Improving Phoneme to Viseme Mapping for Indonesian Language

Anung Rachman[1], Risanuri Hidayat[2], Hanung Adi Nugroho[3]

***Abstract*—The lip synchronization technology of animation can run automatically through the phoneme-to-viseme map. Since the complexity of facial muscles causes the shape of the mouth to vary greatly, phoneme-to-viseme mapping always has challenging problems. One of them is the allophone vowel problem. The resemblance makes many researchers clustering them into one class. This paper discusses the certainty of allophone vowels as a variable of the phoneme-to-viseme map. Vowel allophones pre-processing as a proposed method is carried out through formant frequency feature extraction methods and then compared by t-test to find out the significance of the difference. The results of pre-processing are then used to reference the initial data when building phoneme-to-viseme maps. This research was conducted on maps and allophones of the Indonesian language. Maps that have been built are then compared with other maps using the HMM method in the value of word correctness and accuracy. The results show that viseme mapping preceded by allophonic pre-processing makes map performance more accurate when compared to other maps.**

***Keyword*—Phoneme-to-Viseme Mapping, Allophones, Vowels, Formant Frequencies, Lip-Reading.**

I. INTRODUCTION

Lip synchronization is an important aspect of an animated clip. Researchers have done a lot of research for a better level of accuracy on lip synchronization for voice or text input. Synchronization can be done through a map between phonemes and viseme. The viseme is a visual phoneme, just like a phoneme but in a visual form. Phoneme-to-viseme (P2V) maps are needed because some phonemes have a viseme shape that is similar (visually ambiguous).

Improved performance of P2V maps has been done. The goal is that visual lip movements feel natural like humans when talking. Map improvements follow the most developments in image feature extraction methods. The complexity of facial muscles causes very varied lip movements. Therefore, the accuracy of lip visual feature extraction greatly influences the performance of P2V maps. The more accurate the image feature extraction, the more accurate P2V maps. Others are the improvement of pronunciation detection of articulation so that there is a new method of dynamic phoneme-to-viseme maps (many-to-many) [1]-[3]. Dynamic viseme can improve the accuracy of lip movements, but this method also has a disadvantage, the problem of efficiency. Viseme searches have exponential time complexity, requiring a long time for each sentence [4]. This will be a problem when the system is implemented for real-time animation. Map performance improvements were also made in the phoneme confusion section [5]. Whereas another research uses linguistics to create P2V maps and uses surveys for validation [6].

Vowels take the dominant portion of lip animation because it requires the greatest energy and the longest duration of time in the speech signal [7] so that the presence of vowels on the P2V map takes the most important role. The problem is, often a vowel has a variety of different pronunciations called allophones. Allophones are more complex pronunciation variations. These allophones represent phonemes in certain situations, such as context, tone, and duration [8]. The vowel allophone impact on the P2V map is as follows.

- If an allophone is not accommodated on a P2V map then there is the potential for misclassification of viseme when someone pronounces the allophone in lip-sync animation technology.
- If an allophone is accommodated on a P2V map even though the allophone is similar to the allophone next to it, then the map compression factor becomes of smaller value. The smaller the compression value of P2V maps, the worse the map performance. The explanation of compression factors can be seen in the next sub-chapter.

Therefore, the ideal state of P2V maps is if the existence of allophonic vowels is indeed necessary. A study showed a comparison of five English P2V maps from five previous researchers as shown in Table I [9].

TABLE I
COMPARISON OF P2V MAP [9]

Map	Phonemes	Vowels	Viseme Classes
Jeffers [10]	43	17	4
Neti [11]	42	17	4
Hazen [12]	52	16	5
Bozkurt [13]	45	17	7
Lee [14]	39	14	7

The vowel phoneme column in Table I shows some differences in the initial data used by five researchers. There are those who use vowels a number of 17, 16, and 14. This difference occurs because the understanding of linguistics is also different. Even though, as described above, the number of vowel phonemes affects the performance of P2V maps.

Allophonic vocal variety is also experienced by countries that have neighboring clusters, for example between Korean, Uyghur, and Mandarin [8]. Whereas English as an international language also has allophones, especially when spoken by people from countries that do not use English as a national language. For example, English is spoken by Javanese [15] or

[1,2,3] *Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jl. Grafika No. 2 Kampus UGM, Yogyakarta 55281 INDONESIA (tlp: 0274-552305; fax: 0274-547506;*
e-mail: [1]anung.rachman@mail.ugm.ac.id, [2]risanuri@ugm.ac.id, [3]adinugroho@ugm.ac.id)

by Spanish [16], and vice versa, the person in the main English-speaking languages who speak Korean [17].

In this paper, we study the problem in the certainty of the existence of allophone vowels on P2V maps. The certainty of the existence of vowels is not only important for making maps efficient but also to accommodate the possibility of allophonic vowel input in lip animation. In particular, we study allophonic vowels /e/ in Indonesian for P2V maps.

II. Problem with Allophone of Indonesian Language in Phoneme to Viseme Maps

Indonesian has many regional variations. When spoken as a second language, it is strongly influenced by the local language of the person speaking [18]. Javanese as the most widely used regional language in Indonesia has an influence on pronunciation in Indonesian [19]. Javanese has eight vowel phonemes: six phonemes and two allophonic pairs /e1//e3/ and /o1// o2/ [20]. Whereas the next study states that Javanese vowels are grouped into six phonemes including four allophonic pairs /i1//i2/, /u1//u2/,/e1//e3/, and /o1//o2/ [21]-[23].

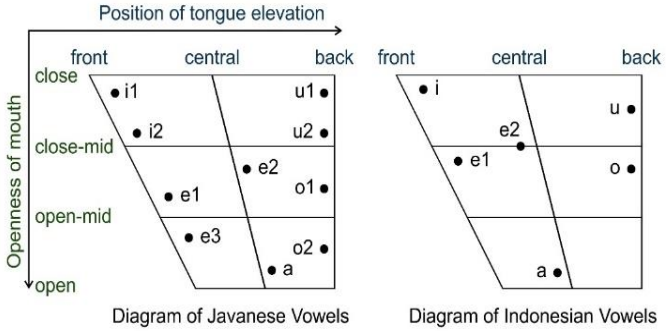

Fig. 1 Javanese vowel system (left) [21]-[23], Indonesian vowel system (right) [24].

Fig. 1 shows the differences in the Javanese and Indonesian vowel systems. It can be seen that vowel variations occur more in the Javanese vowels. Though the most dominant Javanese vowels affect the pronunciation of the Indonesian vowels. This vowel variety should be accommodated in the initial data to build a P2V map. Especially for vowel /e/, seen in Fig. 1, the Java Vowel contains three vowels /e1//e2//e3/ and Indonesian Vowels contain only two vowels /e1//e2/. The pronunciation of words containing vowel /e3/ will potentially reduce the level of lip-sync accuracy by referencing P2V maps which only contain vowels /e1/ and /e2/.

Table II shows the Indonesian P2V map, using two vowels /e/ [25], three vowels /e/ [6], and just one vowel /e/ [26], and this occurs because of different linguistics understanding from researchers. P2V maps can be translated as a ratio called compression factor [27] through (1).

$$(CF = nV/nP(1)) \quad (1)$$

with *CF* is a compression factor, *nV* is the number of visemes, and *nP* is the number of phonemes.

The ideal state of P2V maps is when *CF* = 1, that is, when the number of phonemes is equal to the number of viseme classes produced, meaning that all phoneme characteristics are uniquely distinguished [28]. When the initial vowel phoneme on the P2V map (*nP*) is determined using only linguistics, it will potentially cause the value of the compression factor to be low. This means that it also has the potential to cause map performance to decrease.

Table II
Viseme Classes of Indonesian Vowels in P2V Map

Map	Vowel Viseme Classes
Arifin et al. [25]	[/a/][/o//u/][/i/][/e1/][/e2/]
Setyati et al. [6]	[/a/][/i1//i2/][/e1//e2//e3/][/o1//o2/][/u1//u2/]
Liyanthy et al. [26]	[/a/][/i/][/u/][/e/][/o/]

Table III
Compression Factor of Indonesian Vowels in P2V Map

Map	*nV*:*nP*	*CF*
Arifin et al. [25]	5:6	0.83
Setyati et al. [6]	5:10	0.5
Liyanthy et al. [26]	5:5	1

Table III shows the value of the compression factor for the P2V map in Indonesian specifically for the vowel part. Liyanthy's map [26] seems to have the most ideal *CF* value, but if seen in Table II, the map does not contain allophones, so if there is an input signal containing allophonic vowels it will potentially be classified into the wrong viseme. Therefore, it should be the initial selection of allophonic vowel phonemes can be ascertained, so that their existence is really needed.

III. Phoneme-to-Viseme Mapping

To solve the problem of selecting initial allophonic vowel data, we propose a data-driven pre-processing method that follows linguistics. Pre-processing P2V maps function to find the significance of the difference in allophonic audio features. The underlying concept is the sound traits produced by humans that are closely related to the shape of the mouth (as visual) as the forming part of the tone. Pre-processing on allophones besides being used to ascertain the initial data, is also used to compare the results of visual feature extraction in the classification section, if appropriate, P2V maps are expected to be more accurate. One of the sound characteristics that can be used is formant frequency. In phonetic science, a formant is a spectral formation that results from acoustic resonance on the human vocal tract [27]. But the definition in acoustics is different, a formant is the peak of the spectrum [29]. For harmonic sounds, with this definition, the spectrum in question is in the form of resonance. Formants are often measured as the peak amplitude of the sound frequency spectrum.

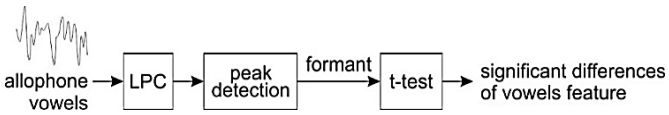

Fig. 2 Pre-processing system overview.

Fig. 2 shows an overview of pre-processing for selecting initial allophonic vowel data. Linear Predictive Coding (LPC) is used to generate resonance signals which are then carried out peak detection to produce formant frequencies as audio features

[30]. Formant frequencies that can be used as features are the first two frequencies F1 and F2 [31], with the input signal current samples $x(n)$ and past samples $x(n\text{-}k)$, $k = 1 \ldots p$ the LPC equation as follows.

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) \quad (2)$$

with a_k is the coefficient that forms a resonance signal.

The formant extraction results were then compared through t-test statistics to determine the significant differences between allophonic vowels.

A. Data Preparation

The complexity of facial muscles causes the shape of the mouth to vary greatly. Visual extraction of the mouth from several people (independent speakers) causes always problematic [32]. In the visual domain, until now there has been no universal approach that can be accepted to overcome this [33]. Therefore, the system to be built in this research is a type of speaker-dependent. And allophones used as samples in this paper are vowels /e/ in Indonesian.

The data used in this research consists of several types as follows.

- Sound recordings containing a total of 60 words that contain predefined words containing vowel /e1//e2/ and /e3/. The recording process uses a sample rate of 24 kHz, mono channel, 32-bit depth, and uses the MP3 format. This recording is used to find significant differences between vowels /e/.
- Recordings of face video containing vowel and syllable utterance containing Indonesian phonemes. Furthermore, the video is extracted into facial images of various forms of the mouth that say phonemes. The resulting face image has a resolution of 720x1280 24bit jpg format. This recording is used to build a P2V map.
- Sound recording contains a variety of sentences that have been determined as many as 50 sentences for HMM training, and another for 20 sentences used for testing phoneme maps to viseme. This recording uses a sample rate of 16 kHz, mono channel, 16-bit depth, and uses the audio waveform format.

B. Classification Method

After knowing the significant differences of allophonic, preliminary data on facial images containing subsequent phonemes were arranged including allophones that were suitable from the results of pre-processing. This initial data is then used to build a P2V map.

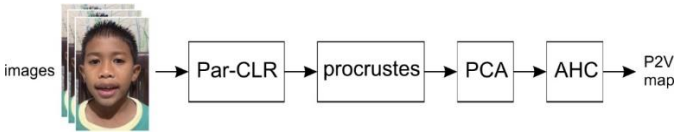

Fig. 3 Phoneme-to-viseme mapping system overview.

Fig. 3 is the process of building a P2V map. Face images as the starting material are extracted by means of Parallel-Cascade Linear Regression (Par-CLR) [34]. This feature is a coordinate point called a landmark with 12 points on the outer lip, and six points on the inner lip. The feature extraction equation for each stage of the Par-CLR regression is as follows.

$$B_t = B_{t-1} + R_t\left(f_t(C, B_{t-1})\right) \quad (3)$$

with B_t is a shape, B_{t-1} is a shape of estimation results from the previous stage, B_0 is a reference landmark, R_t is a regressor, f_t is a function that extracts pixel features in C_t images.

The process starts from the reference landmark B_0 which placed on a face image C_t as input. Using the image features extracted through the f_t function, the Regressor R_t fitting each landmark point from the reference B_0 to the location on the lip of the image feature. The fitting process takes place several times for each landmark point so that the landmark produces a new shape B_t at the location of the lip.

Each landmark then measured its level of similarity using Procrustes Analysis [35]. This method compares the similarities between the two lip shapes. One shape is used as a reference, and the other shape is transformed by placing it in the same place, rotating it, and scaling it to approach the reference shape. After that, the two shapes be measured the distance of similarity. The similarity data of each lip shape is then arranged into a matrix 34x34 according to the number of viseme variables including the shape of the silent lips.

Then the matrix variables were simplified through Principal Component Analysis (PCA) [36] to produce the location of each phoneme in a 3-dimensional Cartesian diagram derived from factor loadings.

$$Factor\ Loadings = Eigenvector \cdot \sqrt{Eigenvalue}. \quad (4)$$

The 34x34 matrix variable is simplified becomes to the 34x3 matrix. Simplified 3-column data (as the axis of the Cartesian diagram) should of course contain 34 simplified column data characteristics. These characteristics are eigenvectors and eigenvalues. Both of these features are obtained from the value of the correlation between variables.

The position of 3-dimensional coordinates (factor loadings) of each phoneme is then grouped into the viseme class through the Agglomerative Hierarchical Clustering (AHC) method [37], [38], based on the degree of dissimilarity. The linkages used are Unweighted Pair Group Average linkage (UPGA). At each step of the UPGA, the two closest clusters are combined with a higher cluster. The distance between two clusters a and b, each size |a| and |b| (number of variables/digits) is considered to be the average of all distances $d(x,y)$ between pairs of objects x at a, and y in b, meaning the average distance between variables from each cluster.

$$\frac{1}{|a| \cdot |b|} \sum_{x \in a} \sum_{y \in b} d(x, y) \quad (5)$$

and the distance metric used is Manhattan distance.

$$ManhattanDistance[\{a, b, c\}, \{x, y, z\}] = \text{Abs}[a - x] + \text{Abs}[b - y] + \text{Abs}[c - z]. \quad (6)$$

C. Comparison of Phoneme-to-Viseme Map

The map that has been produced as a result of the construction is then tested to see the effect of the allophone pre-

processing. Map testing was carried out using the HMM method [39] as had been done other researches [9], [13], [28], [40]. Data transcription is changed from phoneme to viseme. This change is adjusted to the map that has been generated in the previous process. Also included as a reference map is a map belonging to previous researchers with the aim of comparing map performance.

The step taken in testing is determining training and testing sentences by taking into account vowel allophones /e/ which are then manifested in the form of sound recordings and transcriptions. The training section will produce a model set that will be used as a pattern to recognize words in the testing section. The next step is to convert the transcription of the phoneme into a viseme classes according to each map to be tested. Testing will result in the form of the number of words recognized.

The metrics used to measure success in this research are word correctness and accuracy through the following equation.

$$H = \frac{N-D-S}{N} \quad (7)$$

$$Acc = \frac{N-D-S-I}{N} \quad (8)$$

with H is word correctness, Acc is accuracy, N is the number of test samples, D is deletion error, S is substitution error, and I is insertion error.

IV. Experiments Results and Discussion

The word used as the research sample for one vowel was 20 voice records which were divided into two parts, i.e. vowels on the front and back, with ten words voice recorded for each part. The vowels studied are as much three (/e1//e2//e3/), so the total word data is 60.

To see more about the difference in formant frequencies values with respect to the vowel location in words, then a comparison of three populations of vowel /e/ can be seen graphically through a distribution plot.

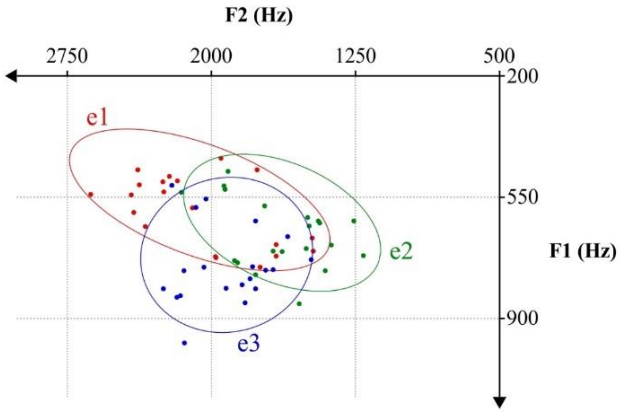

Fig. 4 Formant distribution of vowel /e1//e2/ and /e3/.

As seen at Fig. 4, the three vowels /e1//e2//e3/ have adjacent frequency regions. Some frequencies of sound samples tend to have the same features shown by overlapping areas. But for a few words that contain all three vowels have different angles of frequency area, /e1/ tend to low F1 and high F2, /e2/ tend to F1 and F2 which are low, and /e3/ tend to high frequencies F1 and F2.

Although there are parts that overlap on the formant distribution plot for the three vowels in the word, the difference in each /e/ level of significance needs to be further examined. If one part is significantly different from the other parts, then the average formant for that part is not feasible to be used as a vowel reference on the system.

Comparison is done for each of the two sets of words with different of vowels /e/, i.e. /e1/ with /e2/, /e2/ with /e3/, and /e3/ with /e1/. This aims to get a more detailed level of difference. To get the significance of these differences, the t-test statistical function is used, the results can be seen in Table IV.

TABLE IV
T-TEST OF VOWELS /E/

Vowel Comparison		t-test	
		F1	F2
e1	e2	**0.04517**	**0.00006**
e2	e3	**0.00845**	**0.00052**
e3	e1	**0.00004**	0.13118

Bold: significant different

In the t-test, the difference is generally considered significant if the probability is less than 0.05. In Table IV formant F1, the comparison of vowel /e1/ and /e2/ shows that the value of the t-test is 0.04517, which means that the difference is significant or heterogeneous. In F1 and F2, only between vowel /e3/ and /e1/ on F2 are homogeneous.

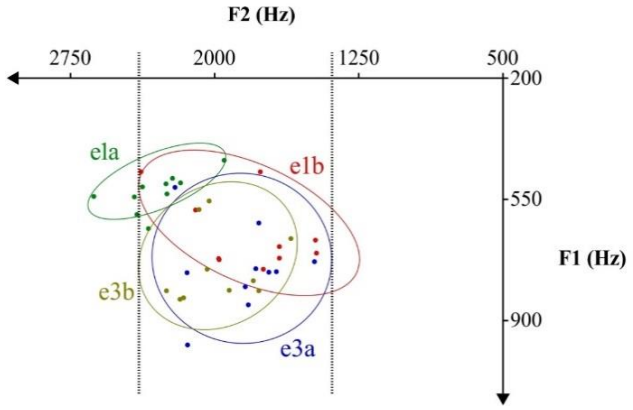

Fig. 5 Formant frequency range F2, between /e1b/ (vowel /e1/ which is at the back of a word) has the same tendency compared to /e3a/ and /e3b/ (vowel /e3/ which is both in front of and back a word).

Voice data samples of the vowel /e/ in the words on this research consist of two types, i.e. /e1/ which is at the beginning of the word and at the end of the word, as well as vowel /e2/ and /e3/. To see further why between vowels /e1/ and /e3/ homogeneous on F2, Fig. 5 shows a plot of formant distribution that is separate between vowel /e/ in the front and the back of the word. The Indonesian vowel /e1/ which is on the front of the word is labeled as /e1a/, and the one at the back of the word is labeled as /e1b/. While label /e3a/ is vowel /e3/ which is at the front of the word, and /e3b/ is vowel /e3/ which is on the back of the word. It can be seen in Fig. 5 that /e1b/ has formant frequency range F2 which tends to be equal to /e3a/ or /e3b/.

This means that Indonesian vowel /e1/ on the back of a word has similar characteristics to vowel /e3/ in a word.

However, in Fig. 5 it also shows why /e1/ has a significant difference to /e3/ in F1, which tends to be different in /e1a/ and a small portion /e1b/, and the distribution circle /e1a/ also is alone. To accommodate /e1a/ then it should be /e1/ and /e3/ also on the P2V map.

Fig. 6 shows the phoneme location between the population as a result of the variable simplifying from the Procrustes result through the PCA method. Phonemes are mapped on 3-dimensional coordinate diagrams i.e. F1, F2, and F3 (XYZ).

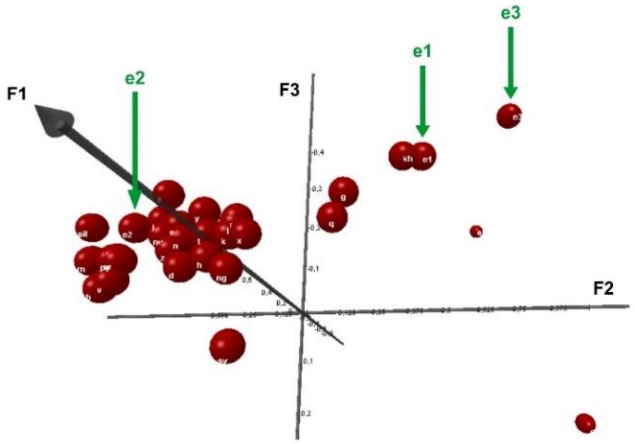

Fig. 6 Vowel location of /e1/, /e2/, and /e3/ at 3-dimensional coordinates, factor 1 (F1), factor 2 (F2), and factor 3 (F3).

Differences in visual features between vowels /e/ proved to be suitable with pre-processing vowel allophones. This confirms that the P2V mapping algorithm has appropriate. While the P2V map results of the mapping in this research (Anung et al.) It can be seen in Table V. While the comparison of vowel specific compression factors from the P2V Indonesian map can be seen in Table VI.

TABLE V
PHONEME TO VISEME MAP (ANUNG ET AL.)

Viseme Class	Phonemes
V01	/a/
V02	/b//e2//f//m//p//sil//v//w/
V03	/c//d//h//j//k//l//n//ng//ny//r//s//t//x//y//z/
V04	/e1//kh/
V05	/e3/
V06	/g//q/
V07	/i/
V08	/o1/
V09	/o2/
V10	/sy/
V11	/u/

It can be seen in Table VI that the value of the map compression factor (CF) of this study is better than the two other maps. This happens because the number of vowel viseme classes is equal to the number of vowel phonemes involved. The number of viseme classes especially for vowels /e/ is in accordance with the pre-processing which distinguishes the vowel /e/ characteristics significantly.

TABLE VI
COMPARISON OF THE COMPRESSION FACTOR VALUE

Map	Only Vowels		Vowels and Consonants	
	nV:nP	***CF***	***nV:nP***	***CF***
Anung et al.	8:8	1	11:33	0.333
Arifin et al. [25]	5:6	0.833	9:32	0.281
Setyati et al. [6]	5:10	0.5	11:49	0.224

A comparison of map performance can be seen through the value of word correctness and accuracy as in Table VII. While Fig. 7 is a graph of the relationship between word correctness and accuracy with compression factors.

TABLE VII
COMPARISON OF WORD CORRECTNESS AND ACCURACY

Map	% Correctness	Accuracy
Anung et al.	52.04 %	38.46
Setyati et al. [6]	41.63 %	35.29
Arifin et al. [25]	33.48 %	22.17

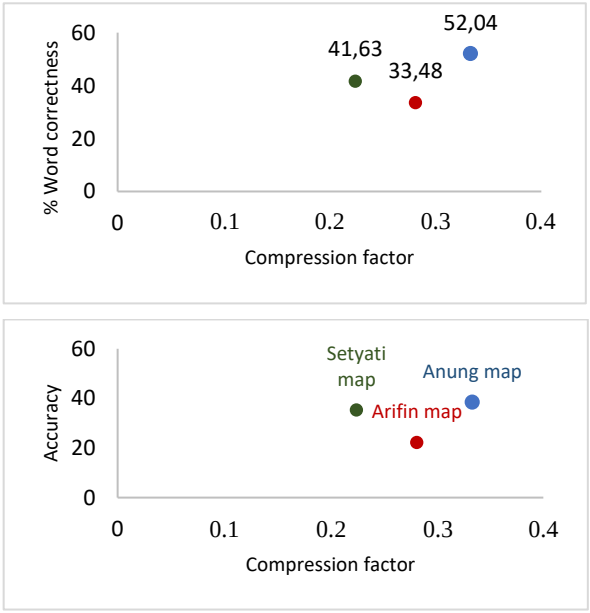

Fig. 7 The relationship between compression factors of the map with word correctness (top) and accuracy (bottom).

Fig. 7 shows the improved performance of the map constructed in this study compared to the two P2V Indonesian maps that have existed before. Anung map with pre-processing shows a word correctness value of 52.04% better than the Setyati map by 41.63% and Arifin map by 33.48% without pre-processing. While the level of accuracy also shows the same thing. Through testing with the test words that take into account the three vowels /e/, Anung map with three different viseme classes is able to recognize words more accurately than the other two maps. Anung and Setyati maps each contain 11 viseme classes, but three vowels /e/ on Anung map are separated in three different viseme classes (Table V), while three vowels /e/ on Setyati maps occupy only one viseme class (Table II). While the factor that influences the performance of the Arifin map is located on maps that only contain nine classes of viseme.

V. CONCLUSION

The pre-processing method is designed to deal with the problem of the certainty of vowel allophones on P2V maps for lip animation. So far, the selection of allophones is done randomly, or only through linguistic knowledge. The certainty of the existence of allophone vowels is needed to support the performance of phoneme-to-viseme maps. If it is supposed to exist but not, or vice versa, it will potentially affect the performance of the map.

Pre-processing is done by finding for the significance of differences vowel allophone sounds feature, and the results of pre-processing of the three vowels /e/ Indonesian in this study show a significant difference in sound characteristics as indicated by the results of the t-test below 0.05 for formant values. Therefore, the lip visual of the three vowels /e/ should also be different, and this shown in this study with a different viseme class for each vowel /e/.

This research proves that the pre-processing of vowel allophones contributes to improving the performance of phoneme-to-viseme maps compared to other maps. The map performance test results showed the word correctness value of 52.04% and an accuracy of 38.46 higher compared to the other two Indonesian P2V maps.

REFERENCES

[1] S. Taylor, B.J. Theobald, and I. Matthews, "A Mouth Full of Words: Visually Consistent Acoustic Redubbing," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4904–4908.

[2] C.F. Rademan and T. Niesler, "Improved Visual Speech Synthesis Using Dynamic Viseme k -means Clustering and Decision Trees," *Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, 2015, pp. 169–174.

[3] Arifin, S. Sumpeno, Muljono, and M. Hariadi, "A Model of Indonesian Dynamic Visemes from Facial Motion Capture Database Using a Clustering-based Approach," *IAENG International Journal of Computer Science*, Vol. 44, No. 1, pp. 41–51, Feb. 2017.

[4] S.L. Taylor, "Discovering Dynamic Visemes," Doctoral dissertation, University of East Anglia, Norwich, UK, 2013.

[5] P. Shih, A. Paul, J. Wang, and Y. Chen, "Speech-Driven Talking Face Using Embedded Confusable System for Real Time Mobile Multimedia," *Multimedia Tools and Applications*, Vol. 73, No. 1, pp. 417–437, Nov. 2014.

[6] E. Setyati, S. Sumpeno, M.H. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo, "Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation," *IAENG International Journal of Computer Science*, Vol. 42, No. 3, pp. 233–244, Jul. 2015.

[7] S.-M. Hwang, H.-K. Yun, and B.-H. Song, "Korean Speech Recognition Using Phonemics for Lip-Sync Animation," *Information Science, Electronics and Electrical Engineering (ISEEE)*, 2014, pp. 1011–1014.

[8] J. Xu, J. Pan, and Y. Yan, "Agglutinative Language Speech Recognition Using Automatic Allophone Deriving," *Chinese Journal of Electronics*, Vol. 25, No. 2, pp. 328–333, Mar. 2016.

[9] L. Cappelletta and N. Harte, "Phoneme-to-Viseme Mapping for Visual Speech Recognition," *1st International Conference on Pattern Recognition Applications and Methods (ICPRAM) Volume 2*, 2012, pp. 322–329.

[10] J. Jeffers and M. Barley, *Speechreading (lipreading)*, 1st ed. Springfield, USA: Charles C. Thomas Publisher, 1971.

[11] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio Visual Speech Recognition," IDIAP, Workshop Final Report, 2000.

[12] T.J. Hazen, "Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp. 1082–1089, May 2006.

[13] E. Bozkurt, Ç.E. Erdem, E. Erzin, T. Erdem, and M. Özkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic lip Animation," *3DTV Conference*, 2007, pp. 1–4.

[14] S. Lee and D. Yook, "Audio-to-Visual Conversion Using Hidden Markov Models," *Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 2002, pp. 563–570.

[15] A.A. Perwitasari, M. Klamer, J. Witteman, and N.O. Schiller, "Vowel Duration in English as a Second Language Among Javanese Learners," *International Conference on Phonetic Sciences*, 2015, pp. 1-4.

[16] L. Burrows, L. Jarmulowicz, and D.K. Oller, "Allophony in English Language Learners: The Case of Tap in English and Spanish," *Language, Speech, and Hearing Services in Schools*, Vol. 50, No. 1, pp. 138–149, Jan. 2019.

[17] S.-G. Bae, B.-M. Lim, and M.-J. Bae, "A Comparative Analysis on Allophone of Korean for English Natives," *Information*, Vol. 20, No. 5(A), pp. 3291–3298, May 2017.

[18] E. van Zanten, "The Indonesian vowels : Acoustic and Perceptual Explorations," Doctoral dissertation, Rijksuniversiteit te Leiden, Netherlands, 1989.

[19] N. Adisasmito-Smith, "Phonetic and phonological influences of Javanese on Indonesian," Doctoral dissertation, Cornell University, New York, USA, 2004.

[20] E.C. Horne, *Beginning Javanese*, 3rd ed. London, UK: Yale University Press, 1961.

[21] K.M. Dudas, "The Phonology and Morphology of Modern Javanese," Doctoral dissertation, University of Illinois, Urbana-Champaign, USA, 1976.

[22] K. Hayward, "Lexical Phonology and the Javanese Vowel System," *SOAS Working Papers in Linguistics*, Vol. 9, pp. 191–225, 1999.

[23] Wedhawati, W.E.S. Nurlina, E. Setiyanto, R. Sukesi, Marsono, and I.P. Baryadi, *Tata Bahasa Jawa Mutakhir*, Revisi ed. Yogyakarta, Indonesia: Penerbit Kanisius, 2006.

[24] C.D. Soderberg and K.S. Olson, "Indonesian," *Journal of the International Phonetic Association*, Vol. 38, No. 2, pp. 209–213, Aug. 2008.

[25] Arifin, Muljono, S. Sumpeno, and M. Hariadi, "Towards Building Indonesian Viseme: A Clustering-Based Approach," *IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, 2013, pp. 57–61.

[26] M. Liyanthy, H. Nugroho, and W. Maharani, "Realistic Facial Animation Of Speech Synchronization For Indonesian Language," *3rd International Conference on Information and Communication Technology (ICoICT)*, 2015, pp. 563–567.

[27] I.R. Titze, R.J. Baken, K.W. Bozeman, S. Granqvist, N. Henrich, C.T. Herbst, D.M. Howard, E.J. Hunter, D. Kaelin, R.D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D.G. Miller, H. Noé, R.C. Scherer, J.R. Smith, B.H. Story, J.G. Švec, S. Ternström, and J. Wolfe, "Toward a Consensus on Symbolic Notation of Harmonics, Resonances, and Formants in Vocalization," *J. Acoust. Soc. Am.*, Vol. 137, No. 5, pp. 3005–3007, May 2015.

[28] H.L. Bear and R. Harvey, "Phoneme-to-Viseme Mappings: The Good, the Bad, and the Ugly," *Speech Communication*, Vol. 95, pp. 40–67, Dec. 2017.

[29] "American National Standard Acoustical Terminology," ANSI S1.1-1994 (ASA 111-1994), Standards Secretariat, Acoustical Society of America, New York, 1994.

[30] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, Vol. 7, No. 1, pp. 29–32, Feb. 1988.

[31] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 7th ed. Stamford, USA: Cengage Learning, 2014.

[32] S.J. Cox, R.W. Harvey, Y. Lan, J.L. Newman, and B.-J. Theobald, "The Challenge of Multispeaker Lip-Reading.," *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2008, pp. 179–184.

[33] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A Review of Recent

Advances in Visual Speech Decoding," *Image and Vision Computing*, Vol. 32, No. 9, pp. 590–605, Sep. 2014.

[34] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1859–1866.

[35] C. Goodall, "Procrustes Methods in the Statistical Analysis of Shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 53, No. 2, pp. 285–339, Jan. 1991.

[36] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, No. 1–3, pp. 37–52, Aug. 1987.

[37] K. Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm-A Review," *International Journal of Scientific and Research Publications*, Vol. 3, No. 3, pp. 1–3, Mar. 2013.

[38] W.H.E. Day and H. Edelsbrunner, "Efficient Algorithms for Agglomerative Hierarchical Clustering Methods," *Journal of Classification*, Vol. 1, No. 1, pp. 7–24, Dec. 1984.

[39] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. (Andrew) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK version 3.4)," Cambridge University Engineering Department, 2006.

[40] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," *19th European Signal Processing Conference (EUSIPCO)*, 2011, pp. 2109–2113.