualization of changes in the mouth height of the use of static viseme of pronunciation the word '*buku*' shows a drastic change. In the use of dynamic visemes, the mouth height changes are

not too drastic so that the visualization of changes of the mouth shape looks smoother. The mouth shape changes of certain phoneme pronunciation are strongly influenced by the precede and follow phonemes. Therefore, the changes in the mouth shape using dynamic visemes look smoother than static viseme. Animation of the mouth shapes presented in the examples in Table III shows that changes in the mouth shape of the word pronunciation using dynamic visemes looks smoother than static visemes. This example shows that the animation of visual speech resulted from the use of dynamic visemes looks smoother than animation from the static visemes.

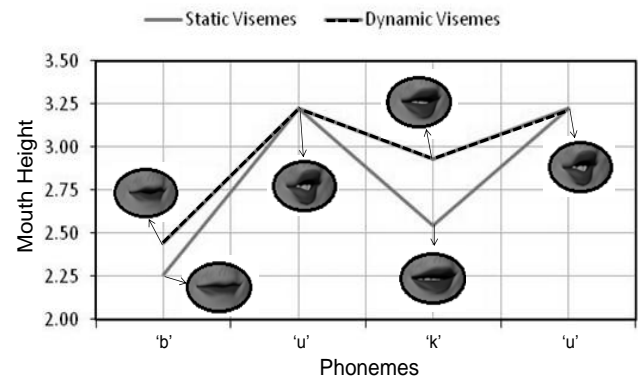


Fig. 10. A Comparison of The Mouth Height of Pronunciation of The Word '*buku*' (book) using (a) Static Visemes and (b) Dynamic Visemes.

The Indonesian dynamic viseme class structure formed in Table V, we use to build animated talking as seen in Fig. 11. It is evaluated by entering Indonesian texts as shown in Table IV. Each text assessed the level of concordance between the pronunciation of syllables and the mouth shapes generated. This test involved 30 respondents who assess and observe the results of animated talking.

TABLE IV
INDONESIAN TEXTS USED IN TESTING

No	The Indonesian Texts
1	saya suka baju baru dari ibu
2	boneka rusa di toko itu lucu sekali
3	sepatuku selesai di cuci dari tadi
4	toko itu ramai sekali dari pagi hari
5	lusa aku mulai menyanyi lagu baru
6	ibu menyirami bunga di pagi hari
7	Saya suka baca buku



Fig. 11. The Results of Animated Talking

TABLE V
 INDONESIAN DYNAMIC VISEME CLASS STRUCTURE








































Class #	Associated Syllables	Mouth Shape	Class #	Associated Syllables	Mouth Shape	Class #	Associated Syllables	Mouth Shape
0	'silent'		13	'ci', 'ji', 'nyi', 'si'		26	'nga'	
1	'ba', 'ma', 'pa'		14	'dai', 'lai', 'tai', 're', 'de', 'te', 'ne', 'rai', 'nai', 'le'		27	'fi', 'vi'	
2	'wa'		15	'ngi'		28	'fe', 've'	
3	'be', 'me', 'pe'		16	'co', 'ju', 'so', 'su', 'sau', 'jau', 'syu', 'jo', 'cu', 'cau'		29	'wo', 'wu'	
4	'ya'		17	'ce', 'je', 'nye', 'se'		30	'hi'	
5	'de', 'le', 'ne', 're', 'te'		18	'ge', 'ke'		31	'yo', 'yu'	
6	'gai', 'kai'		19	'da', 'la', 'na', 'ra', 'ta'		32	'ye'	
7	'bo', 'bu', 'mo', 'po', 'pu', 'mu'		20	'be', 'mai', 'me', 'pe', 'pai'		33	'ngu', 'ngo'	
8	'gi', 'khi', 'ki'		21	'do', 'du', 'lo', 'no', 'nu', 'to', 'lu', 'ro', 'ru', 'tu'		34	'se', 'sai', 'ce', 'je', 'jai'	
9	'fa', 'va'		22	'go', 'gu', 'ko', 'ku'		35	'we'	
10	'bi', 'pi', 'mi'		23	'ca', 'ja', 'nya', 'sa'		36	'nge'	
11	'ka', 'ga', 'kha'		24	'ha'		37	'fo', 'fu', 'vu', 'vo'	
12	'di', 'li', 'ni', 'ti', 'ri'		25	'ho', 'hu'		38	'wi'	

 TABLE VI
 MOS ASSESSMENT CRITERIA

The Value Of MOS	Quality	Description
5	Excellent	The mouth movement is exactly appropriate with the syllable pronunciation
4	Good	The mouth movement is appropriate with the syllable pronunciation
3	Adequate	The mouth movement is adequate appropriate with the syllable pronunciation
2	Bad	The mouth movement is less appropriate with the syllable pronunciation
1	Very Bad	The mouth movement is not appropriate with the syllable pronunciation

The method used to measure the level of correspondence between the syllable pronunciation and the mouth shape is MOS (Mean Opinion Score). This method can be calculated

using Eq. (13). The result of the calculation is the average assessment of respondents to animated talking. Respondents are students and teachers who are knowledgeable about the linguistic, especially pronunciation. The respondents provide an assessment in accordance with 5 criteria the level of correspondence as shown in Table VI. Recapitulation of assessment results by respondents can be seen in Table VII.

$$MOS = \sum_{i=1}^n \frac{x(i).k}{N} \quad (13)$$

where $x(i)$ is the sample value i^{th} , k is the number of weight and N is the number of respondents.

Based on the recapitulation of assessment results by respondents, The result of MOS calculation is 4.329 in the range of values of 1 to 5. It indicates that the level of

concordance between the syllable pronunciation and the mouth shape in animated talking is good.

TABLE VII
RECAPITULATION RATING BY RESPONDENTS

The Indonesian Texts	The number of votes for each level				
	Very Bad	Bad	Adequate	Good	Excellent
saya suka baju baru dari ibu	0	0	3	7	20
boneka rusa di toko itu lucu sekali	0	0	1	12	17
sepatuku selesai di cuci dari tadi	0	0	4	6	20
toko itu ramai sekali dari pagi hari	0	1	5	10	12
lusa aku mulai menyanyi lagu baru	0	2	6	8	14
ibu menyirami bunga di pagi hari	0	2	7	10	11
Saya suka baca buku	0	0	2	7	21

VI. CONCLUSION AND FUTURE WORK

Based on several results of these experiments, it can be concluded that the best quality of clusters is obtained at $k = 38$. It is based on a metric calculation Sum of Squared Error (SSE) and a ratio of BCV and WCV as one of the indicators to determine the quality of the clustering. The result of the clustering process is mapped into the classes of dynamic viseme as the basis for the formation of the structure the Indonesian dynamic viseme classes. The structure of Indonesian dynamic viseme classes is shown in Table V. This structure consists of 39 classes (38 classes from the clustering process result and 1 class of 'silent' is added). The data features used in the clustering process to cover all types of the Indonesian syllables performing a syllable pattern of 'CV'.

The structure of the Indonesian dynamic viseme class of this research is formed through the clustering process of data features performing the syllable pattern of 'CV'. Therefore, it can be discussed for another syllable pattern on a regular basis in the future. The results of this research can also be used as a reference for the research of establishing the structures of Indonesian dynamic viseme based on linguistic knowledge.

In the future, the dataset should be added to the data features of the result of audio data extraction process. The dataset used in the process of cluster consists of data features of audio and visual. The result of the clustering process in the next research should consider data features of visual and audio, in order to achieve more realistic viseme models.

ACKNOWLEDGMENT

The author would like to thank the Scholarship Program of the Bureau of Planning and foreign cooperation of the Ministry of National Education Republic of Indonesia whom has given the author's scholarship to study in the Doctoral Program of Electrical Engineering of the *Institut Teknologi Sepuluh Nopember (ITS) Surabaya*. The author was conducting the research on motion capture technology in the laboratory of HCS (Human Centric System) in the Electrical Engineering Department from *Institut Teknologi Sepuluh*

Nopember (ITS) Surabaya. The author also would like to thank 'Program Penelitian Hibah Bersaing Direktorat Penelitian dan Pengabdian Kepada Masyarakat Direktorat Jenderal Pendidikan Tinggi Kementerian Pendidikan dan Kebudayaan Republik Indonesia' then this paper to be completed.

REFERENCES

- [1] 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.
- [2] David Noonan, Peter Mountney, Daniel Elson, Ara Darzi and Guang-Zhong Yang, "A Stereoscopic Fibroscope for Camera Motion and 3D Depth Recovery During Minimally Invasive Surgery", pp. 4463-4468, In proc ICRA 2009.
- [3] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Ianin Matthews, "Dynamic Units of Visual Speech", ACM SIGGRAPH Symposium on computer Animation, 2012.
- [4] H. Li and C.J. Tang, "Dynamic Chinese Viseme Model Based on Phones and Control Function", Communication Network Security and Confrontation Laboratory, College of Electronic Science and Engineering, National University of Defense Technology, China, 2011.
- [5] Hui Zhao and Chaojing Tang, "Visual Speech Synthesis based on Chinese Dynamic Visemes", IEEE International Conference on Information and Automation, Zhangjiajie, China, 2008.
- [6] Goranka Zoric, Igor S. Pandzic, "Automatic Lip Sync and Its Use in The New Multimedia Services for Mobile Devices", Proceedings of the 8th International Conference on Telecommunications ConTEL, 2005.
- [7] Arifin, Mulyono, Surya Sumpeno, Mochamad Hariadi, "Towards Building Indonesian Viseme : A Clustering-Based Approach", CYBERNETICSCOM 2013 IEEE International Conference on Computational Intelligence and Cybernetics, Yogyakarta, December 2013.
- [8] Endang Setyati, Surya Sumpeno, Mauridhi Hery Purnomo, Koji Mikami, Masanori Kakimoto, and Kunio Kondo, "Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation," IAENG International Journal of Computer Science, vol. 42, no.3, pp233-244, 2015.
- [9] 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 N. 11, November 2015.
- [10] Shinji Maeda, "Face models based on a guided PCA of motion-capture data: Speaker dependent variability in /s/ - /S/ contrast production", pp. 95-108, ZAS Papers in Linguistics 40, Paris, France, 2005.
- [11] Wang, J., Lee, H., "Recognition of human actions using motion capture data and support vector machine", Proc. WCSE. vol. 1, pp. 234-238, IEEE, 2009.
- [12] M. Turk and A. Pentland, "Eigenfaces for Recognition", J. Of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.
- [13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegeman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 711-720, July 1997.
- [14] X. D. Jiang, "Linear subspace learning-based dimensionality reduction," IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 16-26, Mar. 2011.
- [15] 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.
- [16] I.T. Jolliffe, Principal Component Analysis, 2nd Edition, Springer series in statistics, pp 1-6, 2002.
- [17] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, "Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)", Balai Pustaka, Jakarta, Indonesia, 2003.
- [18] Subaryani D.H. Soedirjo, Hasballah Zakaria, Richard Mengko, "Indonesian Text-to-Speech Using Syllable Concatenation for PC-based Low Vision Aid", International Conference on Electrical Engineering and Informatics, 17-19 July 2011, Bandung, Indonesia, 2011.

- [19] Basuki, Thomas Anung, “*Pengenalan Suku Kata Bahasa Indonesia Menggunakan Finite-State Automata dalam Integral*”, Vol. 5, No. 2, Oktober, 2000.
- [20] Chaer, Abdul, “*Fonologi Bahasa Indonesia*”, Jakarta : Penerbit Rineka Cipta, 2009.
- [21] K. A. Abdul Nazeer, M. P. Sebastian, “Improving The Accuracy and Efficiency of The K-means Clustering Algorithm”, Proceedings of the World Congress on Engineering Vol I WCE 2009, London, U.K., July 1-3, 2009.
- [22] Panitia Pengembangan Bahasa Indonesia, *Pedoman Umum Ejaan Bahasa Indonesia Yang Disempurnakan*, Pusat Bahasa Departemen Pendidikan Nasional, 2000.



Arifin graduated arned his bachelor’s degree in Information System from Dian Nuswantoro University, Semarang in 1997 and received M.Kom. degree in 2004 from Informatics Engineering Departement of Dian Nuswantoro University Semarang. Since 2011, he has been studying at Graduate School of Electrical Engineering ITS Surabaya Indonesia as a doctoral student.

He works as a lecturer of Informatics Engineering Departement of Dian Nuswantoro University Semarang (email : arifin@dsn.dinus.ac.id). His research interest includes natural language processing and human-computer interaction. He is a member of IAENG.



Surya Sumpeno is with the Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia. He earned his bachelor degree in Electrical Engineering from ITS, Surabaya-Indonesia in 1996, and M.Sc. degree from the Graduate School of Information Science, Tohoku University, Japan in 2007 (email : surya@ee.its.ac.id). He earned 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.



Muljono is with Informatics Engineering Department, Dian Nuswantoro University, Semarang, Jawa Tengah, Indonesia (e-mail : muljono@dsn.dinus.ac.id). He received Bachelor degree in Mathematics Department at Diponegoro University, Semarang, in 1996 and master degree in Informatics Engineering, STTIBI, Jakarta in 2001. Since 2010, he has been pursuing a Ph.D. degree at Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. His research interests

are speech processing, artificial intelligence, and natural language processing.



Mochamad Hariadi received the B.E. degree in Electrical Engineering Department of Sepuluh November Institute of Technology, Surabaya, Indonesia, Surabaya, Indonesia, in 1995. He received both M.E. and Ph. D. degrees in Graduate School of Information Science Tohoku University Japan, in 2003 and 2006 respectively. He is currently teaching at the Department of Electrical Engineering, Sepuluh November Institute of Technology, Surabaya, Indonesia. His research interests are Video and Image Processing, Data

Mining and Intelligent System. He is a member of IEEE and a member of IEICE.