Developing an Online Self-learning System of Indonesian Pronunciation for Foreign Learners

http://dx.doi.org/10.3991/ijet.v11i04.5440

Muljono[1,2], Surya Sumpeno[1], Dhany Arifianto[1], Kiyoaki Aikawa[3] and Mauridhi Hery Purnomo[1]
[1] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2] Dian Nuswantoro University, Semarang, Indonesia
[3]Tokyo University of Technology, Tokyo, Japan

***Abstract*—The main part of learning a language is pronunciation. In language learning method, pronunciation practice requires more portion than the language theory. There are some obstacles experienced by foreign learners to learn Indonesian because they are still strongly influenced by their mother tongue which is really different from Indonesian. There are some courses of learning Indonesian, indeed, but the foreign learners have to stay in Indonesia to join them. On the other hand, the researchers have successfully proven that the use of Information and Communication Technology (ICT) can help the learners in learning a language. In this paper, we have developed a system of the Online Self-Learning of Indonesian Pronunciation for Foreign Learner using Indonesian Text to Audio Visual Speech which is able to help the foreign learners to overcome their obstacles in learning Indonesian, especially the pronunciation. This system consists of 2(two) application modules: Indonesian Text to Speech (ITTS) and Indonesian Text to Audio Visual Speech (ITTAVS). In order to find out whether this system is feasible or not for foreign learners' skill in pronouncing the Indonesian words, a subjective measurement using subjective test Mean Opinion Score (MOS) is used. We organized native speakers (Indonesians) as the participants of this test. Some of them are lecturers of Indonesian language and can be considered as experts. The average scores (using MOS scale) of the tests given showed a promising result. This system is dedicated to the foreign learners who need to improve their skill in pronouncing the Indonesian words accurately and to change the classical method of learning into a self-learning method.**

***Index Terms*—multimodal, ITTAVS, ITTS, online self-learning, pronunciation, Indonesian**

I. INTRODUCTION

Indonesia has to be ready for the challenges of the global market, especially ASEAN Economic Community (AEC). Many foreigners come to and work in Indonesia in almost all business sectors, such as: education, service, industry, and so forth. Also, there are a lot of foreign companies established in Indonesia. Some of them are joint-venture companies as well. Moreover, some of Indonesian companies have been taken over by foreign investors. These phenomena cause the rise of the population of foreign workers in Indonesia.

The Indonesian Government implements a policy that a company which hires foreign workers has to do a technology transformation from the foreign workers to the local workers. The foreign workers will deliver the technology transformation smoothly if they can speak Indonesian well. This is one example of an online-learning system of Indonesian pronunciation can be a good use.

We have developed and dedicated a system of learning Indonesian for the foreign workers. The foreign workers mentioned herein means those who have different background of language and culture from the local workers and meet a major obstacle in learning Indonesian caused by the influence of their mother tongue [1].

The process of learning Indonesian for foreign learners and natives are significantly different. The natives will take advantage because Indonesian is their mother tongue, but this is not the case for the foreign learners. The major obstacle faced by the foreign learners; it is not only the pronunciation of a word but also the pronunciation of a sentence. It can be said that it is naturally caused by the number of different phonemes that every language has.

Some Indonesian courses for the foreign learners are held in Indonesia. The learners who want to join must come to Indonesia. It becomes a main problem for them.

A discourse on holding an online Indonesian-learning course for the foreign learner has been issued lately[1]. Unfortunately, it has not been realized, yet. On the other hand, for other language, such as English, the online English-learning system has been set up in many developed countries in America and Europe. The learners can do a self-learning and can interact with other people worldwide using an online learning application; not only a single modal but also multimodal.

In classic class, some modals can be used to create an interaction between teacher and learners. Modal is a method used to interact with other people using the sense organs. An interaction using a sensory organ is called a single modal, while an interaction using some sensory organs simultaneously is called multimodal. Multimodal represents the content of knowledge in different modes; it generally occurs in verbal form (such as text and the word that is pronounced) and in non-verbal form (such as: illustration, film, video, and animation) [2].

Multimodal needs not only to access the information in different formats but also to build an interaction among these representations [3]. Multimodal interactive learning environment is an environment where the learner during their learning time can interact with the available contents (such as to play/stop, to skip back/forward the video which they are watching and others) and can click on the hyperlink to obtain further information [2].

[1] Agency for Development and Improvement of Language, Ministry of Education and Culture of The Republic of Indonesia. http://badanbahasa.kemdikbud.go.id/bipa/v2/

Some researchers have developed some techniques of learning a language. Sun [4] has created a multimodal learning application for an online English learning and enriched it with the latest technology. Royce [5] has researched and synergized distinctive symbols in multimodal and implemented some modes in a language course. According to Stein [6], each of communication and interaction come up in the class is a multimodal and the teaching must be concentrated on the environment of multimodal characteristic. Martinec [7] proposed a framework to research a discourse of multimodal which is built by images and language. Based on Martinec's opinion, the linguists must pay more attention to the relationship between the discourse of multimodal and some synthesis of communication modes (such as language, dynamic and static pictures, and 3D objects).

A new method of learning Indonesian language for the foreign learner, especially in its pronunciation, is proposed in this paper. It focuses on 2 (two) ways: the online learning method using multimodal and changing the face-to-face meeting method. In the online learning method using multimodal, we use an Indonesian Text To Audio Visual Speech (ITTAVS) application which we have built to generate audio visual speech from the given text.

II. INDONESIAN TEXT TO AUDIO VISUAL SPEECH

Text To Speech (TTS) is a system that is able to convert a text message into a relevant verbal message [8]. It is widely used to help build an easier communication, especially for whoever has a lack of visibility. It is used for screen reader, speech to speech machine translation system, e-mail reader, SMS reader, storytelling, talking head system, call center, and others as well. TTS will give more benefits to the blind and the deaf because it can be applied as a self -learning method to obtain the speech instantly by giving a text input.

TTS facilitates the normal users to listen to (not to read) the received information from a text message. They do not need to read the text anymore [9]. So, they can free their eyes from reading the text as well.

In communicating with each other through speech, it is not only the audio but also the human non-verbal information (visual speech, face expression, and body language) needed. They contribute to the representation of the people's emotion and expression. In the same way, this system must be designed to convert the text into the sound signal and ideally added a visual speech which must be synthesized together at the same time. This system is called Text To Audio Visual Speech (TTAVS). We have developed TTAVS for Indonesian as a part of our proposed method. It is called ITTAVS.

ITTAVS is essentially a system that consists of ITTS module, visual phoneme (viseme) sub-system and web-based presentation sub-system . In order to make ITTAVS, we need Indonesian speech corpus and build Indonesian Visemes.

A. Indonesian Data Preparation

We prepare the data and tools to make the ITTAVS, such as Indonesian phonemes and visemes, speech corpus, tools to convert text into phonemes, and tools to generate sounds and to synchronize phonemes, sounds, and visemes.

1) Indonesian phonemes

The Indonesian Alphabet System contains 26 alphabets or letters. It is similar to the Latin alphabet system, whereas the Indonesian phonemes consist of 33 phonemes including allophones[10]. Table 1 outlines the Indonesian phonemes.

TABLE I.
PRONUNCIATION OF INDONESIAN PHONEME

Phoneme	IPA Symbol	Pronounced
a	ɑ	like *a* in *father*
e 1	ə	like *a* in *about* or *ago*.
e 2	ɛ	like *e* in *bed* or *dress*
i	i	like *ee* in *meet*
o	o	like *o* in *low*
u	u	like *oo* in *moon*
ai	ɑi	like *i* in *five*
au	ɑu	like *o* in *now*
ei	ei	like ay in say
oi	oi	like oy in boy
b	b	like *b* in *bed*
c	tɕ	like *ch* in *church*
d	d	like *d* in *dog*
f	f	like *f* in *fat*
g	g	like *g* in *gun*
h	h	like *h* in *happy*
j	dʑ	like *j* in *judge*
k	k	like *k* in *keep*
l	l	same as in English
m	m	same as in English
n	n	same as in English
p	p	like *p* in *push*
q	k	more like *k* rather than *q*
r	r	more like the trilled or rolled Spanish *r*
s	s	same as in English
t	t	like *t* in *tap* or *top*
w	w	same as in English
x	ks	same as in English.
y	j	like *ye* as in *yes* or *yet*
z	z	same as in English
kh	x	*ch* in the Scottish word *loch*
ng	ŋ	*ng* in *singing*
ny	nj	*ny* in *canyon*

2) Indonesian Visemes

In our previous research [11], we built Indonesian Viseme using clustering method. The data clustered were the Visual Speech Images taken from a 6:36 minute video covering all Indonesian Phonemes. After preprocessing the video, we derived 1000 frame visual speech images as the data training to be clustered. The combination of Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) were used to extract the features. K-Means algorithm was used as the clustering method. The result was 10 classes of Indonesian Viseme.

We enhanced the result by conducting further experiments and validated through a survey. As a final result, we obtained 14 classes of Indonesian Viseme. These 14

visemes are utilized to visualize the speech in our ITTAVS application throughout this paper.

3) Indonesian Speech Corpus

To build TTS a speech corpus is needed. We use the speech corpus which is called Id1 and made by Arry Akhmad Arman. It is available publicly in the website MBROLA [12]. This speech corpus is used in Indonesian Text To Speech based on concatenation diphone method [13].

B. Building Indonesian Text To Audio Visual Speech

The Indonesian Text To Speech (ITTS) module which is used to convert text into speech has to be created before building the Indonesian Text To Audio Visual Speech (ITTAVS). Figure 1 describes ITTS module as part of ITTAVS sub-system. The ITTS module uses a diphone synthesizer. There are 2(two) main components installed in the ITTS module in order to produce a high quality of synthesized sound. They are the natural language processor (front-end) component to convert text into phoneme; and sound generator (back-end) component to generate phoneme to sound. Front-end component analyzes the text and predicts the prosody procedure needed to convert the orthographic text into an accurate phonetic transcription (i.e. phoneme) and the expected intonation and rhythm (i.e. prosody). There are a number of methods suggested and applied to front-end component [14]. Back-end component receives the input as a sequence of phonemes and prosody description which is the output of the front-end component, and generates them into a sound. Some techniques of generating a sound have been suggested and applied in some speech synthesis research[14].

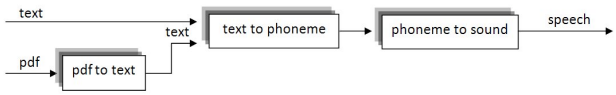

Figure 1. ITTS module. Front-end component converts text into phoneme and predicts the prosody built using Microsoft Speech SDK 5.1 and back-end component generates speech developed by MBROLA

We use Microsoft Speech SDK 5.1 and MBROLA to implement the front-end and the back-end components in the making of ITTS. Microsoft Speech SDK 5.1 is used to analyze the text in forming the sequence of phonemes and to predict the prosody from the text input. Afterwards, the order of phonemes and the integrated prosody, obtained from the Microsoft Speech SDK 5.1, are sent to MBROLA, as the back-end component, to generate speech based on the given text input. E-speak is used to integrate corpus Id1 into Microsoft SDK 5.1 and MBROLA.

After completing the ITTS module, the ITTAVS sub-system is built to generate the visual speech. ITTAVS sub-system is illustrated in Figure 2. Text or pdf file can be used as the input data. If the input data is a pdf file, the sub-system converts it first into a text. Then, the text is changed into a sequence of phonemes. The sequence of phonemes will be converted into a sound, so that the speech will be generated based on the given text input. At the same time, the sequence of phonemes formed will be mapped onto a sequence of viseme defining a visual speech. The speech and visual speech will be generated simultaneously.

Figure 3 shows the display of the menu of the ITTAVS system. The user can type a text inside a "Text Box" or upload the pdf file. To upload pdf file user can browse file from local folder by clicking on “Browse” button, select it and click on "Upload PDF" button. Pdf file will be automatically converted into text and displayed inside a "Text Box".

User then click on "Speak" button to listen generated speech from the text and look at generated visual speech which is visualized by the animated talking head. The speech and visual speech are generated simultaneously based on the input given, so that the learner can learn to pronounce the inputted word or sentence by imitating the speech and visual speech represented by the animated talking head.

To increase or decrease the speech tempo, the user can click on the ‘+’ or ‘-‘ buttons in the Rate menu. To turn up or turn down the volume, the user can click on the ‘+’ or ‘-‘ buttons in the Volume menu.

C. Web-based Presentation Sub-System

We use XAMPP to develop web-based presentation sub-system in our system. XAMPP is a software package which provides Apache (used for the web-server), PHP (as a script language program to develop the web application), and MySQL (as a database management system to save all of the uploaded pdf files).

The next process is installing the ITTAVS sub-system in the web-server. We tested our system using major browsers such as Internet Explorer, Firefox and Chrome. When the system is activated, the user can input a text or upload a pdf file into the “text box”. The system will generate speech and visual speech based on the input given and represent them in the animated talking head simultaneously.

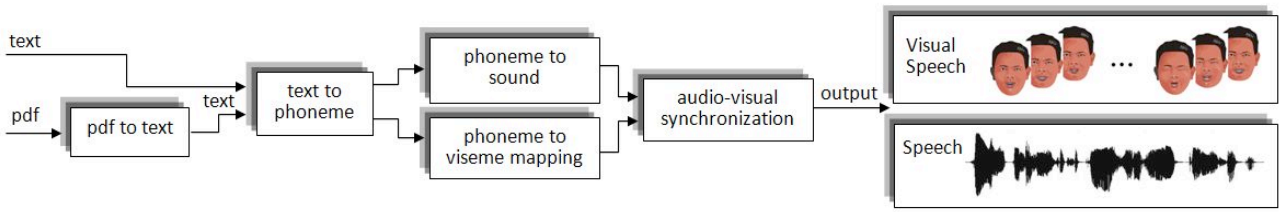

Figure 2. Indonesian Text To Audio Visual Speech sub-system

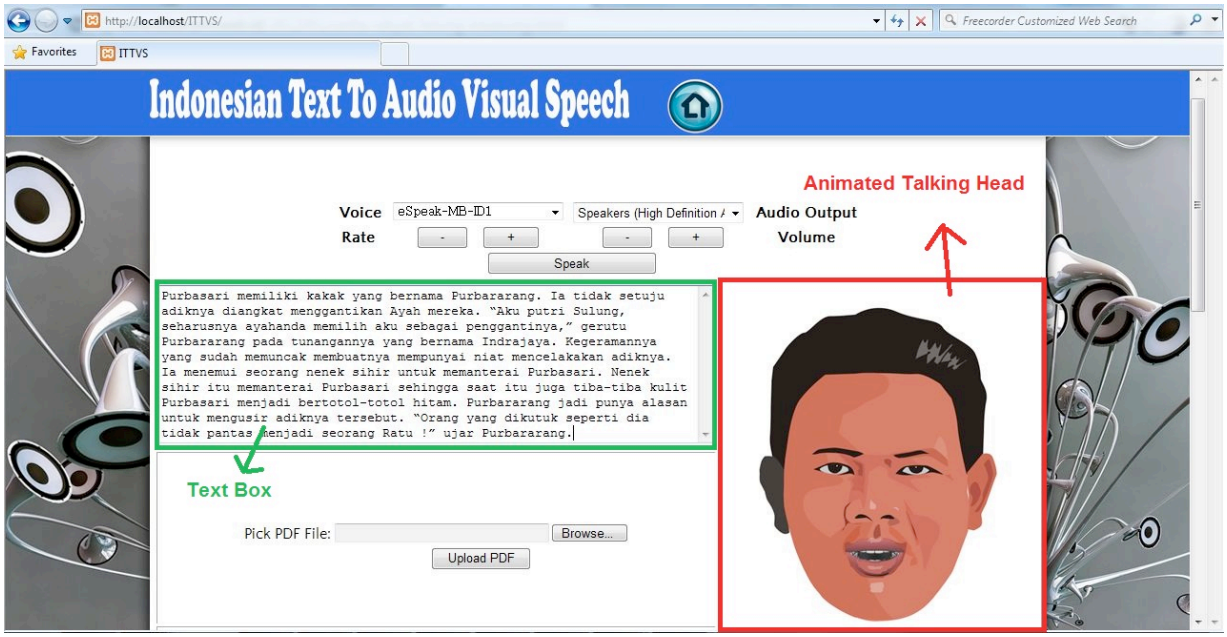

Figure 3. Indonesian Text To Audio Visual Speech system menu

III. BUILDING THE ONLINE SELF-LEARNING SYSTEM OF INDONESIAN PRONUNCIATION

After completing the ITTAVS sub-system, we built an online self-learning system of Indonesian Pronunciation. We used ADDIE model in building that system. ADDIE model is a design framework of Instructional System containing a list of general processes using instructional design and training development [15]. It is a descriptive guidance in building a training tool and effective performance support. There are 5(five) phases in building an online self-learning system of Indonesian pronunciation: analysis, design, development, implementation, and evaluation.

A. Analysis Phase

This phase is to analyze the importance of the model development and the more advanced method of Indonesian pronunciation learning. There are some obstacles occurred to the foreign learners to learn Indonesian because they are still strongly influenced by their mother tongue which is really different from Indonesian.

Moreover, they must join an Indonesian course, mostly held in Indonesia, and attend all meetings if they want to learn Indonesian. This face-to-face learning method will obstruct them to learn Indonesian, most notably in practicing the pronunciation further on.

The more advanced learning system is the online self-learning system of Indonesian pronunciation, as an alternative to replace the face-to-face learning method. This system is expected to be able to help foreign learners adapt the Indonesian pronunciation by themselves through the online system because it generates audio and visual speech based on the given text or pdf file instantly.

B. Design Phase

The main aim of designing our system is to provide a self-learning system of Indonesian pronunciation for the foreign learners. Since the multimedia technology grows rapidly, the researchers have a wide opportunity to develop a self-learning system using various multimodal online [16].

The online learning system provides flexibility for the learner to use and to access. No matter where they are. By functioning the ICT devices and the internet connection, the learner can connect to our proposed system with no boundaries. In addition to that, in the learning process, the learner can exploit his or her sense organs by using the multimodal as the structure of interaction with the online learning system.

The learning model is defined as the cognitive characteristic, affective attitude, physiological behavior as a relatively stable indicator, and how the learners perceive, interact, and respond with their learning atmosphere. The learners will have their own comfortable place to study as well[17]. They will prefer learning something using a certain modality, such as writing, reading, speaking, visualizing, or doing a physical response. Most learners are also keen on combining some modalities abovementioned.

We have designed the system by combining multimodal: text, sound, and visual speech. It will provide some advantages for the learners. The learners can access and interact with the ITTAVS self-learning system online. It also provides menu to input a text or a pdf file that will be converted into speech and visual speech. The text input can be numbers, alphabets, words, phrases, even sentences. The learners are able to learn about the pronunciation of the text inputs containing the intonation and visual speech which is represented by animated talking head.

C. Development Phase

The Development applied in the ADDIE model contains activities of realizing the product design. A conceptual framework of the model application, the method of this system, has been composed in this phase. In the development phase, the conceptual framework will be realized as a product that is ready to be implemented. An Online Self-learning Tool of a Language Pronunciation has been created in this phase.

Figure 4 describes the system of the Online Self-learning of Indonesian Pronunciation for Foreign Learner using Indonesian Text To Audio Visual Speech. The learners can access this system through the internet connection. Then, they have to input texts or upload a pdf file into the system. The system converts the text into speech and visual speech and sends them as the output from the system to the learner. Thus, through that system, the learners can do an Online Self-learning of Indonesian Pronunciation using Indonesian Text To Audio Visual Speech.

D. Implementation Phase

The Implementation phase is the actual phase to set up the learning system which we have developed. It means that in this phase we have installed and set up the related software, modules, and others simultaneously in our system, so that it will work properly.

For prepare the online self-learning of Indonesian Pronunciation system, we install XAMPP as the software for the web server. Afterwards we continue to install the ITTAVS sub-system into the web server. PHP software which support a web application is used to create the ITTAVS. For the database need, we use MySQL software.

The system can be accessed by running the Internet Explorer 8, as the web browser, through the local server. We had conducted an implementation of our system to 21 participants (6 of them are lecturers of Indonesian) before we implemented it in the actual purpose (to public). Each of them is asked to use that system.

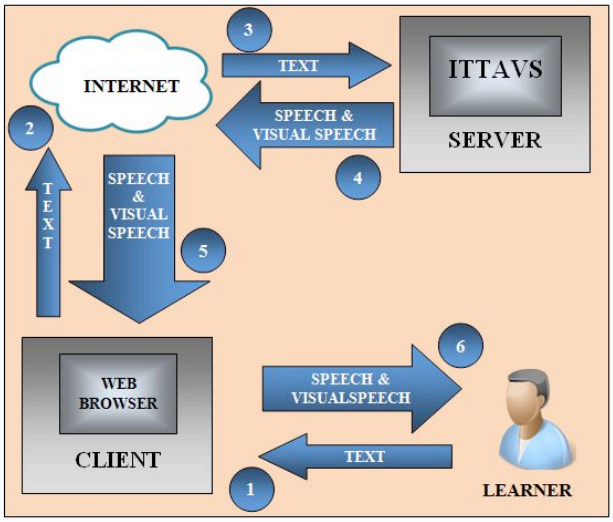

Figure 4. System of The Online Self-learning of Indonesian Pronunciation

E. Evaluation Phase

A subjective test, a test based on the people's opinion, is used to evaluate this system. By using a subjective test, we can obtain the valid feedback which also represents the human perception in this system. It can be used to justify the standard quality of a system as well.

Before delivering the test, we had provided the users' needs such as a quiet and convenient room, standard LCD screen placed in the normal distance, hi-quality earphones or headphones, and even the participants' mood, to avoid the bias scoring or the result [18]. The test was conducted to gain some information whether this system was suitable for the users or not. The number of the participants and the materials tested must be representative to obtain the average score based on the certain criteria. The number of respondents participating on this test were 21 native adults (21 to 50 years old); 6 of them were lecturers of Indonesian. Some points tested on this system used subjective measurement, based on the people's opinion and the analysis of them. The advantage of using subjective measurement is the people can interact with the system directly and the can convey their perception of the system which can be considered as a standard of quality of the system.

The criteria used to represent the performance of this test are as follows:

1) Intelligibility of the synthesized speech

We use word recognition test [18] to measure the intelligibility. In this test, the participants will listen words ("word test") or sentences ("sentence test") synthesized and be asked to write what they have listened. The 'word test' consists of 10 groups of words in which each group contains 3 different words. The words in each group are almost similar. They differ in only a single consonant used in each word, such as 'aman' (secure), 'akan' (will), 'amal' (charity), or 'bapak' (father), 'banyak' (many), 'bajak' (plowing). The 'sentence test' used consists of 10 Semantically Unpredictable Sentences (SUS), sentences which are syntactically correct but have no meaning such as 'Kuda salah belajar di kebun minggu aman' (Horse learns wrong in the garden sunday securely). There are 30 words and 10 sentences totally delivered to the participants. Each of them will be played once. The score is taken from the accumulation of the correct answers compared to the whole words and sentences tested. The result of word -recognition tests is that 83.86% of the given words and 84.74% of the given sentences can be caught accurately.

2) Naturalness of the synthesized speech

We use Mean Opinion Score (MOS) [19] to measure the naturalness. The participants are asked to input 30 different words to the system and they will be converted into speech output. How natural the output speech generated by system has to be scored by the participants. They can score the result ranging from 1 to 5, namely scale 1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent. This test can describe the naturalness of the synthesized speech produced by the system. The result of this test (Test-1) shown in Figure 5 is 4.00, which means that the naturalness of the synthesized speech produced by this system is good.

3) Naturalness of the mouth movements

The participants are directed to score the naturalness of the mouth movements which is displayed in the visual speech. They are ordered to score the visual speech without worrying the sound of the test data sample produced so that the sound will not influence them in their scoring process. However, the volume of the sound is still on [20].

In this test, the participants can input their score using MOS scale ranging from 1 to 5, namely scale 1: Bad (extremely rough), 2: Poor (rough), 3: Fair (fairly smooth), 4: Good (smooth), and 5: Excellent (truly smooth). It means that in scale 5, the mouth movement is really smooth resembling the real visual speech. On the contrary, scale 1 means that the mouth movement is extremely rough and really out of what is expected. The result of this test (Test-

2) shown in Figure 5 is 3.71 which means that the naturalness of the mouth movements displayed by the animated talking head is close to good.

4) The accuracy of the synchronization of the sound with the mouth movements

The participants are asked to give their score the accuracy of the synchronization of the sound with the mouth movements [20] . The measurement tool used is MOS scale ranging from 1 to 5, namely scale 1: Bad (extremely asynchronous), 2: Poor (asynchronous), 3: Fair (fairly synchronous), 4: Good (synchronous), and 5: Excellent (truly synchronous). The result of this test (Test-3) shown in Figure 5 is 3.72 which means that the accuracy of the synchronization of the sound with the mouth movement displayed by the animated talking head is nearly synchronous.

5) The design of the user interface menu

The participants are required to score the design of the user interface menu using MOS scale ranging from 1 to 5, namely scale 1: Bad (extremely uninteresting), 2: Poor (uninteresting), 3: Fair (fairly interesting), 4: Good (interesting), and 5: Excellent (absolutely interesting). The result of this test (Test-4) described in Figure 5 is 3.67 which means that the design of the user interface menu is nearly interesting.

6) The user- friendliness of the system

The participants are asked to score the user-friendliness of this system using MOS scale ranging from 1 to 5, namely scale 1: Bad (extremely difficult to operate), 2: Poor (difficult to operate), 3: Fair (quite easy to operate), 4: Good (easy to operate), and 5: Excellent (absolutely easy to operate). The result of this test (Test-5) shown in Figure is 3.95 which means that the system is close to user-friendly.

7) Properness of the system

In this test, the participants are ordered to score whether this system is proper to be applied to the Self-learning of Indonesian Pronunciation for the foreign learners or not. The MOS scale is used in this test. It ranges from 1 to 5, namely scale 1: Bad (extremely improper), 2: Poor (improper), 3: Fair (fairly proper), 4: Good (proper), and 5: Excellent (absolutely proper). The result of this test (Test-6) outlined in Figure 5 is 3.81 means that the system is close to proper.

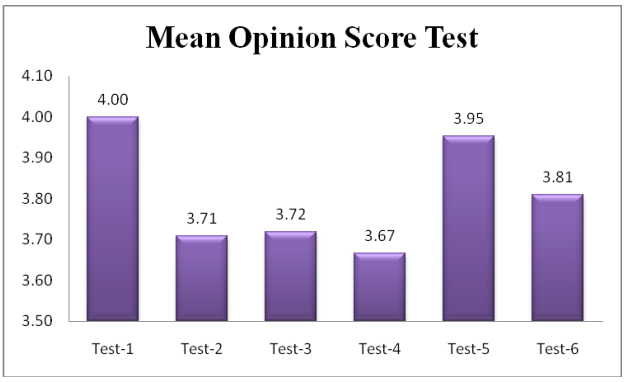

Figure 5. The Result of Mean Opinion Score Test (Test-1 : Naturalness of the synthesized speech , Test-2 : Naturalness of the mouth movements , Test-3 : The accuracy of the synchronization of the sound with the mouth movements, Test-4 : The design of the user interface menu , Test-5 : The user-friendliness of the system . Test-6 : Properness of the System)

IV. CONCLUSION

This research shows that our proposed ITTAVS based on the multimodal method is able to be applied to support the learners in learning Indonesian pronunciation. An Online Self-learning System of Indonesian Pronunciation for The Foreign Learner has been designed and built using ITTAVS. This system has been tested to and evaluated by 21 participants; 6 of them are lecturers of Indonesian. The MOS of the given tests explains that the intelligibility of the synthesized speech, naturalness of the synthesized speech, naturalness of the mouth movements, the accuracy of the synchronization of the sound with the mouth movements, design of the user interface menu, user-friendliness of the system, and properness of the system, has shown the promising results. We hope that this system is able to be applied and used as an online self-learning system of Indonesian pronunciation for foreign learners.

In addition, this system can be developed as the Indonesian learning aid to the deaf (as the speaking training tool) and to the blind who has a lack of visibility.

REFERENCES

[1] R. Ellis, *Understanding Second Language Acquisition*, Some Highlighting edition. Oxford ; New York: OUP Oxford, 1985.

[2] R. Moreno and R. Mayer, "Interactive Multimodal Learning Environments," *Educ. Psychol. Rev.*, vol. 19, no. 3, pp. 309–326, Jun. 2007. http://dx.doi.org/10.1007/s10648-007-9047-2

[3] N. Guichon and S. McLornan, "The effects of multimodality on L2 learners: Implications for CALL resource design," *System*, vol. 36, no. 1, pp. 85–93, Mar. 2008. http://dx.doi.org/10.1016/j.system.2007.11.005

[4] M. Sun, "Application of Multimodal Learning in Online English Teaching," *Int. J. Emerg. Technol. Learn. IJET*, vol. 10, no. 4, pp. 54–58, Sep. 2015. http://dx.doi.org/10.3991/ijet.v10i4.4697

[5] T. Royce, "Multimodality in the TESOL Classroom: Exploring Visual-Verbal Synergy," *TESOL Q.*, vol. 36, no. 2, pp. 191–205, Jul. 2002. http://dx.doi.org/10.2307/3588330

[6] P. Stein, "Rethinking Resources in the ESL Classroom: Rethinking Resources: Multimodal Pedagogies in the ESL Classroom," *TESOL Q.*, vol. 34, no. 2, pp. 333–336, Jul. 2000. http://dx.doi.org/10.2307/3587958

[7] R. Martinec and A. Salway, "A system for image–text relations in new (and old) media," *Vis. Commun.*, vol. 4, no. 3, pp. 337–371, Oct. 2005. http://dx.doi.org/10.1177/1470357205055928

[8] D. Govind and S. R. Prasanna, "Expressive speech synthesis: a review," *Int J Speech Technol*, vol. 16, no. 2, pp. 237–260, Jun. 2013. http://dx.doi.org/10.1007/s10772-012-9180-2

[9] Y. Zhixiao, C. Fengjuan, and Z. Dexian, "A virtual classroom platform based on text to visual speeches," in *IEEE International Symposium on IT in Medicine and Education, 2008. ITME 2008*, 2008, pp. 312–315.

[10] A. Chaer, *Fonologi Bahasa Indonesia*. Rineka Cipta, 2009.

[11] Arifin, Muljono, S. Sumpeno, and M. Hariadi, "Towards building Indonesian viseme: A clustering-based approach," in *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, 2013, pp. 57–61.

[12] "The MBROLA Project [Online]." Available : http://tcts.fpms.ac.be/synthesis/mbrola.html.

[13] A. A. Arman, "Konversi dari Teks ke Ucapan." Sep-2004.

[14] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, vol. 3. Dordrecht: Springer Netherlands, 1997. http://dx.doi.org/10.1007/978-94-011-5730-8

[15] *Designing Effective Instruction*, 5 edition. Hoboken, NJ: Wiley, 2006.

[16] D. Birch and M. Gardiner, "Students' perceptions of technology-based marketing courses," in *Proceedings of Australia and New Zealand Marketing Educators Conference*, Fremantle, Australia, 2005.

[17] A. P. Gilakjani, H. N. Ismail, and S. M. Ahmadi, "The Effect of Multimodal Learning Models on Language Teaching and Learning," *Theory Pract. Lang. Stud.*, vol. 1, no. 10, Oct. 2011. http://dx.doi.org/10.4304/tpls.1.10.1321-1327

[18] A. W. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 77–80.

[19] P. Taylor, *Text to speech synthesis*. Cambridge: Cambridge University Press., 2009. http://dx.doi.org/10.1017/CBO9780511816338

[20] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Commun.*, vol. 55, no. 7–8, pp. 857–876, Sep. 2013. http://dx.doi.org/10.1016/j.specom.2013.02.005

AUTHORS

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

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 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 an IAENG and IEEE member.

Dhany Arifianto is with Physic Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia (email : dhany@ep.its.ac.id). He earned his bachelor degree in Physic Engineering from ITS in 1997. His M.Eng., and Ph.D degrees was received from Tokyo Institute of Technology, Tokyo, Japan in 2002, and 2005, respectively. His current interests include digital signal processing, speech technology and underwater acoustics

Kiyoaki Aikawa is with School of Media Science, Tokyo University of Technology, Japan (email: aik@stf.teu.ac.jp). He received his Ph.D. from University of Tokyo in 1980. He engaged in the NTT Basic Research Laboratory in 1980. He was a visiting scientist of Carnegie Mellon University in 1990. From 1992 to 1995 he was a senior researcher in Advanced Telecommunications Research Laboratories. He stayed in NTT Laboratories from 1996 to 2002. He is a professor of School of Media Science and the director of Media Center at Tokyo University of Technology. He received the Sato Award from the Acoustical Society of Japan. He received the Telecom-System Technology Award from the Electrical Communication Foundation. He is a member of IEICE, ASJ, IPSJ, ASA and IEEE

Mauridhi Hery Purnomo is with the Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya Indonesia (email : hery@ee.its.ac.id). He received bachelor degree from ITS, in 1985. His M.Eng., and Ph.D degrees was received from Osaka City University, Osaka, Japan in 1995, and 1997, respectively. He joined ITS in 1985 and has been a Professor since 2003. His current interests include intelligent system applications, image processing, medical imaging, control and management. He is a Member of IAENG and IEEE.

Submitted 07 January 2016. Published as resubmitted by the authors 23 February 2016.