Mohamad Ali, A. Z., Segaran, K., & Wee Hoe, T. (2015). Effects of Verbal Components in 3D Talking-head on Pronunciation Learning among Non-native Speakers. *Educational Technology & Society, 18* (2), 313–322.

Effects of Verbal Components in 3D Talking-head on Pronunciation Learning among Non-native Speakers

Ahmad Zamzuri Mohamad Ali*, Kogilathah Segaran and Tan Wee Hoe
Faculty of Art, Computing and Creative Industry, Universiti Pendidikan Sultan Idris, Tanjong Malim, 35900, Perak, Malaysia // zamzuri@fskik.upsi.edu.my // m20102001398@siswa.upsi.edu.my // whtan@fskik.upsi.edu.my
*Corresponding author

(Submitted March 14, 2014; Revised June 30, 2014; Accepted July 25, 2014)

ABSTRACT

This study was designed to investigate the benefit of inclusion of various verbal elements in 3D talking-head on pronunciation learning among non-native speakers. In particular, the study examines the effects of three different multimedia presentation strategies in 3D talking-head Mobile-Assisted-Language-Learning (MALL) on the learning achievement of students with low English pronunciation skills. Pre-test and post-test scores were utilized to determine the students' overall performance. 60 college students with low pronunciation skill were involved, in which they were divided into three equal groups of 20 students. Scores obtained were analyzed statistically with one-way analysis of covariance (ANCOVA), with the pre-test scores as covariate. The findings revealed that 3D talking-head with spoken text and on-screen text MALL has significant contribution in retaining the correct pronunciation acquisition in comparison with 3D talking-head with spoken-text alone MALL and spoken text with on-screen text MALL. Therefore, it may be concluded that multiple sources of verbal information to identify and decode language input are advantageous for effective pronunciation learning, specifically among non-native speakers.

Keywords

Animation, Language-learning, Talking-head, MALL, Non-native

Introduction

Discussion of animated characters typically revolves around cartoons, special effects, and movies. However, it appears to be many positive results indicating that animation has an important role in the field of education as well (Balasubramanyam, 2012; McMenemy & Ferguson, 2009; Doyle, 2001). Recently, animations have been included as a part of multimedia learning aid in numerous subject matters, including in language learning (Lin & Tseng, 2012; Kayaoğlu, Dağ Akbaş & Öztürk, 2011). In the past, students who learn English as a second language depended heavily on printed text and audio materials such as cassette tapes, audio CDs and radio broadcast (Xiao & Jones, 1995). Speech and language technology evolved under the assumption that speech was merely auditory event (Massaro, Liu, Chen & Perfetti, 2006). Nevertheless, numerous research findings revealed that, students' understanding of the language are also influenced by speaker's face and accompanying gestures, in addition to the actual sound of the speech (Massaro et al., 2006). Therefore, inclusion of animated character that functions as pedagogical agent in language learning instructional aids seems meaningful.

An appropriate teaching approach is important when learning a second language with the assistance of animated character (Massaro, Bigler, Chen, Perlman & Ouni, 2008). It is very meaningful for overcoming problems faced by students who need to practice clear pronunciations of new words pertaining to a second language, which is different from their first language (Cook, 1996). Notably, numerous instructional approaches were undertaken in establishing English as a second language being acquired globally, which eventually resulted in the introduction of Computer-Assisted-Language-Learning (CALL) and Mobile-Assisted-language-Learning (MALL). In CALL or MALL, the embodied agent, or talking-head animation, becomes the prominent virtual aid for teaching pronunciation, vocabulary, articulation and so forth (Wik & Hjalmarsso, 2009). Generally, talking-head animation acts as a visual character that functions by saying a word or telling a story to the students (Dey, Maddock & Nicolson, 2010). The talking-head character is limited in a setting that the display on the screen only shows the section from the top of its head to the shoulders (Dey et al., 2010). In addition, talking-head animation was developed by combining the principles of linguistics, pedagogy and replete with a good audio system that is capable of helping students to optimize their pronunciation skills (Massaro, 2003).

3D talking-head

Nowadays, there are increasing educational research interests on the animated pedagogical agent in aiding language learning. (Atkinson, 2002; Baylor & Ruy, 2003; Moreno & Mayer, 2000). Pedagogical agents are animated characters designed to function in educational settings to facilitate learning (Shaw, Johnson & Ganeshan, 1999). Some regard the agents as talking-heads which feature speech, facial expressions, and gestures to implement pedagogical strategies (Graesser, Chipman & King, 2008). The talking-heads may be featured as virtual tutors or teachers in language learning applications, supporting various facets of the learning process, including read aloud and practice conversation (Busa, 2008).

Acquiring a new language could be challenging matters to people from all walks of life, especially adult learners (Tan, Lin & Wang, 2013). In a study that evaluated synthetic and natural Mandarin visual speech, Chen and Massaro (2011) revealed that 3D talking-head can improve the learning process of new language acquisition. Massaro (2006a) identified potential of animated virtual tutors in language learning, specifically for individuals who possess language learning disabilities and require extra instruction. Nonetheless, the requirements have hardly been fulfilled due to the shortage of professionals and skilled teachers who can offer individual or personalized attention (Massaro, 2006a). Teachers commonly used books and conventional media in their attempts to fulfill the personalization needs, but the outcomes were not promising because those media are not customized according to individual learners (Massaro, 2006a). According Massaro (2006b), the face is an indispensable component of a human body, especially for conveying a correct message. The combined movements of lips, tongue, and jaws deliver visual information that could increase the lucidity of aural understanding, even in a noisy situation (Massaro, 2006b). Therefore, visual speech information can play an essential role in assisting learners to differentiate words, which are difficult to distinguish when referring to aural information alone (Jesse & Massaro, 2010).

The gestures, facial expressions and the associated emotions made by a speaker would reinforce the speech (Massaro, 1998). In an experiment carried out by Liu, Massaro, Chen, Chan and Perfetti (2007) to examine the use of visual speech in Chinese pronunciation through a web-based learning program revealed that visual speech has significant advantages in improving learners' pronunciation in comparison with audio material. However, issues related to the use of text as verbal support in talking-head application were left unsolved. As justified by Ahmad Zamzuri and Kogilathah (2013), excluding text in pronunciation learning might cause difficulties among learners in identifying the syllable breaks for proper pronunciation, specifically the non-native speakers. Hence, research to identify the ideal solution of verbal support in applications that feature talking-head animation seems important.

On-screen text with spoken text

The theoretical framework of this study was mainly grounded on Mayer's cognitive theory of multimedia learning. According to Mayer's cognitive theory of multimedia learning, human processes information through dual channels, namely the visual channel that processes visually represented materials and verbal channel that processes audio materials (Mayer, 2001). Mayer (2001) believes that human understanding occurs when learners are able to mentally integrate visual and verbal representations of a subject matter as both channels being activated simultaneously. There are three assumptions highlighted in Mayer's cognitive theory of multimedia learning. As discussed earlier, first assumption states that multimedia learning is *dual channel activities*, which are visual-pictorial channel and auditory-verbal channel. For instance, in the 3D talking-head language learning condition, animation will be processed in the visual-pictorial channel and the pronounced word will be processed in the auditory-verbal channel.

The second assumption is *limited capacity*, in which each channel in the human cognitive system has limited capacity in processing information. Therefore, designing pedagogically effective multimedia aids for language learning which grounded on theories, continuously been an important issue (Kim & Gilman, 2008). Taking this into consideration, according to Mayer (2001), students able to acquire knowledge better from animation and narration than from animation, narration and on-screen text. Moreover, when image and text both presented visually, the visual channel will be overloaded, besides the redundancy effects due to dual verbal information (Mayer, 2001). Therefore, if the text is presented in audio form alone, it can be processed by the verbal channel, while the visual channel processes the visual information (Mayer, 2001). In some respects, when on-screen text located together with talking-head, students' demand on the visual modality could lead to split-attention effects (Kim & Gilman, 2008; Craig,

Gholson & Driscoll, 2002). The reason for this effect is that learners might focus their visual attention on animation, rather than on the on-screen text (Kim & Gilman, 2008; Craig, Gholson & Driscoll, 2002).

However, question arises; do these principles apply for 3D talking-head linguistic learning aid? Whereby, text might be helpful in assisting learners in determining the syllable break for correct pronunciation. The study reveals that visual texts can be more preferable under certain conditions, and one of such conditions is instructional pacing. By examining the effects of instructional pacing and text modality on cognitive load and performance, Stiller, Freitag, Zinnbauer and Freitag (2009) have proven that the learner paced visual text instruction was the most efficient condition. Such condition could be applied to applications that feature 3D talking-head with audio and text. Adding to this, learners would pronounce a word accurately if they know the word stress which can be shown in syllable breaks. The correct number of syllables must be created to clearly pronounce the stress pattern of a word (Jones, 2011). Thus, syllable breaks should be placed as text in learning applications.

Finally, the third assumption is *active processing*, in which learners are involved in active processing in the channels, which includes media selection (verbal and visual), organizing the media into the verbal and visual mental model and finally integrating them with preexisting knowledge which results in meaningful schema acquisition. This happens when corresponding verbal and visual representations are in the working memory at the same time (Mayer, 2001). The issue of integrating visual and verbal information in order to retain it in the long term memory is important in talking-head language learning condition. The application must have the capability in assisting learners to integrate the visual form of the 3D talking-head with facial expression and lip movement with the spoken text or spoken text complemented with on-screen text. Likewise, they are able to store the knowledge acquired from the sensory memory (listening and watching the 3D talking- head) and working memory (integrating 3D talking-head action and the pronounced word) in the long term memory and apply it precisely when required.

Facial expression and lip synchronization

Facial expression has been recognized as one of the key concerns to make the language learning efficient and robust when talking-heads are used (Wik & Hjalmarsson, 2009). Research conducted in neuroscience, cognitive science, and psychology justify that emotions which are formed through facial expressions have a significant role in capturing attention, planning, reasoning, learning, memory, and decision making (Picard, 1997). Emotions also play an important role that influences perception, cognition, and creativity (Johnson, Rickel & Lester, 2000; Picard, 1997). In the instructional design for pronunciation learning, gestures and facial expressions have been identified as important features that would facilitate learning (Brown, 2007). In addition, Sime (2006) has determined that facial expression and gestures are two forms of non-verbal communication that can make the classroom learning interesting and motivate learners to engage in the classroom activities actively. Facial expressions particularly could enhance learning capability, such as the ability to recall information (Allen, 2000; Lazaraton, 2004). This would lead to the retention of knowledge captured by learners when learning pronunciation.

In terms of teaching pronunciation, Rodgers (2001) introduced novel methods and methodological prediction, e.g., full-frontal communicativity which is one of the 10 scenarios that engages all aspects of human communicative capacities. Meanwhile, lip synchronization or lip sync is recognized as one of the key features of talking-heads (Lun, n.d.). As English is a language that depends heavily on lip shape, tongue position, teeth position, jaw movement and air flow (Baxter, 1993), the process of learning pronunciation could be practiced by reading the lip sync (Sumby & Pollack, 1954; Benoît & Le Goff, 1998).

Method

Research objective

Based on the literature overview above, the main objective of this research is to identify if a 3D talking-head with spoken text and on-screen text has significant contribution in retaining the correct pronunciation acquisition in comparison with 3D talking-head with spoken text alone and spoken text with on-screen text. The hypothesis derived from the discussion is as follows:

Ha1. 3D Talking-head with spoken text and on-screen text MALL has significant contribution in retaining the correct pronunciation acquisition in comparison with 3D talking-head with spoken text alone MALL and spoken text with on-screen text MALL.

Teaching material

The multimedia presentation used in this application was mainly based on human computer interaction design principles. The interface design of the MALL, which is the medium of choice, will be simple and the usage of heavy graphics is less, since considering the download time and the launching of the application. The application will start with a welcoming screen. The content of the welcoming screen would be the title of the application, and a button to continue to the next screen. In the following screen, there will be ten words option (Table 1) for students to choose from for pronunciation practice.

Table 1. Ten words option

Full word	Word with syllable break	Meaning
Aegis	ae.gis	the protection, backing, or support of a particular person or organization
Archipelago	ar.chi.pe.la.go	a sea or stretch of water having many islands
Cache	cache	a collection of items of the same type stored in a hidden or inaccessible place
Cavalry	ca.val.ry	soldiers who fought on horseback
Foliage	fo.li.age	plant leaves collectively
Mischievous	mis.chie.vous	causing or showing a fondness for causing trouble in a playful way
Pronunciation	pro.nun.ci.a.tion	the way in which a word is pronounced
Ubiquitous	u.bi.qui.tous	present, appearing, or found everywhere
Suite	suite	a set of rooms designated for one person's or family's use or for a particular purpose
Voluptuous	vo.lup.tu.ous	describes a woman who has a soft, curved, and attractive body

There would be instructions on the top of the practice screen to inform the students to select the options. This will be implemented for all three types of prototype. In each prototype there will be a menu which is the list of the difficult to pronounce and commonly mispronounced words. When the students select the particular word, they will be navigated to the text with syllable break and 3D talking-head screen. When the students select the "play" button on this screen, the 3D talking-head will pronounce the word by following the syllable break. Apart from the play button, there are also other navigational buttons on the screen such as, home, next, previous and exit button. The exit button will be displayed in every screen to allow the students to exit whenever they feel like stop using the application (Alessi & Trollip, 2001). Furthermore, application of the exit button suits well for the instruction with drills methodology, which is the pedagogical approach of the study (Alessi & Trollip, 2001). When the students select the "next" button, they will be directed to the full pronunciation practice screen. In this screen, the students can select the "play" button to repeat the 3D talking-head pronouncing the word again. Following that, in the full pronunciation screen, the students can select the "next" button to see the word's meaning or select the "home" button to go back to the menu screen, or select the "exit" button to see the credit screen and leave the application. This design is similar to the remaining two prototypes, which are the 3D talking-head with spoken text alone MALL and spoken text with on-screen text MALL. The only difference would be, in the practice screen of 3D talking-head with spoken text, will not have the text and as for the spoken text and on-screen text MALL, the screen will only contains audio and text alone. Figure 1, Figure 2 and Figure 3 depicts the difference between the three prototype developed.

Essentially, the 3D talking-head character design was developed into non-realistic facial proportion yet with proper modeled features of human to show the movement of the lip. The 3D talking-head animation was only limited to lip syncing and basic facial expressions. The automated-lip-sync technic using the special plug-in in the 3D software was utilized to synchronize the lip movement with audio. Meanwhile, the non-realism appearance has its own advantages such as, expressions can be exaggerated using non-realistic features, much freedom in designing relatively the human face resemblance, creative touch which makes them more preferable compared to real human

face, users won't have high expectations towards its realism performance as they would compare to the real human being face appearance, and cost saving since it does not require special equipment and technology to model and animate (Ruttkay & Noot, 2000).

Figure 1. 3D talking-head with spoken text and on-screen text MALL

Figure 2. 3D talking-head with spoken text alone MALL

Figure 3. spoken text and on-screen text MALL

Procedure

Our experimental study investigated the effects of three different multimedia presentation strategies in 3D talking-head MALL on the learning achievement of students with low English pronunciation skills. Independent variable is the multimedia presentation strategies of 3D talking-head MALL, whilst the dependent variable is the post-test performance. Basically, the research design is three groups pre-test post-test approach, which all the three groups are experimental group. In detail, group one was assigned with the first strategy which is 3D talking-head with spoken

text and on-screen text, group two was assigned with the second strategy which is 3D talking-head with spoken text alone and group three was assigned with the third strategy which is spoken text with on-screen text. The study was conducted separately for all the groups in the controlled lab environment. Brief explanations were given about the features of the app to ensure the smoothness of students' exploration throughout the learning process. Prior to the study, the students were required to undergo the pre-test to determine their pre-existing pronunciation skills of the selected words. Students in each group were given 15 minutes of learning time which was deemed suitable based on the literature and observation done during the development phase. Each student was provided with one 7 inch tablet for the mobile learning purpose. Post-test was conducted immediately upon completion of the learning process. The pre-test and post-test are identical, and involve the same words. The same instructor conducted the study session of all the groups. No verbal guide from the instructor has occurred throughout the learning process.

The research participants were 60 college students, whose age ranged from 18 to 20, enrolled in the Diploma in Multimedia Application program and undergoing similar Essential English Communication course. 60 students with pronunciation difficulty were identified through the oral test done prior to the study, which was conducted by a specifically designated linguistic tutor. These sixty students were then divided into three groups with 20 students each, with balanced proportions of gender in each group. Random stratified sampling technique was used to group them equally.

Test instruments

Pre-tests and post-tests were used on the three groups that undergo the three different multimedia presentation strategies respectively. Pre-test and post-test were oral test that only measure the pronunciation performance of the selected words. The tests required students to read the word list given loudly within 15 seconds. Since it consumed time to assess all the three groups immediately, which contains sixty students in total, the test was videotaped and later given to the examiners to assess. By videotaping, also ensures the examiners have ample time to evaluate the students' performance precisely, whereby, they could replay the video accordingly. Three English lecturers who are expert in linguistic were appointed to assess the students' performance to ensure the reliability of the result. The examiners assessed the students' performance based on the oral test score sheet and schema which was adapted from the existing National Education Certificate oral test score sheet. The score sheet and schema was also validated by two English linguistic lecturers. The average grade from the three examiners was used as the final score of the students.

Results

A one-way ANCOVA was conducted in order to examine whether 3D talking-head with spoken text and on-screen text MALL has significant contribution in retaining the correct pronunciation acquisition in comparison with 3D talking- head with spoken text alone MALL and spoken text with on-screen text MALL. The independent variable was the 3D talking-head with three different presentation strategies and the dependent variable was the post-test scores administered following completion of the learning process. Scores of the pre-test administered prior to the commencement of the learning process were used as the covariate.

Preliminary checks were conducted to ensure that there was no violation of the assumptions of normality, linearity, homogeneity of variances and homogeneity of regression slopes. From one-way ANCOVA test, Levene's test for homogeneity of variances was not significant ($p > 0.05$), and therefore the data do not violate the assumption of equality of error variances. After adjusting for pre-test scores, there was significant difference on post-test achievement of 3D talking-head with spoken text and on-screen text MALL in comparison with 3D talking-head with spoken text alone MALL and spoken text with on-screen text MALL application, $F (2,56) = 5.65$, $p < 0.05$, *partial eta squared = 0.17*, with large effect size according Cohen's 1998 guidelines (Pallant, 2007) as shown in Table 2. The adjusted mean scores indicate that students in the 3D talking-head with spoken text and on-screen text MALL application obtained a better mean score ($M = 40.70$, $SE = 0.96$) than students in the 3D talking-head with spoken text alone MALL application ($M = 36.73$, $SE = 0.97$) and spoken text with on-screen MALL application ($M = 36.77$, $SE = 0.96$) as shown in Table 3. This clearly suggests that 3D talking-head with spoken text and on-screen text MALL application has significant contribution in retaining the correct pronunciation acquisition in comparison

with 3D talking-head with spoken text alone MALL and spoken text with on-screen text MALL. In summary, the hypothesis derived from the literature overview done is accepted.

Table 2. Summary of Tests between-subjects effects

Source	df	Ms	F	Sig.	Partial Eta Squared
Pre	1	156.50	8.51	.01	.13
Group	2	103.86	5.65	.01	.17
Error	56	18.39			

Note. R squared .28 (Adjusted R squared = .24).

Table 3. Summary of Descriptive Statistic

Mode	Unadjusted			Adjusted	
	N	M	SD	M	SE
3D Talking-head with audio and text MALL	20	40.85	4.30	40.70	.96
3D talking-head with audio alone MALL	20	36.32	5.29	36.73	.97
Audio and text alone MALL	20	37.03	4.00	36.77	.96

Note. Covariates appearing in the model are evaluated at the following values: Pre = 22.91.

Discussions and conclusions

Animation plays potential role in improving human learning process, particularly in promoting profound understanding of the subject matter (Mayer & Moreno, 2002). In line with this, animated pedagogical agent such as 3D talking-head or virtual language tutor, which simulates real human like tutor in the computer, has been created for aiding language learning. The use of 3D talking-head also subsequently improved the learning process of a new language or second language (Chen & Massaro, 2011). Even so, there are dozens of audio and text based talking dictionary available online, the inclusion of animated pedagogical agent seems beneficial for pronunciation learning; specifically for non-native in learning difficult to pronounce English words. English is a language which depends upon; airflow, lip shape, tongue position, teeth position and jaw movement (Baxter, 1993), where the process can be practiced through watching the lip syncing activities (Sumby & Pollack, 1954; Benoît & Le Goff, 1998). Besides that, facial expression is also among the concern to make the language learning more efficient and robust (Wik & Hjalmarsson, 2009). These features, which could be incorporated in the 3D talking-head, might contribute to effective pronunciation learning, specifically among non-native speakers. This research finding further affirms the argument. Whereby, the finding showed that students in the 3D talking-head with spoken text and on-screen text MALL learning condition, outperformed students in the 3D talking-head with spoken text alone MALL, and spoken text with on-screen text MALL, which appears to be the traditional language learning setting.

The belief that speech and language science and technology evolved under the assumption that speech was a solely auditory event seems is not totally true (Massaro, Liu, Chen & Perfetti, 2006). Multiple sources of information to identify and interpret the language input, undoubtedly, promote better pronunciation learning. It could be concluded that 3D talking-head with spoken text and on-screen text MALL, which combines visual information in the form of 3D talking-head and verbal information in the form of spoken text audio and on-screen text display promote better pronunciation learning, instead of one of the element is removed (i.e., talking head or on-screen text). The value of the 3D talking-head is mostly on depicting the lip syncing activities. Whereby, visual information from the movements of the lips enhances lucidity of the audio understanding, specifically in a noisy environment (Massaro, 2006b). Nonetheless, the 3D talking-head character in this study only displays the lip sync in comparison to Massaro's series of studies that also display the tongue and palate; which were shown by making the skin transparent. Further study comparing these two strategies seems interesting.

Beside the lip sync, facial expression of the 3D talking-head character may also possibly influence the outcome of the study. Emotions which are formed through facial expression have a significant role in capturing attention and facilitate learning (Picard, 1997), which might contribute in enhancing recall ability that would lead to better retention of the knowledge captured (Allen, 2000, Lazaraton, 2004). From the observation throughout the study,

students in the 3D talking-head with spoken text and on-screen text MALL, seemed more excited in exploring the application in comparison with spoken text with on-screen text MALL. This might have been due to the 3D talking-head as pedagogical agents have potential in increasing the motivation and confidence level of the students in learning, since it acts as a virtual tutor. Therefore, further study in looking on the motivation and confidence level aspects and its effects on pronunciation learning should answer this belief.

Besides observing the talking-head in visual terms, the importance of listening in language learning and its role in helping the learner to overcome the barrier of learning a new language has long been acknowledged. Likewise, in this study, spoken text plays an important role in teaching the students to pronounce. When students are exposed to visual information and verbal together, they are able to integrate them with the pre-existing knowledge which results in meaningful schema acquisition in comparison if the information is in verbal condition alone (i.e., spoken text and on-screen text). This happens when corresponding visual and verbal representations are in the working memory at the same time (Mayer, 2001). However, in learning pronunciation apart from 3D talking-head as visual information, visual text or on-screen text also plays an important role. To pronounce the stress pattern of a word clearly, the right number of syllables needs to be produced (Jones, 2011). Therefore, in order to show this, the syllable break was placed as text in the application. From the observation and the result obtained, it was proven that the 3D talking-head with on-screen text has shown significant difference in the interest and score achieved by the students.

On the other hand, the possible redundancy and split attention effects does not manifest any noticeable consequences on the learning performance of students in the 3D talking-head with spoken text and on-screen text condition. This might due to the repetition process which facilitates the learners in developing a more accurate mental model in the cognitive structure for meaningful schema acquisition; as also pointed by Stiller et al. (2009). Even so, the repetition occurs in all the three learning conditions, students in the 3D talking-head with spoken text and on-screen text condition seems benefited more. Multiple sources of verbal information that complemented the visual information might trigger more sufficient active processing in the memory structure for effective pronunciation learning. Possibly, similar learning outcome could be achieved if the learning time was added for the remaining two conditions. However, increased time does not necessarily support learning. Increased learning time may be attributed to other consequences which might impede learning, such as getting bored, decreasing the learning interest, demotivated, etc.

Even though the study revealed that the 3D talking-head with spoken text and on-screen text MALL plays an important role in helping learners distinguish how to pronounce difficult words, the question arises whether the same retention rate can be achieved when examining words in context comparing words in isolation? In this study, students only practice to pronounce the words in isolation. Thus, this study suggests a potential future study should be in comparing words in context besides words in isolation in retaining the correct pronunciation acquisition. Surprisingly, students in the 3D talking-head with spoken text alone MALL, and spoken text with on-screen text MALL learning conditions obtained almost similar mean score. As has been discussed, both of these applications are actually lacking in including one of the elements, which would be crucial for effective pronunciation learning.

Acknowledgements

The authors wish to acknowledge the support of the Ministry of Higher Education and Universiti Pendidikan Sultan Idris, who awarded a RAGS research grant for this study.

References

Ahmad Zamzuri, M. A, & Kogilathah, S. (2013). 3D Talking-head mobile app: A conceptual framework for English pronunciation learning among non-native speakers. *English Language Teaching, 6*(8), 66-76.

Alessi S., & Trollip S. (2001). *Multimedia for Learning: Methods and Development* (3rd ed.). Boston, MA: Allyn & Bacon.

Allen, L. Q. (2000). Nonverbal accommodations in foreign language teacher talk. *Applied Language Learning, 11*(1), 155-176.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94* (2), 416-427.

Balasubramanyam, V. (2012). Animations in medical education. *Medical Journal of Dr. D.Y. Patil University, 5*(1), 22.

Baxter, B. (1993). *Studying tips: Speaking tips*. Retrieved from http://www.musicalenglishlessons.org/tips-speaking.htm

Baylor, A., & Ryu, J. (2003). Does the presence of image and animation enhance pedagogical agent persona? *Journal of Educational Computing Research, 28*(4), 373-395.

Benoît , C., & Le Goff, B. (1998). Audio–visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP. *Speech Communication, 26*(1-2), 117–129.

Brown, A. (2007). The use of nonverbal features in teaching phonetics. *Proceedings of the Phonetics Teaching & Learning Conference.* Retrieved from http://www.phon.ucl.ac.uk/ptlc/proceedings/ptlcpaper_24e.pdf

Busa, M. G. (2008). New perspectives in teaching pronunciation. Retrieved from http://www.openstarts.units.it/dspace/bitstream/10077/2850/1/bus%C3%A0.pdf

Chen, T. H., & Massaro, D. W. (2011). Evaluation of synthetic and natural Mandarin visual speech: Initial consonants, single vowels, and syllables. *Speech Communication, 53* (7), 955-972.

Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology, 94*(2), 428-434.

Cook, V. (1996). *Second language learning and language teaching* (2nd ed). London, UK: Edward Arnold.

Dey, P., Maddock, S., & Nicolson, R. (2010). Evaluation of a viseme-driven talking head. *Proceedings of the EG UK Theory and Practice of Computer Graphics* (pp. 118-116). Retrieved from http://staffwww.dcs.shef.ac.uk/people/S.Maddock/publications/DeyEtal2010_TPCG.pdf

Doyle, A. (2001). Web animation: Learning in motion. *Tech & Learning, 22*(2), 30-42.

Graesser, A. C., Chipman, P., & King, B. G. (2008). Computer-mediated technologies. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 211-224). New York, NY: Lawrence Erlbaum Associates.

Jesse, A., & Massaro, D.W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics, 72*(1), 209-225. doi: 10.3758/APP.72.1.209

Johnson, W. L., Rickel, J. W. & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*(1), 47-78.

Jones, D. (2011). *Cambridge English pronouncing dictionary* (18th ed.). Cambridge, UK: Cambridge University Press.

Kayaoğlu, M. N., Dağ Akbaş, R., & Öztürk, Z. (2011). A small scale experimental study: Using animations to learn vocabulary. *Turkish Online Journal of Educational Technology, 10*(2), 24-30.

Kim, D., & Gilman, D.A. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Educational Technology & Society, 11*(3), 114-126.

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A micro-analytic inquiry. *Language Learning, 54*(1), 79-117.

Lin, C. & Tseng, Y. (2012). Videos and animations for vocabulary learning: A study on difficult words. *Turkish Online Journal of Educational Technology, 11*(4), 346-355.

Liu, Y., Massaro, D. W., Chen, T. H., Chan, D., & Perfetti, C. A. (2007). *Using visual speech for training Chinese pronunciation: An in-vivo experiment*. Retrieved from http://www.cs.cmu.edu/~max/mainpage_files/files/SLaTE07_Liu_Training_Chinese_Pronunciation.pdf

Lun, E. V. (n.d.). *Talking head*. Retrieved from http://www.chatbots.org/talking_head/

Massaro, D. W., Bigler, S., Chen, T., Perlman, M., & Ouni, S. (2008). Pronunciation training: The role of eye and ear. *Proceedings of INTERSPEECH 2008* (pp. 2623-2626). Red Hock, NY: Curran Associates.

Massaro, D. W., Liu, Y., Chen, T. H., & Perfetti, C. (2006). A multilingual embodied conversational agent for tutoring speech and language learning. *Proceedings of INTERSPEECH 2006-ICSLP* (pp. 825-828). Retrived from http://www.learnlab.org/uploads/mypslc/publications/massaro-multilingualembodied.pdf

Massaro, D. W. (2006a). Embodied agents in language learning for children with language challenges. In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.), *Proceedings of the 10th International Conference on Computers Helping People with Special Needs, ICCHP 2006*,(pp. 809–816). Berlin, Germany: Springer.

Massaro, D.W. (2006b). The psychology and technology of talking heads: Applications in Language Learning. In O. Bernsen, L. Dybkjaer, & J. van Kuppevelt (Eds.), *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems* (pp.183-214). Dordrecht, The Netherlands: Kluwer Academic Publishers

Massaro, D.W. (2003). A computer animated tutor for spoken and written languange learning. *Proceedings of 5th International Confrences on Multimodal Interfaces*. (pp. 172-175). NewYork, NY: ACM. doi: 10.1145/958432.958466

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Mayer, R. E. (2001). *Multimedia Learning.* Cambridge, UK: Cambridge University Press.

Mayer, R. E., & Moreno, R. (2002). Animation as an aid to multimedia learning. *Educational Psychology Review, 14*(1), 87-99.

McMenemy, K., & Ferguson, S. (2009). Enhancing the teaching of professional practice and key skills in engineering through the use of computer animation. *International Journal of Electrical Engineering Education, 46*(2), 164-174.

Moreno, R., & Mayer, R. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology and Technology, 92*(4), 724-733.

Pallant, J. (2007). *SPSS survival manual: A step-by-step guide to data analysis using SPSS for Windows (Version 15)* (3rd ed.). Crows Nest, Australia: Allen & Unwin.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Rodgers, T. (2001). Language teaching methodology. Retrieved from http://www.cal.org/resources/Digest/rodgers.html

Ruttkay, Z., & Noot, H. (2000). Animated CharToon faces. *Proceedings of 1st International Symposium on Non Photorealistic Animation and Rendering, Annecy, France,* (pp. 91-100). doi: 10.1145/340916.340928

Shaw, E., Johnson, W. L., & Ganeshan, R. (1999). Pedagogical agents on the web. *Proceedings of the Third International Conference on Autonomous Agents* (pp. 283-290). Retrieved from http://www.isi.edu/~shaw/publications/agents99.htm

Sime, D. (2006). What do learners make of teachers' gestures in the language classroom? *International Review of Applied Linguistics in Language Teaching, 44*(2), 211-230.

Stiller, K. D., Freitag, A., Zinnbauer, P., & Freitag, C. (2009). How pacing of multimedia instructions can influence modality effects: A case of superiority of visual texts. *Australasian Journal of Educational Technology, 25*(2), 184-203.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212–215.

Tan, W. H., Lin, C. Y., & Wang, Y. (2013). Mandarin Communication Learning app: A proof-of-concept prototype of contextual learning. *Journal of Research, Policy & Practice of Teachers & Teacher Education, 3*(2), 38-48.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication, 51*(10), 1024-1037.

Xiao, X., & Jones, M. G. (1995, October). *Computer animation for EFL learning environments*. Paper presented at the Annual Conference of The International Visual Literacy Association, Chicago, IL.