

Towards Building Indonesian Viseme: A Clustering-Based Approach

Arifin¹, Muljono²

Department of Information Technology
Universitas Dian Nuswantoro
Semarang, Indonesia

email: {arifin¹, Muljono²}@research.dinus.ac.id

Surya Sumpeno³, Mochamad Hariadi⁴

Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember Surabaya
Surabaya, Indonesia

email: {surya³, mochar⁴}@ee.its.ac.id

Abstract—Lips animation plays an important role in facial animation. A realistic lips animation requires synchronization of viseme (visual phoneme) with the spoken phonemes. This research aims towards building Indonesian viseme by configuring viseme classes based on the clustering process result of visual speech images data. The research used Subspace LDA, which is a combination of Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), as the extraction feature method. The Subspace LDA method is expected to be able to produce an optimal dimension reduction. The clustering process utilized *K-Means* algorithms to split data into a number of clusters. The quality of clustering result is measured by using Sum of Squared Error (SSE) and a ratio of Between-Class Variation (BCV) and Within-Class Variation (WCV). From these measurements, we found that the best quality clustering occurs at $k=9$. The finding of this research is the Indonesian viseme consisting of 10 classes (9 classes of clustering result and one neutral class). For a future work, the result of this research can be used as a reference to the Indonesian viseme structure that is defined based on linguistic knowledge.

Keywords—viseme; clustering; subspace LDA; feature extraction; *K-Means*; Sum of Squared Error

I. INTRODUCTION

Visual speech synchronized with spoken phonemes can produce a more realistic human lips animation. Lips animation is closely related to viseme indicating certain phoneme articulation. Viseme is a visual representation of the phonetic speech [1]. There are a lot of Indonesian animation products. One particular movie that attracts our attention is 'Sing to the Dawn', which is an Indonesian animation movie produced by Infinite Frameworks (IFW). It is the first 3D animation movie aired in the cinema. Unfortunately, the lips animation in this movie is not good. Its lips animation does not look realistic because the viseme does not synchronize with the spoken phoneme. Therefore, it is important to define Indonesian viseme to sound articulation. Up to now, there has been no established Indonesian viseme standard defined.

There are two ways to build viseme. The first is by using an approach of Linguistic Viseme classes which can be defined linguistically and manually construct visually similar phonemes. The second second approach is Data Sets Driven [2]. This research uses the second approach to build Indonesian viseme by configuring viseme classes based on the clustering process result of visual speech image data. We can not use an already established viseme such as English viseme because the number of viseme classes in each language is

different. For example, English needs 15 viseme classes [3] and Persian needs 7 viseme classes [4]. The relationship of phoneme and viseme is "one to many" type. It means that one viseme can cover many different phonemes [5].

The building of the Indonesian viseme using data sets driven approach is started with feature extraction using PCA method. PCA is a statistical method to analyze data sets [6]. PCA method aims to reduce the dimension by conducting linear transformation from high-dimensional space to low-dimensional space. The disadvantage of PCA method is its less optimal class partition. The uses of PCA in this research are to extract features from the visual speech images, to reduce the dimension, and to project the data to the direction which has the biggest variance.

The purpose of LDA (Linear Discriminant Analysis) is to find an optimal projection which can project input data to a smaller space dimension wherein all patterns can be maximally separated. The use of LDA in this research is to maximize between-class scatter (S_B) and to minimize within-class scatter (S_W). Therefore, the viseme classes can be maximally separated and tightly grouped.

Projection matrix is formed from the LDA method used in clustering process. The purpose of clustering is to discover natural grouping of Indonesian viseme.

II. RELATED WORK

There are a lot of researches related to English viseme [3][5][7][8][9], unfortunately there hasn't been one related to Indonesian viseme. The following are some works on English viseme that have been done.

Werda et al [7] develop Automatic Lip Feature Extraction Prototype (AliFE) to extract the lip images data which are used in classification process for visual speech recognition.

There is also a research on the lips animation synchronized to the spoken phoneme for speech driven realistic lip animation [8]. One of the tasks in this study was English phonemes to visemes mapping.

Taylor, Mahler, Theobald, and Matthews present a new method for generating a dynamic, concatenative, unit of visual speech that can generate realistic visual speech animation [9]. Dynamic visemes are applied to speech animation by simply concatenating viseme units.

This research builds an Indonesian viseme class structure. Therefore, we need to understand phonemes, visemes and phoneme-viseme mapping to form viseme classes.

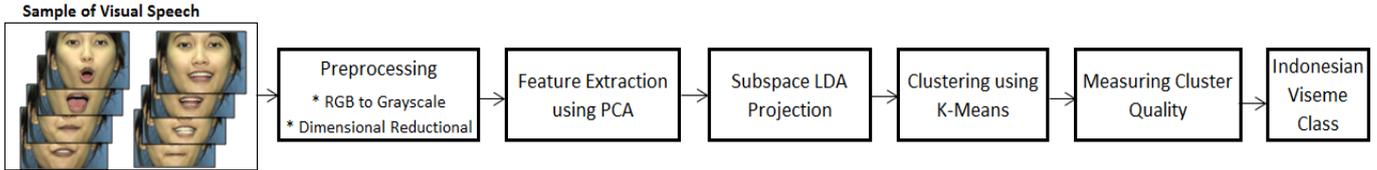


Fig. 1. System Overview

III. INDONESIAN PHONEME SET

Phoneme is the smallest unit of sound that can differentiate meaning. Indonesian phoneme consists of vowels and consonants. Vowel is a speech sound that does not meet any obstacle when it flows from the lungs. Vowel is divided into monophthong, which consists of 'a', 'i', 'u', 'e', 'o', and diphthong, which consists of 'ai', 'au', 'oi'. Consonant is a speech sound that is produced from the lungs and experiences an obstacle in its way out. The examples of consonants are 'p', 'b', 'm', 'w', 'f', 'v', 't', 'd', 'n', 'c', 'j', 'k', 'g', 'h'.

Indonesian phoneme set consists of 33 phonemic symbols which comprise 10 vowels (including diphthongs), 22 consonants and 1 silence [10]. The vowel articulation pattern of the Indonesian language indicating the first two resonances of the vocal tract is shown in Fig. 2. The Indonesian consonant sounds are distinguished by the positions of articulator and the pattern of articulation. It is displayed in Table I [11]. Table II shows the Indonesian phoneme set.

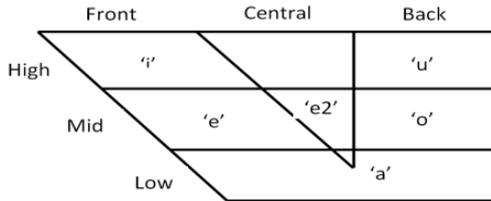


Fig. 2. Articulation Pattern of Indonesian Vowels. Image Taken From [11]

TABLE I. ARTICULATION PATTERN OF INDONESIAN CONSONANTS

	Bilabial	Labiodental	Dental	Palatal	Velar	Glotal
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 II. 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	Silence

IV. METHODOLOGY

A. Overview

This research employed several steps as follows: preprocessing to prepare the images data so that they could be processed in the next step, feature extraction and projecting

the data to the direction having the biggest variance using PCA, projecting the data to a smaller space dimension wherein all patterns could be maximally separated and tightly grouped using LDA, and followed by K-Means clustering. The quality of clustering result was measured by using SSE and a ratio of BCV and WCV. The clustering result was used for mapping into viseme classes, so that the Indonesian viseme class structure could be formed. The system overview of the proposed method is shown in Fig. 1.

B. Preprocessing

A 6:36 minute video that visualizes a person talking in Indonesian was used in this research. From this video, more than 9000 frames were extracted. Each frame was checked by using video editing software to determine which frames represent particular phonemes. Frames which were not required were manually discarded so that 1000 frames were obtained. Next, of all frames of particular phonemes resulted from the same visual speech recurring, one unique frame was selected. Therefore, 225 unique frames were obtained. In selecting a unique viseme, the sound context of a particular phoneme was considered. Therefore, it is important to notice a series of three consecutive phonemes, they are: current phoneme, and phonemes preceding and following the current phoneme [5]. This process can be seen in Table III. The selected phoneme was the current phoneme which served as a representation of unique viseme from the articulation.

The preprocessing step aims to prepare the images data so that they could be processed in the next step. There are several preprocessing steps: converting from RGB color to grayscale, cropping, and scaling of all images data so that they have the same image size.

TABLE III. SAMPLES OF VISEMES

Word Samples	Phoneme		
	before	current	after
bantu			
itu			

The next preprocessing step done was dimension reduction of image 2D into 1D. It aims to reduce the image size and the reduction result was column matrices which were combined into T matrices.

C. Feature Extraction Using PCA

Visual speech image data sets were processed into feature extraction by using the PCA. Given M sets of image data from

image database of visual speech (A_j), where $A_j = [A_{j1}, A_{j2}, \dots, A_{jM}]$, ($j = 1, 2, \dots, M$), every image was converted into 2 dimensional matrix of ($X_m \times X_n$). This matrix was then converted into an image vector T with the size of ($U \times 1$) where $U = (X_m \times X_n)$. It resulted in a set of image vectors with the size of ($U \times M$):

$$T = [T_1 T_2 \dots T_M] \quad (1)$$

Next, an arithmetic average was calculated from image vectors of ($U \times 1$) sized pixels using Eq. (2):

$$\bar{A} = \frac{1}{M} \sum_{j=1}^M T_j \quad (2)$$

where M is the number of data and $\sum_{j=1}^M T_j$ is the number of each row.

The next step was calculating A_{Train} matrix which contains difference of each value of T matrix with \bar{A} matrix, Eq. (3) is used:

$$A_{Train} = T_j - \bar{A} \quad (3)$$

After that, the value of S_T covariant matrix (total of Scatter S_T matrix) was calculated by using Eq. (4).

$$S_T = A_{Train} \times A_{Train}' \quad (4)$$

From S_T covariant matrix, eigenvalue (D) and eigenvector (V) were calculated. Eigenvalue is a characteristic value of a square matrix, while eigenvector is a value taken from eigen value greater than 0. In this research, eigenvalue (D) and eigenvector (V) were obtained by using $\text{eig}()$ function implemented in Matlab. The next step was calculating eigenfaces value which were characteristics of image data by using Eq. (5).

$$\text{Eigenfaces} = A_{Train} \times V \quad (5)$$

PCA's next task was reducing characteristics that still presented in image data. The data feature having unimportant characteristics were removed and the result of this process was 50 projection which was data feature of each object. Eq. (6) was used to calculate PCA projection matrix.

$$\text{PCA_Projected} = \text{Eigenfaces}' \times T \quad (6)$$

PCA_Projected projection matrix was the result of PCA process which is used for LDA projection next.

D. Subspace LDA Projection

Data sets of PCA_Projected obtained from PCA process were used for LDA projection. The scatter matrix in (S_w) class and between class scatter (S_B) are defined as follows:

$$S_w = \sum_{i=1}^c \sum_{a_j \in A_i} (A_j - \bar{A}_i)(A_j - \bar{A}_i)^T \quad (7)$$

$$S_B = \sum_{i=1}^c N_i (\bar{A}_i - \bar{A})(\bar{A}_i - \bar{A})^T \quad (8)$$

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.

S_B and S_w matrix was used to compute eigenvector (V) and eigenvalue (D) on LDA Projection. The data feature having

unimportant characteristics were removed. The result of this process was 10 projection, a data feature of each object.

E. Clustering Using K-Means

K-Means is an algorithm used to classify the data set into k number of clusters [12]. Euclidean distance is generally used to determine the distance between data points and the centroids. The steps of K-Means clustering algorithm are as follows:

- Determining the value of k randomly.
- Determining the value of the centroid.

In the beginning of iteration, centroid values were determined randomly. At the next iteration step, the centroid value was determined by calculating the mean of each cluster by using Eq. (9).

$$\bar{V}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (9)$$

where \bar{V}_{ij} is the centroid of the i^{th} cluster for the j^{th} variable. N_i is the number of data in the i^{th} cluster, while X_{kj} is the k^{th} data for the j^{th} variable.

- Calculating the distance of centroids and each feature data by using Eq. 10.

$$d_{(x,y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 \dots + (x_n - y_n)^2} \quad (10)$$

where $d_{(x,y)}$ is *Euclidean Distance*, while $x = (x_1, x_2, \dots, x_n)$ is data points and $y = (y_1, y_2, \dots, y_n)$ is centroid points.

- Grouping data based on the minimum Euclidean Distance.
- Going back to step b, repeating the steps until the centroid value is fixed and the cluster members do not move to another clusters.

F. Measuring Cluster Quality

One of the methods to determine a well-defined cluster is by using criteria 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 will be.

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

where k is the number of cluster, p is the data points of member of each cluster of C_i and $d(p, m_i)$ is the distance of each p data point to m_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 between centroids and WCV is the Sum of Square Error [13]. A greater ratio value shows better clustering quality. The ratio of BCV and WCV is formulated by using Eq. (12).

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

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

V. EXPERIMENTS RESULTS AND DISCUSSION

There are 225 image data used in clustering process. K-Means was used in the clustering process and Euclidean Distance was used to measure between data neighborhood. Data sets used in the clustering process were projection matrix of PCA and Subspace LDA. In this research, different k values were tested repeatedly to obtain k values which could give an optimal clustering. The optimal cluster was measured based on clustering quality.

Fig. 3 demonstrates two dimensional space of clustering result at $k = 6$ for PCA projection and Subspace LDA projection result at $k = 9$ can be seen in Fig. 4. From the figures, it can be seen that the PCA projection is scattered with great feature values compared to that of Subspace LDA projection. The result of PCA projection is not sufficient to discriminate, while a more discriminative feature can be obtained by using Subspace LDA projection which is done by calculating within class matrix (S_w) and between class matrix (S_B). Therefore, in this research the next calculation is based on Subspace LDA projection result only.

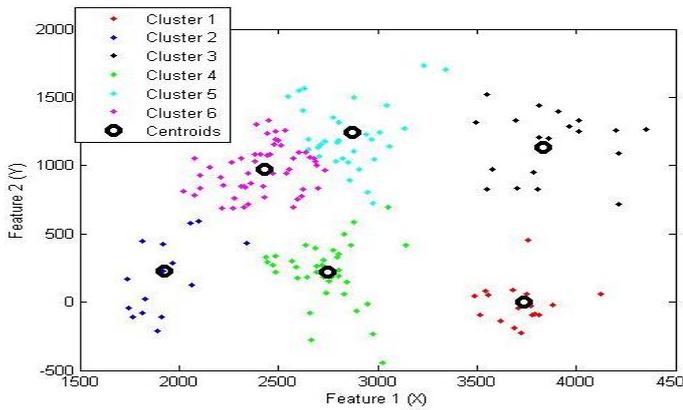


Fig. 3. Two Dimensional Space of Clustering Process for 6 Clusters of PCA Projection

The result of SSE calculation and the ratio of BCV and WCV can be seen in Table IV. The formulas used to calculate SSE and the ratio of BCV and WCV are Eq. (11) and Eq. (12). Table IV shows that the most optimal cluster result is at $k = 9$, with the smallest SSE value and the greatest ratio of BCV and WCV. Fig. 4 demonstrates scatter graphic of clustering result at $k = 9$ and description of the clustering result is shown in Table V.

TABLE IV. SSE CALCULATION AND RATIO OF BCV AND WCV

K value	Mean of Centroid Distance (BCV)	SSE (WCV)	$\frac{BCV}{WCV}$
k=5	0.88	66.58	0.0132
k=6	0.95	54.37	0.0175
k=7	0.88	50.28	0.0176
k=8	0.83	47.86	0.0174
k=9	0.79	42.38	0.0187
k=10	0.77	46.87	0.0165
k=11	0.75	43.59	0.0171
k=12	0.75	42.81	0.0175
k=13	0.70	42.73	0.0162

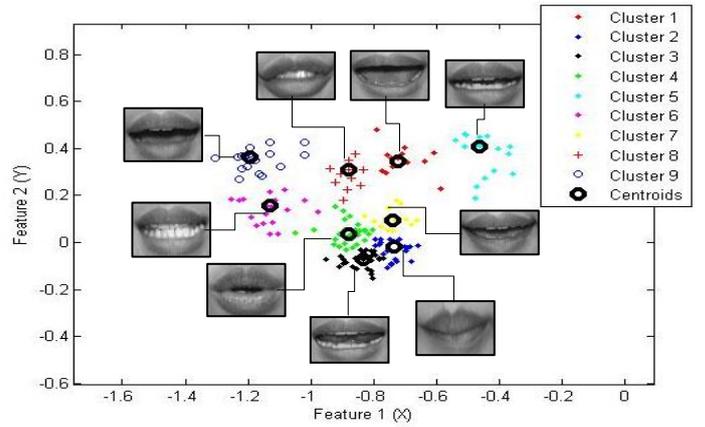


Fig. 4. Two Dimensional Space of Clustering Process for 9 Clusters of Subspace LDA Projection

TABLE V. DESCRIPTION OF CLUSTER RESULT AT $k = 9$

Cluster	Associated Visemes	(%)	Cluster	Associated Visemes	(%)
Cluster#1	'a'	52.0	Cluster#6	'c'	25.0
	'h'	48.0		'j'	21.4
Cluster#2	'p'	42.3	'y'	10.7	
	'b'	26.9	'z'	17.9	
	'm'	30.8	'ny'	10.7	
Cluster#3	'd'	28.0	'i'	14.3	
	't'	20.0	Cluster#7	'E'	29.2
	'n'	16.0		'y'	25.0
	'l'	16.0		'oi'	20.8
	'r'	16.0		'ai'	16.7
'y'	4.0	'a'		4.2	
Cluster#4	'u'	40.0	Cluster#8	'h'	4.2
	'o'	24.0		'f'	52.0
	'au'	16.0	'v'	48.0	
	'w'	20.0	Cluster#9	'ng'	36.4
Cluster#5	'k'	44.0		'e'	40.9
	'g'	24.0		'g'	9.1
	'kh'	24.0		'kh'	13.6
	'ny'	4.0			
	'y'	4.0			

Table V displays the membership percentage of each viseme which varies. This membership percentage can determine the membership level. A viseme with high membership percentage means its membership level to that cluster is very strong. A viseme with weak membership level (less than 10%) is removed from the cluster membership, but its membership to other clusters whose membership level is high is maintained or placed on the percentage of larger clusters as in the case of the phoneme 'kh'. The result of viseme membership after removal can be seen in Table VI. The membership level of each cluster is then used as a basis of mapping into viseme classes.

In this research, visemes which states 'silence' condition were not included in the clustering process. These visemes would still be formed into a separate viseme class, as in viseme classes of English [3]. Based on the clustering result, a viseme class structure was formed as shown in Table VI. The clearer visual representation of each viseme class is shown by viseme models in Fig. 5.

TABLE VI. VISEME CLASS STRUCTURE

Viseme Classes	Associated Phoneme	Viseme	Example
Class#0	Silence	-	-
Class#1	'a', 'h'	'a'	awal, hawa
Class#2	'p', 'b', 'm'	'b'	pola, baru, mandi
Class#3	'd', 't', 'n', 'l', 'r'	'd'	dari, tanda, nanas, lari, rata
Class#4	'o', 'au', 'u', 'w'	'u'	obat, gurau, lagu, warga
Class#5	'k', 'g', 'kh'	'k'	kata, galau, khayal
Class#6	'c', 'j', 's', 'i', 'z', 'sy', 'ny'	'c'	cara, jarak, saja, sepi, ijazah, syarat, nyanyi,
Class#7	'E', 'y', 'oi', 'ai'	'E'	Enak, sayang, amboi, santai
Class#8	'f', 'v'	'f'	format, motivasi
Class#9	'ng', 'e'	'ng'	yang, penulis,

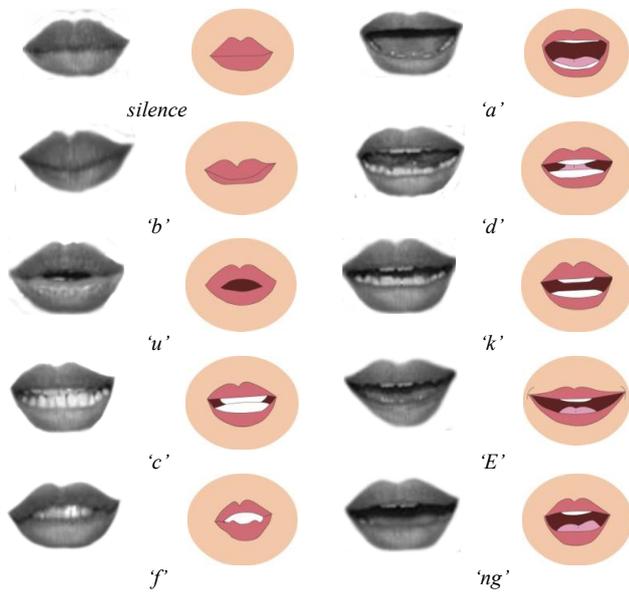


Fig. 5. Viseme Models as Results From Proposed Clustering Method

The research found that Indonesian visemes are similar to English visemes. Some English consonant and vowel phonemes have the same articulation as Indonesian phonemes [14] such as 'b', 'f', 'd', 'g', 'k', 'l', 'm', 'n', 'p', 'r', 's', 't', 'w', 'z', 'ng', 'y' as in the following examples : '*r-red*, '*m-man*, '*p-pen*, '*s-so*, '*ng-sing*, '*g-got*, etc. While the difference is that there are some phonemes pronounced with certain articulation according to the word context such as *happy*-pronounced as 'i', *to*-pronounced as 'u', *but*-pronounced as 'a', etc., and phonemes such as '*th-thin*, '*dh-then*, '*ao-saw*, '*ch-chain*.

VI. CONCLUSION AND FUTURE WORK

Several experiments have shown that the best clustering quality is obtained at $k = 9$. This is based on metric the calculation of Sum of Squared Error (SSE) and the ratio of BCV and WCV as clustering quality. The clusters resulted from this process are mapped into viseme classes as a basis of Indonesian viseme structure formation. It creates Indonesian

viseme class structure consisting of 10 classes (9 classes of clustering result and 1 neutral class). This Indonesian viseme class structure covers all of the Indonesian phonemes.

The Indonesian viseme class structure in the research is formed through a clustering process to discover natural grouping. Therefore, in the future it can be used as a reference to an Indonesian viseme class structure that is defined based on linguistic knowledge.

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] Luca Capellena, Naomi Harie, "Viseme Definitions Comparison for Visual-Only Speech Recognition", 19th European Signal Processing Conf., Barcelona, Spain, 2011.
- [3] 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 Telecommunication ConTEL, 2005.
- [4] Mohamad Aghaahmadi, Mohammad Mahdi Dehshibi, Azam Bastanfard, Mahmood Fazlali, "Clustering Persian Viseme Using Phoneme Subspace for Developing Visual Speech Application", Multimedia Tools and Applications an International Journal, ISSN 1380-7501, 2012.
- [5] M. Leszczynski, W. Skarbek, "Viseme Recognition – A Comparative Study", Faculty of Electronics and Information Technology, Warsaw University of Technology, Poland, pp. 287-292, 2005.
- [6] 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.
- [7] SalahWerda, Walid mahdi and Abdelmajid Ben Hamadou, "Lip Localization and Viseme Classification for Visual Speech Recognition", International Journal of Computing & information Sciences, Vol. 5, No. 1, pp. 62-75, 2007
- [8] Elif Bozkurt, Cigdem Eroglu Erdem, Engin Erzin, Tanju Erdem, Mehmet Ozkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation", IEEE Transaction on Audio and Speech, 2007.
- [9] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Ianin Matthews, "Dynamic Units of Visual Speech", ACM SIGGRAPH Symposium on computer Animation, 2012.
- [10] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, "Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)", Balai Pustaka, Jakarta, Indonesia, 2003.
- [11] 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.
- [12] 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.
- [13] Daniel T. Larose, "Discovering Knowledge in Data", A John Wiley & Sons, Inc. Publication, USA, pp. 153–157, 2005.
- [14] Suyanto, "An Indonesian Phonetically Balanced Sentence Set for Collecting Speech Database", Journal of Industrial Technology, Vol. XI No. 1, pp. 59- 68, 2007.