# Development of Indonesian Text-to-Audiovisual Synthesis System Using Syllable Concatenation Approach to Support Indonesian Learning

https://doi.org/10.3991/ijet.v12i02.6384

Arifin
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Dian Nuswantoro University, Semarang, Indonesia
arifin@dsn.dinus.ac.id

Surya Sumpeno
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
surya@ee.its.ac.id

Mochamad Hariadi
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
mochar@ee.its.ac.id

Arry Maulana Syarif
Dian Nuswantoro University, Semarang, Indonesia
arry_m@dsn.dinus.ac.id

**Abstract**—This study aims to develop of Indonesian Text-to-Audiovisual synthesis system using a syllable concatenation approach to support Indonesian learning. This system can visualize the syllable pronunciation synchronized with the speech signal so that it can provide a realistic illustration of the articulator movement when each phoneme is pronounced. Syllable concatenation approach is used to realize a realistic visualization by assembling articulation and coarticulation in the form of syllables. In the development of the system, we have recorded speech database in the syllables form which refers to the patterns of syllables in Indonesian. The syllable concatenation approach is used to concatenate viseme of each phoneme, and to form the visualization of syllable pronunciations. It is synchronized with the corresponding speech from the speech database. Evaluation of this system is conducted based on a "lips-reading" of the 10 Indonesian sentences entered into the system. Ratings are based on the degree of correspondence between the syllable pronunciation and the speech produced. Assessment of all respondents is calculated using MOS (Mean Opinion Score). The calculation results show that the Indonesian text-to-audiovisual system has produced the pronunciation visualization more realistic and smoother.

**Keywords**—Indonesian Text-To-Audiovisual synthesis system, Indonesian texts, Syllable concatenation approach, viseme

# 1 Introduction

In general, humans communicate through language by writing or speaking. One communication implementation is a human-to-human communication in the learning process. A communication in the learning process is a process to convey the message of the speaker to the audience with the aim that the message can be received well, affect the understanding and change the audience behavior. Therefore, the success of the learning activities depends on the effectiveness of the communication process that occurs in this learning [1]. One example of the learning process is the Indonesian learning which is often conducted in formal and non-formal educational institutions.

Today, Indonesian is increasingly in demand by foreigners. It can be seen by an increase in the number of institutions that teach Indonesian as a foreign language in some countries. Since 2005, the Government of Indonesia through the Bureau of Planning and Foreign Cooperation of Ministry of National Education organizes '*Darmasiswa*', a program containing a learning Indonesian for foreign speakers (Indonesian: *Bahasa Indonesia untuk Penutur Asing,* abbreviated BIPA) which is attended by 110 countries of the five continents of Asia, America, Australia, Europe and Africa, where participants can choose one of 45 different universities in Indonesia.

The BIPA program became popular and increasingly in demand as a result of the opening of free trade. Up to now there are still some problems in the learning process, namely the effective teaching methods and materials that should be taught [2]. In the implementation, problems encountered are divided into the linguistic and non-linguistic problems. Some non-linguistic problems are namely the learning atmosphere, adjustment, psychological conditions. Meanwhile, some linguistic problems are namely pronunciation, accent, grammar and vocabulary [3]. The Indonesian pronunciation material is one of the learning materials which are very important.

Some problems of the pronunciation of language speech occur in the use of Indonesian. The pronunciation rules of a language speech are different from other languages, such as English, French, and German. Example of problems that often occur in Indonesian pronunciation, are the word *'teknik'* which is pronounced with *'t-é-h-n-i-k'*, while it is supposed to be pronounced with *'t- é -k-n-i-k'*, and the word *'energi'* which is pronounced with *'é-n-e-r-h-i'*, *'é-n-e-r-s-i'*, or *'é-n-e-r-j-i'*, while it is supposed to be pronounced with *'é-n-e-r-g-i'*. Another example is the phoneme pronunciation of *'e'*, which has different meanings for the word *'teras' ('t- é-r-a-s')* which means a home page, and the word *'teras' ('t-e-r-a-s')* which means officials.

Examples of these problems are just a small part in the pronunciation errors found. There are still many other words which are wrong in the pronunciation, especially on the Indonesian language teaching in basic level. The basic level learners are those who simply do not know about the Indonesian language. Therefore, learning about the introduction of literacy and how the pronunciation (phonological aspects) is required, especially for learners who come from countries that have a different writing system than the alphabet in Indonesian.

Natural facial animation, life-like and realistic is a challenging study field today [4]. Some of the applications that can be developed are a character facial animation for film animation, speech therapy for the deaf and designing Human-Computer Interaction systems. A realistic talking-head is one important part of the character facial animation. In general, talking-head is developed using visemes synchronized with the speech and phonemes spoken. Several different phonemes are visualized by the same viseme, such as phonemes *'m', 'p'* and *'b'*. Therefore, these phonemes can be grouped into one class. The fact is that the visualization of a phoneme can be differentiated based on coarticulation that follows. For example, the visualization of the phoneme *'b'* is different in the word *'buku'* (book) and *'baca'* (read), since the coarticulation that follows the articulation of phoneme *'b'* is different as shown in Fig. 1. The phoneme pronunciation visualization of this type is called dynamic viseme. The use of dynamic viseme is one way to produce a realistic talking-head.

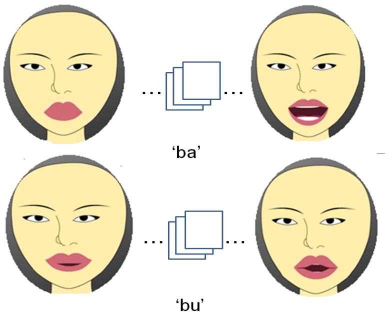

**Fig. 1.** Visualization of the mouth shape the syllables pronunciation of *'ba'* in the word '*baca*' and *'bu'* in the word *'buku'*

A text-to-audiovisual synthesis is an act of mouth movements when someone is communicating with others [5]. One part of this system is the transcription of a text into phonemes. A text consists of prosodic units, such as phrases, clauses, and sentences. The combination of the phonetic transcript and information on prosodic units can be used to create a symbolic unit representing linguistic entities, and is considered the front-end of text-to-speech. Synthesizers use this symbol to represent linguistic entities and convert them into speech.

## 2 Related Works

There are several studies on the dynamic viseme [5][6][7][8]. However, studies on the Indonesian dynamic viseme are rarely conducted. A study of visual speech synthesis method based on Chinese dynamic viseme is conducted by [5]. The dynamic viseme is constructed based on the parameters of the mouth features that are classified using a clustering algorithm. This process obtains 40 basic static visemes combined with the type of consonants and vowels. The experiment results that visual speech generated from the dynamic viseme is smoother and more realistic. An Indonesian text-to-audiovisual synthesis resulted from a study conducted by [6], is constructed based on Indonesian static viseme models. A model of static viseme classes is formed based on the results of a clustering process on a dataset of 2D images of mouth movements. Visualization transitions among units viseme are arranged using the morphing viseme method so that visualization of the phoneme pronunciation is smoother.

In the study conducted by [8], a dynamic viseme is also applied to a visual speech animation by assembling the simple viseme units. Dynamic visual speech gestures are generated from a movement of visual speech articulators. The subjective evaluation is performed to compare static viseme and dynamic viseme. The evaluation results show that dynamic viseme can generate visual speech animation more accurate.

We have conducted a study on a text-to-audiovisual synthesis based on Indonesian dynamic viseme models. In our study, a dynamic viseme model is obtained based on a combination of viseme classes of consonant phonemes and vowel phonemes that form syllables which are in accordance with the Indonesian syllable pattern.

# 3 Indonesian Language

Indonesian was declared as a national language in 1928. Indonesian is the inter-ethnic liaison language (Lingua Franca) which could unify many tribes in Indonesia. At first, the Indonesian language was written based on a Latin-Roman alphabet that followed the Dutch spelling. In 1972 Indonesian was enacted using The Enhanced Indonesian Spelling System *(*Indonesian: *Ejaan Yang Disempurnakan,* abbreviated EYD*)*. Indonesian uses units such as phrases and sentences. Sentence is the smallest unit in the verbal or written form that reveals intact thought. A sentence can consist of several elements such as subject, predicate, object, complement and description. A combination of elements of a sentence form sentences that have meaning.

## 3.1 Phonology

Phonologies are a linguistic knowledge related to sound, production of sound, instrument sounds, etc. There are two parts to phonology, namely phonetic and phonemic knowledge [9]. Phonetics is part of phonological learning of how to produce the sounds of the language or the way the language sounds are produced by the human vocal organs. Meanwhile, phonemics is part of the phonological learning of the speech sound according to its function as a different meaning [9]. Phonology is also related to terms, namely, phonemes, consonants, and vowels. The phone is a speech sound neutral or still unproven to distinguish meaning.

## 3.2 Definition of Phoneme

A phoneme is the smallest unit of sound of a language to distinguish the meaning [11]. For example, the letter *'h'* in the word *'harus'* (must) is a phoneme. If the letter *'h'* in the word is omitted, it will be *'arus'* (current). The word *'harus'* is different from *'arus'*, so the presence of the letter *'h'* can distinguish the meaning. An Indonesian phoneme consists of vowels and consonants. The vowel is a speech sound that does not meet an obstruction when expelled from the lungs. Vowels are divided into a single vowel (monophthongs) which consists of *'a', 'i', 'u', 'e', 'o'*, and a double vowel (diphthongs) which consists of *'ai', 'au', 'oi'*. Meanwhile, a consonant is a speech sound produced from the lungs expelled with obstacles.

Grapheme is the smallest unit as a differentiator in a writing system [12]. Grapheme is the epitome of the letter, grapheme referring to the letter or combination of letters as a unit of the phoneme symbol in the spelling. The relation of grapheme and phoneme is a one-to-one relationship, such as in the word *'kursi'* (chair) which consisting of graphemes *'k', 'u', 'r', 's', 'i'* and its pronunciation also consists of 5 phonemes *'k', 'u', 'r', 's', 'i'*. The other relation of grapheme and phoneme can be formed as many-to-one, such as in the word *'ladang'* (field) which consisting of graphemes *'l', 'a', 'd', 'a', 'n', 'g'*, whereas, its pronunciation consists of phonemes *'l', 'a', 'd', 'a', 'ng'*. Therefore, the graphemes *'n'* and *'g'* represented by the one phoneme *'ng'*. Table 1 shows the letters of alphabet in Indonesian and Table 2 shows the phonemes in Indonesian.

**Table 1.** Indonesian's Letters of Alphabet

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

**Table 2.** Indonesian's Phonemes

| Consonants | | | | Vowels | |
|---|---|---|---|---|---|
| 1 | 'b' | 12 | 'p' | 23 | 'a' |
| 2 | 'c' | 13 | 'r' | 24 | 'i' |
| 3 | 'd' | 14 | 's' | 25 | 'u' |
| 4 | 'f' | 15 | 't' | 26 | 'e' |
| 5 | 'g' | 16 | 'v' | 27 | 'é |
| 6 | 'h' | 17 | 'w' | 28 | 'o' |
| 7 | 'j' | 18 | 'y' | 29 | 'oi' |
| 8 | 'k' | 19 | 'sy' | 30 | 'ai' |
| 9 | 'l' | 20 | 'kh' | 31 | 'au' |
| 10 | 'm' | 21 | 'ny' | 32 | 'silent' |
| 11 | 'n' | 22 | 'ng' | | |

### 3.3 The Syllable Patterns in Indonesian

The syllable is a part of a spoken word in one breath and generally consists of several phonemes. Every syllable in Indonesian is characterized by a vowel (abbreviated V) that can be followed or preceded by a consonant (abbreviated C). The number of syllables is determined by searching the number of vowels in a word. If there is a word that contains 3 vowels, it can be determined that the word is composed of 3 syllables. For example, the word *'cepat'* (fast) is a word composed of two syllables, namely *'ce'* and *'pat'*. Each syllable contains a vowel sound, that is sounds of *'e'* and *'a'*. Every syllable must at least consist of a vowel sound or a combined vowel and consonant sounds. Based on these rules, there are some patterns of syllables in Indonesian. Type of syllable patterns Indonesian is shown in Table 3

**Table 3.** Indonesian's Syllable patterns

| The syllable patterns | Examples |
|---|---|
| V | a-kan, ma-u |
| VC | la-in, un-duh |
| CV | ra-jut, ma-ju |
| CVC | lam-bat, ram-but |
| CCV | su-pra, pu-tri |
| CCVC | blok, prak-tek |
| VCC | eks-por |
| CVCC | teks, pers |
| CCVCC | kom-pleks |
| CCCV | stra-ta, in-stru-men |
| CCCVC | in-struk-si |

## 4 Proposed Method

### 4.1 System Overview

There are several stages of the study, those are: creating a speech database, performing a clustering process of consonant phonemes to obtain consonant viseme classes, building a viseme model, synthesizing text-to-audiovisual by a process synchronizing between text, speeches, and phonemes. Overall, these steps can be seen in Fig. 2.

The dataset used in the clustering process is 2D images visual speech resulted from the extraction process of video containing scenes of the person saying 200 sentences in Indonesian. The sentences used in the recording cover the whole of syllable patterns and phonemes of Indonesian. The focus in recording the video is on the mouth movement the person pronouncing the sentences in Indonesian.

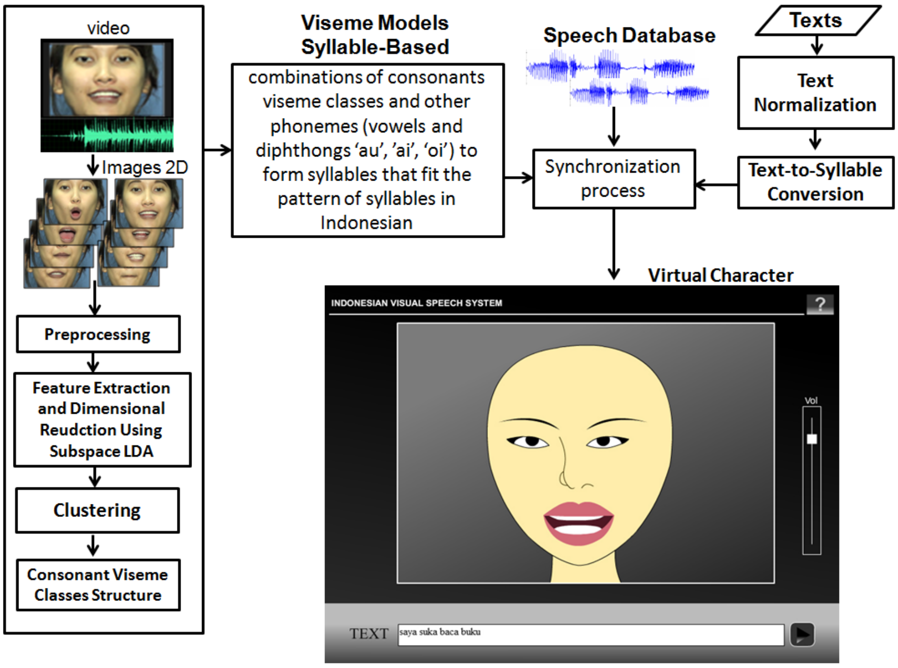
video
Images 2D
Preprocessing
Feature Extraction and Dimensional Reudction Using Subspace LDA
Clustering
Consonant Viseme Classes Structure
Viseme Models Syllable-Based
combinations of consonants viseme classes and other phonemes (vowels and diphthongs 'au', 'ai', 'oi') to form syllables that fit the pattern of syllables in Indonesian
Speech Database
Synchronization process
Texts
Text Normalization
Text-to-Syllable Conversion
Virtual Character
INDONESIAN VISUAL SPEECH SYSTEM
?
Vol
TEXT
saya suka baca buku

**Fig. 2.** System Overview of Our Proposed Method

### 4.2 Formation of Viseme Model

The formation of the Indonesian viseme model is based on the result of the clustering process of 2D images visual speech dataset. We make a video with 6 minutes and 36 seconds duration containing mouth movements of the person saying 200 sentences in Indonesian. The video is extracted to 10.000 frames of 2D images. We choose a unique frame that represents certain consonants viseme by considering before and after phonemes. This process results in 315 unique frames. Next, we change the image color format from the RGB into grayscale, and crop every image frame to focus on the mouth area with the same size for all images.

We use the Subspace LDA method for extracting features and reduce the dimension. This method is a combination of PCA and LDA method. The use of PCA method is to project the data on the direction which has the largest variety, indicated by the eigenvector corresponding to the largest eigenvalue of the covariance matrix [12]. Specifically, the task of PCA method is to reduce the dimensions by performing a linear transformation of a high-dimensional space into a low dimensional space. Whereas, LDA method aims to find a linear subspace which maximizes the matrix distance of between-class distribution ($S_B$) and minimize the matrix distance of within-class distribution ($S_W$). The results of this process are mutually separate classes linearly.

The result of the process of feature extraction and dimension reduction is a dataset used in the clustering process. If there is a set of image dataset as many *M* of the mouth shape image database ($A_j$), where $A_j = [A_1, A_2, ..., A_M]$, ($j = 1, 2, ..., M$) with the image dimension is row x column pixels projected into a two-dimensional matrix (T), then the matrix is :

$$T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix} \tag{1}$$

where x is the value of each pixel of image matrix $A_j$.

To calculate the row average of the matrix T, we used (2).

$$\bar{A}_{im} = \frac{1}{M_i} \sum_{J=1}^{M_i} X_{jm} \tag{2}$$

where $M_i$ is the number of data row $i^{th}$ and $X_{jm}$ is the average of the data row $i^{th}$ .

Next is calculating the matrix *ATrain* containing a value of the difference of the image data of T and an average value of the row:

$$ATrain = T - \bar{A} \tag{3}$$

where $\bar{A}$ is the average value of the row $\bar{A}_{im}$.

Calculating a value of covariance matrix $S_T$ (total matrix Scatter $S_T$) defined using (4).

$$S_T = ATrain\ x\ ATrain' \tag{4}$$

Eigenvalue (D) and eigenvector (V) are calculated from the covariance matrix $S_T$. The eigenvalue is a characteristic value of a square matrix, whereas eigenvector is the value taken based on of eigenvalues greater than 0. In this study, the eigenvalue *(D)* and eigenvector *(V)* are calculated by using Matlab function. Calculation of the eigenfaces value which is characteristic of image data is using (5).

$$\text{Eigenfaces} = \text{ATrain x V} \tag{5}$$

The next task of the PCA method is to reduce the features that are still contained in the image data. The data dimension that has characteristics that are not essential are removed and not used for the next process. The result of this process is the projection matrix PCA, which is calculated using (6).

$$\text{PCA\_Train} = \text{Eigenfaces' x T} \tag{6}$$

The projection matrix *PCA_Train* is used to calculate the projection matrix LDA. The scatter matrix of the within-class distribution *($S_W$)* and the scatter matrix of the between-class distribution *($S_B$)* are defined as (7) and (8).

$$S_W = \sum_{i=1}^{c} \sum_{a_j \in A_i} (A_j - \overline{A}_i)(A_j - \overline{A}_i)^T \tag{7}$$

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

where c is the number of class and $N_i$ is the number of data in class $A_i$. Whereas $\overline{A_i}$ is the average value each class and $A_j$ is *PCA_train* taken each class.

The LDA projection matrix is used as a dataset in the process of clustering using K-Means. K-Means algorithm is an algorithm for clustering n data based on specific attributes into k partitions, where k < n [11]. The initial step clustering process using the K-Means algorithm is to determine the number of clusters. Next is determining the centroid value. In the initial iterations, centroid values are determined randomly. The next iteration, centroid values are determined by calculating the average value of each cluster using (9).

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

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

Calculation of the distance between the centroid and each of the data uses a Euclidean distance method as shown in (10).

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \tag{10}$$

where $D_e$ is a Euclidean Distance and $i$ is the number of data, while *(x, y)* is data coordinates and *(s, t)* is centroid data.

Data is grouped based on the minimum Euclidean Distance. This process is repeated so that the centroid values are fixed and cluster members do not move to another cluster. Each cluster consists of data that are similar to one to another in a cluster compared with data from other cluster members.

The quality of the clustering process results measured by using Sum of Squared Error (SSE) as shown in (11). The smaller the value of SSE shows a better clustering quality [14].

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

where $k$ is the number of clusters, $p$ is data point of each cluster member $C_i$, $d(p, m_i)$ is the distance of data point $p$ to centroid $m$ to cluster $i^{th}$.

**Table 4.** The Calculation Results of SSE And Ratio of BCV And WCV

| K value | Mean of Centroid Distance (BCV) | SSE (WCV) | 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 |

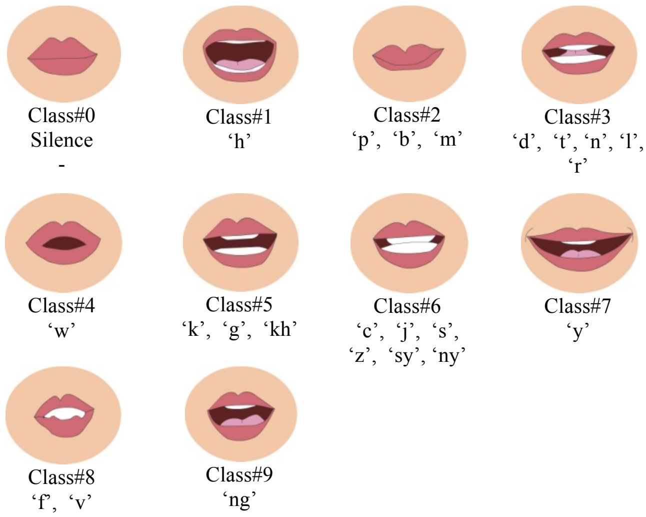

**Fig. 3.** The viseme class models to consonant phonemes

The quality of a cluster can also be seen from a comparison of between-class variation (BCV) and within-class variation (WCV). BCV is the average of the distance between the centroid, whereas WCV is equal to *Sum of Square Error* [12]. The greater the ratio value of BCV and WCV show a better clustering quality. The ratio value of BCV and WCV calculated using (12).

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

where $\frac{1}{n_k}\sum_{i=1}^{k} d(m_i, m_i)$ is the average of distance among centroid.

Table 4 shows the calculation results of SSE and the ratio of BCV and WCV from several experiments [10]. The best cluster quality obtained at k = 9, with the smallest value of the SSE and the greatest value of the ratio of BCV and WCV. The classes formed from the clustering process are used as the basis for the establishment of the viseme class models to the consonant phonemes as shown in Fig. 3.

### 4.3 Stringing visemes using Syllable Concatenation Approach

Data used in developing a system of text-to-audiovisual consist of a speech database, which cover the whole syllables in Indonesian, viseme class models for consonant phonemes and viseme models to vowel phonemes *('a', 'i', 'u', 'e', 'o', 'é')* and viseme models to diphthong phonemes *('ai', 'au', 'oi')*. The visualization of mouth movement of the syllable pronunciation can be generated by stringing visemes. In this study, we use a syllable concatenation approach to concatenate the visemes. Syllable concatenation approach is a method of developing text-to-speech or text-to-audiovisual based on syllables as the smallest unit of speech database [15]. The visual units are generated from stringing viseme of each phoneme to form the visualization of the mouth movements of every syllable. Furthermore, the unit is synchronized with the speech database to produce audiovisual systems.

The viseme class model obtained from the clustering process consonant phonemes is used as the basis for the process of stringing visemes. The formation of the viseme classes aims to reduce the amount of variation visualization mouth shape of each phoneme. Several different phonemes can be visualized by the same viseme, for example, phoneme *'b', 'p', 'm'* visualized by viseme class#2 (see Fig. 3). One class viseme is a visual representation of a phoneme or several different phonemes. Fig. 4 shows the visualization result of stringing visemes for syllable *'ma-ta'* and *'pa-da'*. Although the syllable preceded by a different phoneme ie *'m', 't', 'p',* and *'d',* the visualization can be represented by the same viseme.

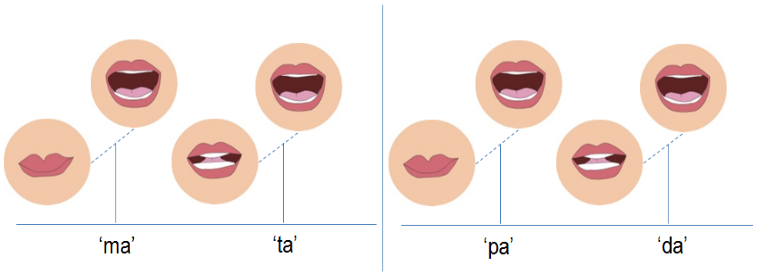

**Fig. 4.** Visualization of the mouth shape the syllables pronunciation of words *'ma-ta'* and *'pa-da'*

### 4.4 Text-to-Syllable Conversion

A text-to-syllables conversion is a separation process of text into words and splitting words into syllables. Generally, the texts have a structure that is not good so a preprocessing step is needed. The preprocessing step consists of case folding and normalization. Case folding is changing all the letters in the text document to lowercase. Whereas, text normalization aims to remove punctuation characters and change the numbers into a series of letters [16].

A word-to-syllable conversion can use simple conversion rules to implement the conversion table containing patterns of syllables in Indonesian. The next process is converting syllables into phonemes to generate phoneme codes, the value of the duration and pitch of each phoneme. The syllable-to-phonemes conversion, in general, can be seen in Fig. 5 [18].

Generally, the words in Indonesian can be converted into a phoneme with simple rules. However, there are some irregular conditions, such as the letter *'e'* can be pronounced as *'e'* or *'é'*, so it must be converted into different phonemes for different conditions. This conversation requires a conditional conversion process by taking into account a series of letters before and after that meet certain requirements so that phonemes can be obtained. This condition can be formulated as in (13).

$$Left_{context}[Letter_{set}]Right_{context} = Phoneme_{string} \tag{13}$$

Certain letters are appointed to the position $[Letter_{set}]$ can be converted into a phoneme in $Phoneme_{set}$, if $Left_{context}$ and $Right_{context}$ fulfilled.

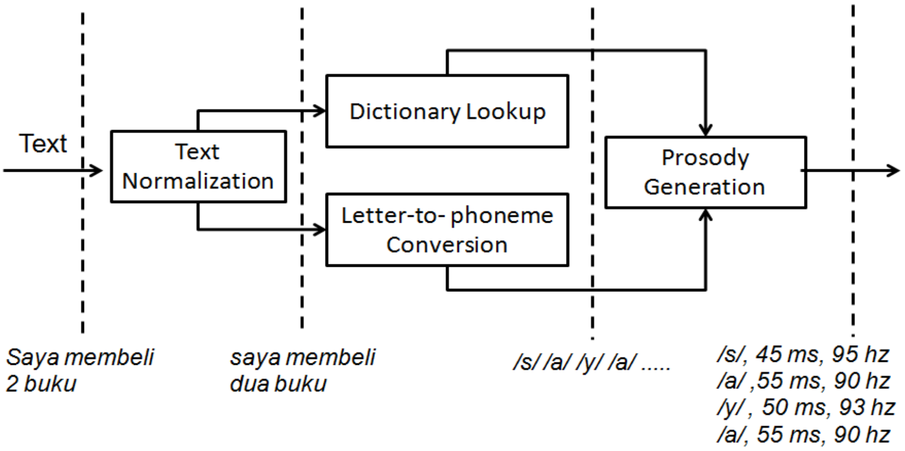

**Fig. 5.** Text-to-phonemes conversion

The value of the duration of each phoneme is obtained based on this process. It is used to determine the number of frames at the time of stringing pronunciation visualization.

### 4.5 The Synchronization Process

The synchronization process aims to align synchronization between models viseme, speech and syllables, and to create a visual flow of the series of frames that are experiencing a transition viseme [17]. One important part of the synchronization process is stringing visemes. The duration value of each phoneme is one of the factors that determine the number of frames for each phoneme in the process of stringing visemes. The number of frames every phoneme affect the results of visualization pronunciation.

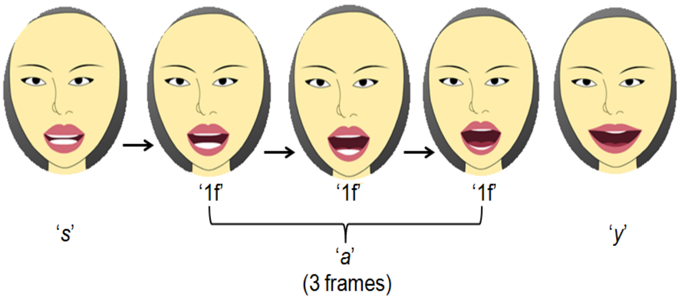

**Fig. 6.** Implementation of the number of frames on the phoneme '*a*' in the word '*s<u>a</u>ya*'

In this study, we use (14) to calculate the number of frames for each phoneme. The duration value of each phoneme is different at the end and the beginning of pronunciation of certain phonemes. The value of this duration can be obtained from the extracted files to the file features speech pronunciation of certain phonemes.

$$FeP = \frac{(End_{d\_value} - Beginning_{d\_value})}{frame\_rate} \tag{14}$$

where $FeP$ is the number of frames each phoneme, $End_{d\_value}$ is duration value at the end of pronunciation of certain phoneme, $Beginning_{d\_value}$ is duration value at the beginning of pronunciation of certain phoneme and $frame\_rate$ is the frame rate value used in the development of the animation. For example, the duration value at the end and beginning of the phoneme pronunciation of 'a' is 425 ms and 355 ms respectively. While, frame rate is 24 fps, then the number of frames is 2.9 frames (rounded to 3). Implementation of the number of frames of each phoneme in the making of animation is illustrated in Fig. 6. The duration value of each phoneme in the syllable is combined, so it can generate the visualization pronunciation syllable smoother. Fig. 7 shows the illustration of synchronization process and stringing visemes to form the pronunciation visualization of syllable based on text input.

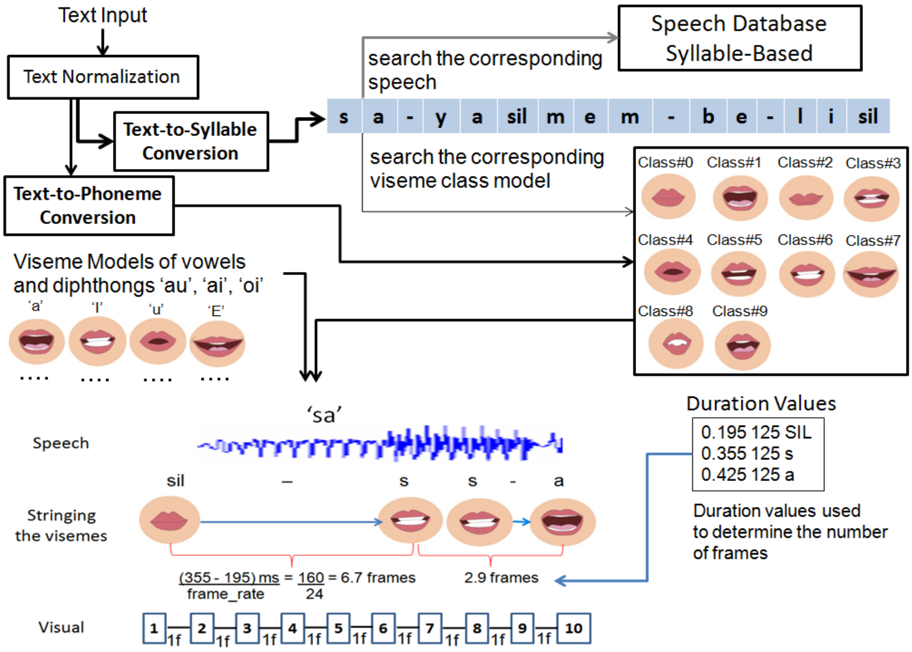
Text Input
Text Normalization
Text-to-Syllable Conversion
Text-to-Phoneme Conversion
search the corresponding speech
Speech Database Syllable-Based
s a - y a sil m e m - b e - l i sil
search the corresponding viseme class model
Class#0 Class#1 Class#2 Class#3
Class#4 Class#5 Class#6 Class#7
Class#8 Class#9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

**Fig. 7.** Illustration of synchronization process and stringing visemes to form the pronunciation visualization of syllable based on text input

## 5 Experimental Results

The system is tested by entering 10 sentences in Indonesian. We observe a correspondence between spoken syllables, visualize the mouth movement and output speech. Indonesian texts which are entered in the experiment are altered in Table 5. The input texts consist of the syllable patterns that encompass the whole of the syllable pattern in Indonesian.

The experimental results showed that the system can visualize the pronunciation of phonemes and syllables precisely and smoothly. Transition animation of visualizing the mouth shape of the phonemes and syllables are presented subtly. It does not look disjointed between one of visualization of mouth shapes to each other, as illustrated in Fig. 8. This system can be used by learners of Indonesian or foreign learners. In general, the level of the learners is divided into beginner-level learners, intermediate and advanced [19]. For the beginner level, the learners can use this system to visualize the words of greeting, forms a simple sentence of active, passive, negative and prepositions. While for the intermediate and advanced levels, learning is distinguished by the complexity of the sentence.

**Table 5.** Indonesian Sentences Used as Input in the System Testing

| No | Indonesian Sentences |
|---|---|
| 1 | baju warna biru itu mahal harganya |
| 2 | perusahaan ekspor impor itu maju pesat |
| 3 | sepatuku kotor belum aku cuci |
| 4 | strategi promosi yang tepat dapat meningkatkan penjualan |
| 5 | prasarana gedung itu sangat memadai |
| 6 | aku pergi ke toko baju bersama ibu |
| 7 | praktek dokter umum itu setiap hari |
| 8 | struktur organisasinya sangat kompleks |
| 9 | buanglah sampah di tempatnya |
| 10 | teks proklamasi dibacakan oleh presiden Sukarno |

In this study, visualizing the pronunciation of phonemes is built based on the syllable concatenation approach. Visualization of phonemes pronunciation in a syllable patterns is strongly influenced by the phoneme before or after. Fig. 8 (b) shows that the visualization of the phoneme pronunciation *'b'* and *'r'* is influenced by the following phonemes. Therefore, the results of the syllable pronunciation visualization look smoother. Because of the pronunciation visualization of displacement phoneme does not occur drastically as in Indonesian Text-to-Audiovisual Synthesis System Based on Phoneme Speech Database [6][20]. This method is an implementation of the dynamic visualization of phoneme pronunciation.

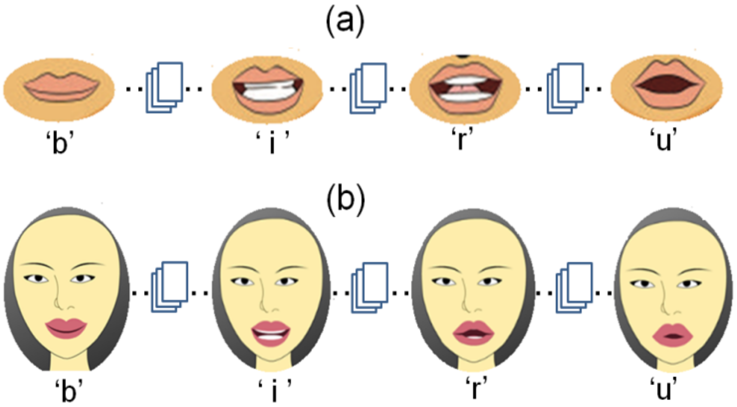

**Fig. 8.** Visualization of the transitions animation each phoneme on (a) Indonesian Text-to-Audiovisual System Based on Phoneme Speech Database, (b) Indonesian Text-to-Audiovisual System Based on Syllable Speech Database

The system is tested on 30 respondents to measure the degree of correspondence between the pronunciation visualization and speech. Respondents are people who understand the science of Indonesian phonology namely students and lecturers of the departments of Indonesian language. Respondents evaluate the degree of correspondence between the pronunciation visualization and speech by providing the criteria value as shown in Table 6. Whereas, the assessment results of all respondents is shown in Table 7. It is averaged by using MOS (mean opinion score) as in (15). The result of the calculation using the MOS to the assessment data from all respondents is 4.283. This shows that the degree of correspondence between the pronunciation visualization and speech of this system is good.

**Table 6.** Criteria of the Degree of Correspondence Between Visualization Pronunciation and Speech

| The Value of MOS | Quality | Description |
|---|---|---|
| 5 | Very Good | very appropriate |
| 4 | Good | Appropriate |
| 3 | Adequate | fairly appropriate |
| 2 | Bad | less appropriate |
| 1 | Very Bad | Not appropriate |

$$MOS = \sum_{i=1}^{n} \frac{x(i).k}{N} \tag{15}$$

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

In this experiment, we also compare the testing results between this system and the system's Indonesian text-to-audiovisual which uses a speech database phoneme-based from our previous study [6]. Fig. 9 illustrates the testing results of system text-to-audiovisual for Indonesian both of the speech database phoneme-based and the speech database syllable-based. Test results show that in general the system text-to-audiovisual for Indonesian based on the syllable speech database can produce the corresponding degree of the pronunciation visualization is better.

**Table 7.** The Results of User Assessment

| No | Indonesian Sentences | The Degree of Correspondence | | | | |
|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* |
| 1 | baju warna biru itu mahal harganya | 0 | 1 | 3 | 6 | 20 |
| 2 | perusahaan ekspor impor itu maju pesat | 0 | 1 | 7 | 7 | 15 |
| 3 | sepatuku kotor belum aku cuci | 0 | 0 | 4 | 6 | 20 |
| 4 | strategi promosi yang tepat dapat meningkatkan penjualan | 0 | 4 | 6 | 10 | 10 |
| 5 | prasarana gedung itu sangat memadai | 0 | 3 | 6 | 9 | 12 |
| 6 | aku pergi ke toko baju bersama ibu | 0 | 0 | 5 | 7 | 18 |
| 7 | praktek dokter umum itu setiap hari | 0 | 0 | 7 | 9 | 14 |
| 8 | struktur organisasinya sangat kompleks | 1 | 2 | 7 | 9 | 11 |
| 9 | buanglah sampah di tempatnya | 0 | 0 | 1 | 9 | 20 |
| 10 | teks proklamasi dibacakan oleh presiden Sukarno | 0 | 0 | 4 | 6 | 20 |

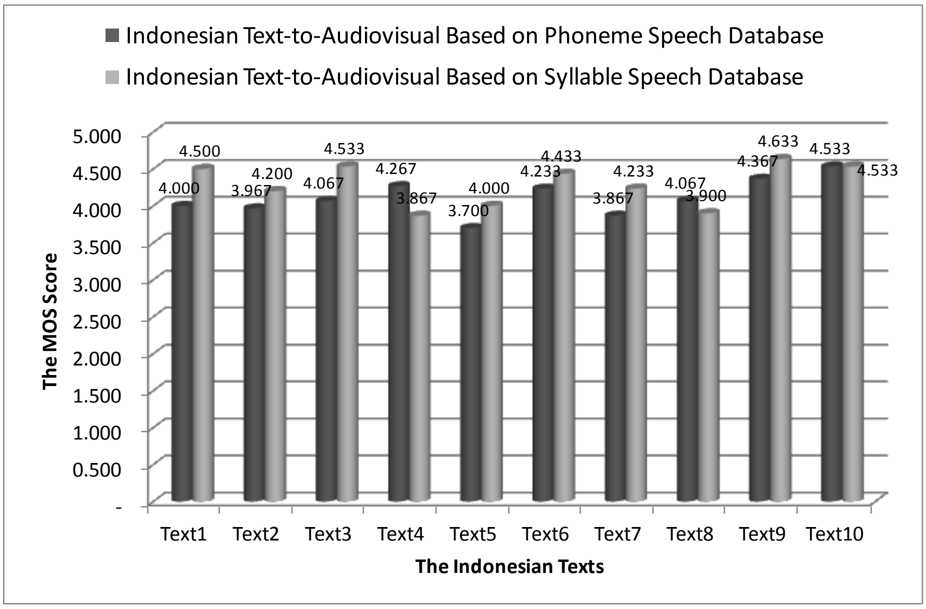

**Fig. 9.** Comparison of the testing results in system's text-to-audiovisual for Indonesian based on the syllable speech database and system's text-to-audiovisual for Indonesian based on a phoneme speech database

The pronunciation visualization of syllables with patterns of ‘V’, ‘CV’, ‘CVC’, ‘VC’, ‘CCV’, ‘CCVC’, ‘VCC’, ‘CVCC’, ‘CCVCC’, can generate visualization more realistic and smoother. However, the syllables with patterns of 'CCCVC' and 'CCCV' as in the words *‘struk-tur’* and *‘stra-te-gi’* occurs overlapping visualization consonant phonemes. The visualization of consonant phonemes *‘r’* in the word ‘*struk-tur’* and *‘stra-te-gi’* still looks overlapping. In fact, the pronunciation of phoneme *'r'* is not clearly visible due to the influence of phonemes before and after the phoneme. Therefore, the pronunciation visualization of the phoneme *'r'* in the word *‘struk-tur’* and *‘stra-te-gi’* should be displayed vague, because it is represented by the phoneme *‘t’* and *‘u’* in the word *‘struk-tur’*.

## 6 Conclusion and Future Work

Indonesian text-to-audiovisual system can be used as one of the multimedia-based applications to learning Indonesian. This system can be run using Internet technologies to support distance learning methods. The system can visualize how to pronounce a phoneme or syllables in Indonesian. Therefore, System's text-to-audiovisual Indonesian can be used by learners of Indonesian and foreign learners.

Based on the result of these experiments, we conclude that the Indonesian text-to-audiovisual system has produced the pronunciation visualization more realistic and smoother than an Indonesian text-to-audiovisual system based on a phoneme speech database. It is based on the calculation result of the MOS methods to system assessment by the respondents. Fig. 9 shows that the Indonesian text-to-audiovisual system has produced the pronunciation visualization more realistic and smoother to the syllables with patterned of ‘V’, ‘CV’, ‘CVC’, ‘VC’, ‘CCV’, ‘CCVC’, ‘VCC’, ‘CVCC’, ‘CCVCC’. Meanwhile, the pronunciation visualization of the syllables with a pattern of 'CCCVC' and 'CCCV' still makes an overlapping visualization mainly in consonant phonemes. The pronunciation visualization of phoneme should be displayed with a minimum number of frames (thin) to avoid an overlapping visualization.

In the future, the formula to calculate the number of frames exactly for the syllables patterned 'CCCVC' and 'CCCV' is required. The exact number of frames of each phoneme impact on smoother and more realistic pronunciation visualization, especially on the syllables pattern of 'CCCVC' and 'CCCV'.

## 7 Acknowledgment

The authors would like to appreciate the Scholarship Program of the Bureau of Planning and foreign cooperation of the Ministry of National Education Republic of Indonesia who 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 conducted the study on motion capture technology in the laboratory of HCS (Human Centric System) in electrical engineering department of the *Institut Teknologi Sepuluh Nopember (ITS) Surabaya*. The author also would like to appreciate ‘*Program Penelitian Hibah Bersaing Direktorat Penelitian dan Pengabdian*

*Kepada Masyarakat Direktorat Jenderal Pendidikan Tinggi'* of the Ministry of the National Education Republic of Indonesia for the help in completing THIS RESEARCH.

## 8 References

[1] Wai-Kim Leung, Ka-Wa Yuen, Ka-Ho Wong and Helen Meng, "Development of Text-to-Audiovisual Speech Synthesis to Support Interactive Language Learning on a Mobile Device", 4th IEEE International Conference on Cognitive Infocommunications, pp. 583-588, December 2–5, 2013. https://doi.org/10.1109/coginfocom.2013.6719170

[2] Wojowasito, S., "*Perkembangan Ilmu Bahasa (Linguistik) Abad 20*", Bandung: Shinta Dharma, 1976.

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

[4] Salil Deena, Shaobo Hou and Aphrodite Galata, "Visual Speech Synthesis by Modelling Coarticulation Dynamic using a Non-Parametric Switching State-Space Model", School of Computer Science, university of Manchester, UK, 2010. https://doi.org/10.1145/1891903.1891942

[5] Hui Zhao and Chaojing Tang, "Visual Speech Synthesis Based on Chinese Dynamic Visemes", Proceeding of the 2008 IEEE International Conference on Information and Automation, June 20-23, Zhanjiajie, China, 2008. https://doi.org/10.1109/ICINFA.2008.4607983

[6] 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, ISSN 1828-6003, pp. 1149-1156, November 2015.

[7] Johan Wouters, Michael W. Macon, "Control of Spectral Dynamics in Concatenative Speech Synthesis", IEEE Transaction on Speech and Audio Processing, Vol. 9, No. 1, pp. 30-38, January 2001. https://doi.org/10.1109/89.890069

[8] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Lain Matthews, "Dynamic Units of Visual Speech", ACM SIGGRAPH Symposium on Computer Animation, 2012.

[9] Chaer, Abdul, "*Fonologi Bahasa Indonesia",* Jakarta: PT. Rineka Cipta, pp. 103, 2007.

[10] 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. https://doi.org/10.1109/cyberneticscom.2013.6865781

[11] Chaer, Abdul, "*Linguistik Umum",* Jakarta: PT. Rineka Cipta, 2003.

[12] Turk MA and Pentland AP., "Face Recognition Using Eigenfaces", *IEEE* Transactions on Pattern Analysis and Machine Intelligence, pp. 586-591, 1991. https://doi.org/10.1109/cvpr.1991.139758

[13] 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.

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

[15] 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. https://doi.org/10.1109/ICEEI.2011.6021792

[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] T. Ezzat and T. Poggio, "Visual Speech Synthesis by Morphing Visemes", International Journal of Computer Vision, vol.38, no.1, pp.45-57, 2000. https://doi.org/10.1023/A:1008166717597

[18] Arry Akhmad Arman, "*Definisi Text to Speech (Departemen Teknik Elektro ITB: Bandung)*", URL: http://indotts.melsa.net.id/ Definisi Text to Speech « Teknologi Bahasa.htm (date accessed 1 Februari 2016).

[19] Imam Suyitno, "*Pengembangan Bahan Ajar BIPA Berdasarkan Hasil Analisis Kebutuhan Belajar*", Wacana Vol. 9 No. 1, pp. 62-78, 2007. https://doi.org/10.17510/wjhi.v9i1.223

[20] Muljono, Surya Sumpeno, Dhany Arifianto, Kiyoaki Aikawa, Mauridhi Hery Purnomo, "Developing an Online Self-learning System of Indonesian Pronunciation for Foreign Learners", International Journal of Emerging Technologies in Learning, Vol. 11, No. 04, pp. 83-89, 2016. https://doi.org/10.3991/ijet.v11i04.5440

## 9 Authors

**Arifin** 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 Departement of Dian Nuswantoro University Semarang, Indonesia (email: arifin@dsn.dinus.ac.id). Since 2011, he has been studying at the Graduate School of Department of Electrical Engineering, Institut Sepuluh Nopember Surabaya (ITS) Indonesia as a doctoral student. He works as a lecturer of Informatics Engineering Departement of Dian Nuswantoro University, Semarang. His research interests includes natural language processing and human-computer interaction.

**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 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.

**Mochamad Hariadi** received the B.E. degree in Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 1995 (email: mochar@ee.its.ac.id). He received both M.E. and Ph. D. degrees in the Graduate School of Information Science Tohoku University Japan, in 2003 and 2006 respectively. He is currently teaching in Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS). His research interests are Video and Image Processing, Data Mining and Intelligent System. He is a member of IEEE and a member of IEICE.

**Arry Maulana Syarif** received his M.Kom. degree in 2013 from the Informatics Engineering Departement of Dian Nuswantoro University Semarang, Indonesia (email: arry_m@dsn.dinus.ac.id). He is currently teaching at the Faculty of Computer Science of Dian Nuswantoro University Semarang. His research interests are data mining and artificial intelligence.

Article submitted 31 October 2016. Published as resubmitted by the authors 06 December 2016.