Accuracy Performance and Potentiality of Real-Time Avatar Lip Sync Animation in Different Languages

N. H. Loh, S.S. Shaharuddin

Abstract: ***With the fast growth in computing power nowadays, the qualities of animation enable an extra layer of visually convincing realism. In lip sync animation, creation of realistic lip movement is arduous in getting the lip shape and position to synchronize with the speech sounds. To note, spending hours in manually generating every single lip movement can be a long and challenging task. Consequently, a comprehensive analysis on viseme based multiple phonemes in English, Bahasa Melayu and Mandarin was carried out, to develop an accurate and potential platform for real time talking avatar in multiple languages. The accuracy performance between human and avatar in real time were compared and evaluated. The findings revealed successful utilization of real time synchronization to drive the synthetic 3D avatars based on live speech input for multiple languages, with satisfied accurate lip motion result. This paper provides useful knowledge for multilingual solution which accurately predicts mouth movement on real human face, when a person is speaking and directs to lip sync process. It contributes to live performances and valuable in open-ended field with tons of potential, such as animation production industry, entertainment, gaming, digital marketing and media education.***

Keywords: Avatar; Human Lip Shape; Lip Sync Animation; Real Time; Speech Recognition.

I. INTRODUCTION

Real-time lip sync animation is an approach to perform the talking of a virtual computer-generated character known asavatar, which synchronizes an accurate lip movement and sound in live animation. To explained, lip synching is often a part of the post-production phase in the making of animation films. However, drawings, clay puppets and computer meshes do not talk; so, when the synthesised characters are required to say something, their dialogues have to be recorded and analysed first before animate them to speak. Therefore, lip synchronisation or 'lip-sync' is the technique of moving a mouth of an animated character in such a way that it appears to speak in synchronism with the sound track. Likewise, real time lip sync is a technique driven by human voice directly to generate an avatar to talk on the screen.

In this context, determination of human's lip pattern and movement showed the significance in generating natural speech. [1] emphasized that it is a necessary step in the process of mapping lip movements to the speech sound.

Revised Manuscript Received on April 15, 2019.

N. H. Loh, Faculty of Applied and Creative Arts, Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak Malaysia

S.S. Shaharuddin, Faculty of Applied and Creative Arts, Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak Malaysia

However, creating an accurate lip sync animation would be significantly more difficult especially in setting key frame value, as shown in Figure 1. In fact, it is particularly challenging in mapping the lip movements and sounds to be synchronized [2]. It is a time consuming process [3] especially in doing multilingual animation. To elaborate, the process is done manually through adjusting frame by frame that often needs several passes of fine tuning to match the sound [4]. As a result, it can be clearly seen that most of the animation films will choose not to redo the lip sync process when republish the animation with second language. The difficulty of the process causes heavy workload, time consuming and costly.

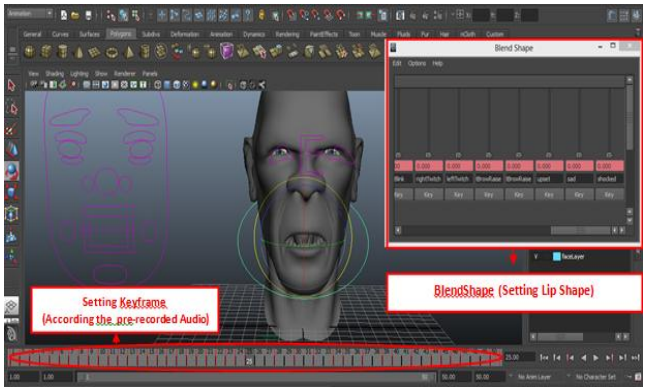


Fig. 1 Screen shot of setting keyframe to create lip sync animation in Autodesk MAYA

Therefore, the real time approach is needed to solve the difficulty of ordinary lip sync techniques and ensure realism in lip sync animation. In order to make character animation believable, correct lip shape corresponds to the sound is essential [5]. This paper provided automated digital speech model of viseme classification mapping to match the key phoneme sounds for English, *Bahasa Melayu* and Mandarin languages. Viseme is short for visible phoneme and refers to the shape of the mouth at the apex of a given phoneme [6]. Different categories of lip shapes in producing different phonemes sound had been analyzed very specifically using viseme categories. The approach reduces the difficulties of the ordinary lip sync technique which involves figuring out the speech timings and animating mouth positions manually to cohere with the sound. As for the making of real time animation which to be broadcasted with different language dialogues, it shortens the duration of production process and ensures an accurate outcome of lip sync.

Moreover, the creation of lip sync can be integrated into the existing animation production pipeline easily in the synthesis phase, asthe speaking virtual character is driven by audio signal in real time. Hence, it can be used in many applications for animator artists, such as multilingual animation reproduction, mass and pre-animation production, avatar speech and lip animation for game pipelines.

II. LITERATURE REVIEW

Concept of Real Time Lip Sync Animation

Real-time lip sync animation is also known as automatic lip sync or digital speech. A symbolic view of real time lip sync animation is shown in Figure 2. To quote [1], lip synchronization or 'lip-sync' is the technique of moving a mouth of an animated character in such a way that it appears to speak in synchronism with the sound track. However, real-time animation is the management of time and the program parameter is considered as manageable by the system [7]. As stated by [8], automatic lip sync is a technology that allows the computer to identify and understand the words spoken by a person by using a communication device.

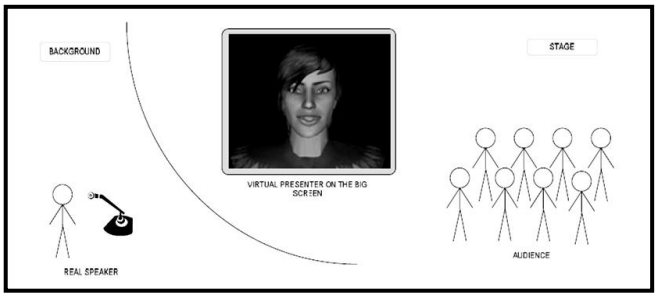


Fig. 2 A symbolic view of Real Time Lip Sync Animation by [8]

Figure 2 demonstrates a scene where a speaker is talking through the microphone while the lip sync is performed at the same time accordingly, the animation is projected on the big screen and interacts with the audiences. The virtual animated character is driven by speech in real time. To achieve real-time lip sync performance, the simulation of avatar mouth's movement must always be synchronized with the physical time and speech sound to produce an output precisely within strict constraints, regardless of computational platform.

Human Lip Shapes and Position

Position of the mouth is regarded as functions of the phonemes. Different intonation of speech in vowels and consonants will make up the difference of mouth position or lip shape. [9]stated that human speech portrays the actual movement of the mouth as it makes the sounds of the dialogue. The mouth shape is changed when we speak, and each sound has a distinct look that is recognizable. Supported by [10], he presented that the reason why mouth shapes are so important during speech is because they are highly noticeable to an audience.

Based on the report of [11], four phonetically human behaviours on the speech provide efficient visual cues for audio-visual speech recognition and bring robustness to the automatic speech recognition system. These four models are the opening jaw, the lips rounding, the lips closure and the lips rising, as shown in Figure 3.

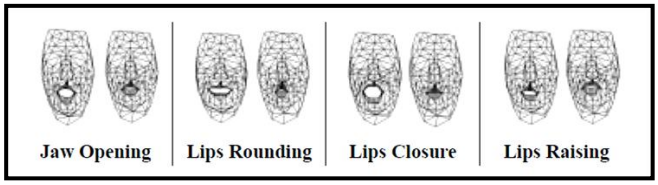


Fig. 3 Phonetically Human Behaviour on the Speech Model by [11]

In addition, referring [12], human lip shape is described normally as being either rounded or unrounded, a slightly more detailed analysis would be rounded, neutral and spread. They explained the phonetic in English language as the following. /i/ is pronounced as a spread lip which is also considered as unrounded lip, with /ʌ/ having a neutral lip shape, and /ɔ/ having a rounded lip shape. Rounding the lips means pulling their corners towards the middle so that the mouth forms an O. Lips can be fully spread, fully rounded, or somewhere in between which is also known as neutral lip shape. To distinguish, unrounded lip shape shows that the corners are not drawn in at all, and rounded lip shows the corners being drawn in to a moderate degree [13]. Likewise, [14] also has the same opinion. He described that the human lips can have many different shapes and positions, but can be categorised as three possibilities only. The possibilities are rounded lip, spread lip and neutral lip when speaking. The illustrations of human lip shapes are shown in Figure 4 and Figure 5.

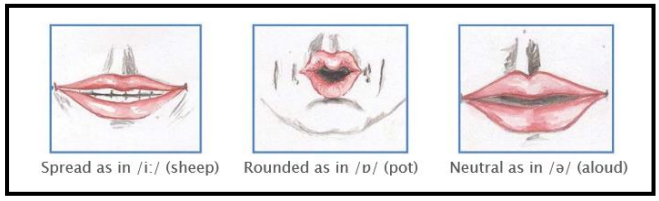


Fig. 4 Illustration of Human Lip Shapes by [15]

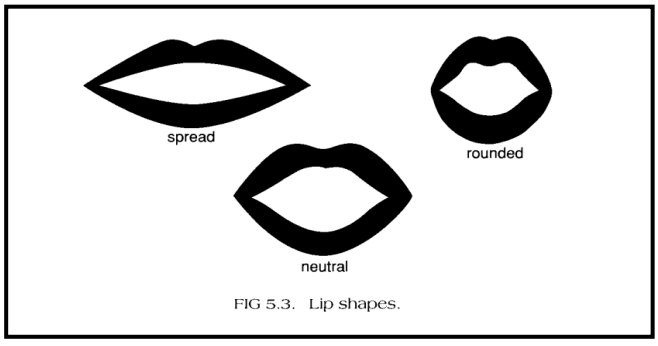


Fig. 5 Illustration of Human Lip Shapes by [16]

Virtual human's lip shape is helpful to viewers by providing visual information of speech sounds. It is analysed in visual cues and categorised in viseme analysis, where each speech sound must be assigned a corresponding shape as viseme approach, this is significant to create an accurate performance of automated lip sync animation.

Phonemes

Synchronizing animation to the spoken word requires an understanding of phonetics, the study of spoken language and speech sounds [17]. To understand speech sound, a common approach is to break it down into a simple set of constituent, atomic sound segments. The most commonly used segment is called phonemes [18].In English, there are 44 different speech sounds [19]. They can be divided into two major categories, vowels and consonants. A vowel sound is one in which the air flow is unobstructed when the sound is made. In contrast, a consonant sound is one in which the air flow is cut off either partially or completely when the sound is produced. The chart of 44 English sounds is shown in Figure 6.

The 44 Sounds of English

Consonant Sounds		Vowel Sounds	
1. /b/	(bat)	23. /th/	(the)
2. /d/	(dog)	24. /hw/	(wheel)
3. /f/	(fan)	25. /ng/	(ring)
4. /g/	(gate)	26. /ā/	(cake)
5. /h/	(hat)	27. /ē/	(feet)
6. /j/	(jump)	28. /ī/	(bike)
7. /k/	(kite)	29. /ō/	(boat)
8. /l/	(leaf)	30. /yo͞o/	(cube)
9. /m/	(mop)	31. /a/	(cat)
10. /n/	(nest)	32. /e/	(bed)
11. /p/	(pig)	33. /i/	(fish)
12. /r/	(rock)	34. /o/	(lock)
13. /s/	(sun)	35. /u/	(duck)
14. /t/	(top)	36. /ə/	(alarm)
15. /v/	(vase)	37. /â/	(chair)
16. /w/	(wagon)	38. /û/	(bird)
17. /y/	(yo-yo)	39. /ä/	(car)
18. /z/	(zebra)	40. /ô/	(ball)
19. /ch/	cheese)	41. /oi/	(boy)
20. /sh/	(shark)	42. /ou/	(house)
21. /zh/	(treasure)	43. /o͝o/	(moon)
22. /th/	(thumb)	44. /o͝o/	(book)

Fig. 6 The Chart of 44 English Sounds by [19]

Besides, *Bahasa Melayu* or Malay language has 6 vowels and 26 consonants [20]. However, consonants in *Bahasa Melayu* are divided into two categories - native consonant and non-native consonant. According to [21], 18 native consonants are primarily derived from the Malay language, and 8 non-native consonants occured in borrowed words, principally from Arabic and English. The phoneme charts of *Bahasa Melayu* are shown as Figure 7, Figure 8 and Figure 9:

Huruf	Fonem	Penggunaan Fonem Vokal		
		Awal	Tengah	Akhir
a	/a/	/awak/ /aku/	/cakap/ /gagap/	/mana/ /pula/
e	/e/	/enak/ /ekor/	/leher/ telen	/tauge/ /ole-ole/
i	/i/	/ikat/ /iman/	/pilih/ /gigih/	/tuli/ /tari/
o	/o/	/oleh/ /oran/	/bohon/ /tolon/	//pidato/ /solo/
u	/u/	/ubat/ /ular/	/bulat/ /pulut/	/sudu/ /palu/
e	/ə/	/emak/ /empat/	/penat/ /tepat/	/nasionalisme/ /mekanisme/

Fig. 7 The Chart of 6 Vowels Sounds in *Bahasa Melayu* by [22]

Huruf	Fonem	Penggunaan Fonem Konsonan		
		Awal	Tengah	Akhir
b	/b/	Batuk	lambat	sebab
c	/č/	Cantik	benci	Mac
d	/d/	Datuk	ladang	abad
g	/g/	Gagal	ligat	monolog
h	/h/	Hapak	sahabat	telah
j	/ĵ/	Jamu	sejuk	garaj
k	/k/	Kilas	sakit	tapak
l	/l/	Lama	Melayu	kebal
m	/m/	Manis	taman	sekam
n	/n/	Nanti	tanda	makin
ng	/ŋ/	Nganga	tangkup	belang
ny	/ɲ/	Nyanyuk	senyap	-
p	/p/	Panas	papan	kudup
r	/r/	Rasa	arus	luhur
s	/s/	Senak	ansur	deras
t	/t/	Tamat	pantun	sukat
w	/w/	Warisan	sawah	takraw
y	/j/	Yuran	wayang	-

Fig. 8 The Chart of 18 Native Consonants Sounds in Bahasa Melayu by [22]

Huruf	Fonem	Contoh Penggunaan
f	/f/	filem
v	/v/	variasi
th	/ɵ/	mithal (misal)
dh	/ð/	dharab (darab)
z	/z/	zaman
sy	/š/	mesyuarat
kh	/x/	khat
gh	/ɣ/	ghaib

Fig. 9 The Chart of 8 Non-Native Consonants Sounds in *Bahasa Melayu* [23]

Revealed by [24], the approach of Malay word is different from English. English word pronunciation depends on a sequence of phonemes. Contrarily, *Bahasa Melayu's* word pronunciation is comprised of Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) combinations. Knowledge on the phoneme sounds presented in *Bahasa Melayu* is crucial for standardizing and transforming multiple languages from more to one mapping. It is used to classify into the viseme to make an accurate lip sync animation possible for speaking in *Bahasa Melayu*. Moreover, in standard Mandarin, there are six initial vowel sounds [25] and 21 initial consonant sounds [26], as shown in Figure 10 and Figure 11.

Pinyin	IPA	Pinyin	IPA	Pinyin	IPA
a	[ɑ]	e	[ɤ]	u	[u]
o	[o]	i	[i]	ü	[y]

Fig. 10 Initial Vowels Sound of Standard Mandarin in Pinyin and IPA by [27]

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IPA
p ph m f t th n l k kh h tɕ tɕh ɕ tʂ tʂh ʂ ɹ ts tsh s

Hanyu Pinyin 漢語拼音
b p m f d t n l g k h j q x zh ch sh r z c s

Tongyong Pinyin 通用拼音
b p m f d t n l g k h ji ci si jh ch sh r z c s

Jhuyin Fuhao 注音符號
ㄅ ㄆ ㄇ ㄈ ㄉ ㄊ ㄋ ㄌ ㄍ ㄎ ㄏ ㄑ ㄒ ㄓ ㄔ ㄕ ㄖ ㄗ ㄘ ㄙ

Table 2. Mandarin consonants represented by IPA, HP, TYP, and Jhuyin Fuhao.

Fig. 11 Initial Consonants Sound of Standard Mandarin in Pinyin and IPA by [28]

Pinyins are regarded as the 'alphabet' of Mandarin. According to [29], Pinyin is currently the most popular tool for encoding Mandarin sounds and is regularly used to transcribe western languages. It is also used for mapping to the International Phonetic Alphabet (IPA) which aims to describe Chinese sound to make it convenient.

Apart from that, International Phonetic Alphabet (IPA) is a set of symbols designed to be representing the speech sounds of languages of the world [30].

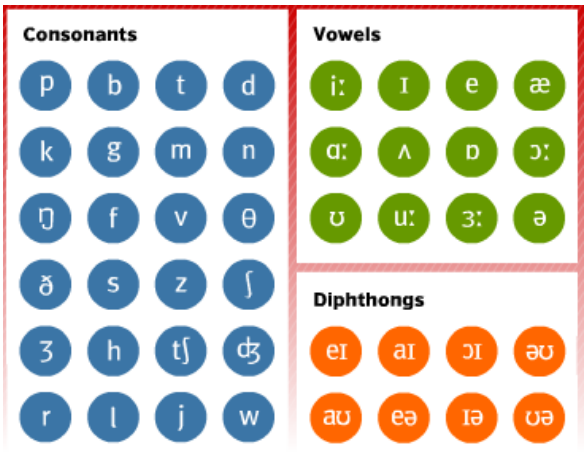


Fig. 12 The International Phonetic Alphabet (IPA) Phonemic Chart [31]

As in Figure 12, the chart shows there are 24 consonant sounds symbols and 12 vowels sounds symbols in IPA. In order to build real-time lip sync with audio output generated by a speech synthesis module, the speech synthesis system is expected to generate phoneme information which will be mapped to appropriated viseme. The speech sounds of phonemes from three languages of English, *Bahasa Melayu* and Mandarin are communicated and represented using the standard symbols of International Phonetic Alphabet (IPA). As a whole, this paper analyses phonemes containing of these three languages and classifies them into lip shape categories. This is done so as to synchronize the lip shapes of a computer-generated face with speeches in multiple languages.

Viseme

[32]defined a viseme as a basic visual unit of speech. It is also an approach of basic animation parameters to estimate visual similarities between different phonemes [33]. Similar visual appearance of mouth positions for different phonemes is collected into classes of visual phonemes. These so called visemes are then used as animation parameters to get key-frames for all possible mouth positions.[18]also stated that visemes can be referred as the corresponding facial poses that produce the phoneme sounds. Similar visemes can be combined into a single unique viseme and the resulting set of facial poses (Figure 13) can be used as blend shapes for a simple type of lip sync animation.

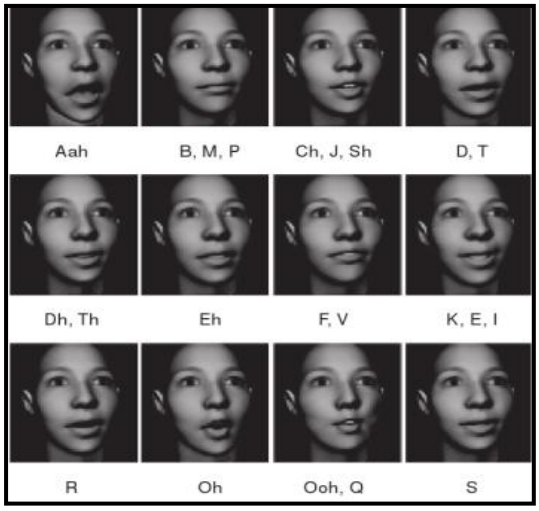


Fig. 13 Viseme Set by [18]

As claimed by [34], a viseme is any of several speech sounds which look the same, for example when lip reading. He is also the one who introduced the word “viseme” which is a compound word of "visual" and "phoneme". The viseme-based human speech technique is applied to this study. It is then developed for performing a virtual computer-generated avatar to talk automatically by using lip synchronization platform. So, viseme approach is used to classify three languages of English, *Bahasa Melayu* and Mandarin phoneme sounds. It is essential to create an accurate virtual lip movement to synchronize with the human speech in three languages of English, *Bahasa Melayu* and Mandarin in the study.

III. METHODOLOGY AND ANALYSIS

Video Content Analysis

Video Content Analysis (VCA) have carried out with professional teaching in English pronunciation, to identify accurate human lip shape and mouth positions in each 12 vowels and 24 consonants sounds, based on English International Phonetic Alphabet (IPA). The videos involved man and woman who teach the correct mouth shape and position to pronounce the sound. It is an essential fundamental process in order to identify the correct human mouth position for each single phoneme sound, and can be used for analyzing the realistic mouth movement when a human was speaking a word or a sentence. All the videos were converted into pictures according to the phoneme sound by using Adobe Premiere Pro software.

Fig. 14 Screen Capture of Lip Shape and Mouth Positions from the Existing Videos in Teaching English Pronunciation

All the pictures captured from the videos were measured and organized into viseme categories. It is an essential process in order to classify the correct lip shapes and position into the technicalities of lip sync animation. It helps to develop the automated digital speech system by increasing speech recognition's accuracy for both human and computer system. Speech recognition was done by visually recognizing the shape of the speakers' lip movements, and make sure the accurate lip shape parameter was performed.

Next, human lip shapes for each phoneme sound were measured and determined the viseme classification, so as to simulate accurate lip sync animation. Based on the study of human lip shapes for each phoneme sound, it found out that accurate motion of lip sync requires precise data to classify lip shape and its position into viseme categories that collected from the mean values. Therefore, this stage has broken down into four important steps to analyze the details of measuring human lip shapes for each phoneme sound, according to 12 vowel sounds and 24 consonant sounds, based on English International Phonetic Alphabet (IPA). as shown in Figure 15.

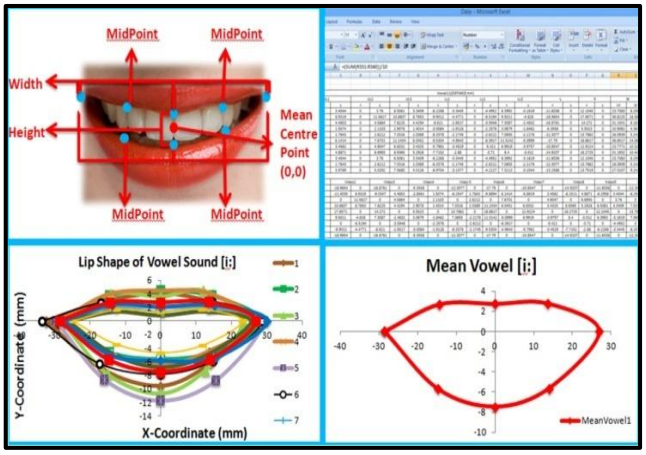


Fig. 15 Progress Measurement on Human Lip Shapes for Each Phonemes Sound

In first phase, all the lip shapes and mouth positions for each sound were measured by using active shape models (ASM) [35]. Human lip shape contained complex articulated deformation shapes that had various probabilities of movement and position. Therefore, ASM model was required in this stage to take measurement. It is a statistical model of shape and appearance that uses a set of labelled points or landmarks to examine the statistics of the lip shapes' coordinates. The lips' feature point was determined, the characteristic pattern of a shape class was described by the mean shape vector and resulting the shape parameter.

Secondly, the mean value of human lip shapes are analyzed and studied as shown in Figure 16, by comparing the results that obtained from the analyzed graphs done in previous phase.

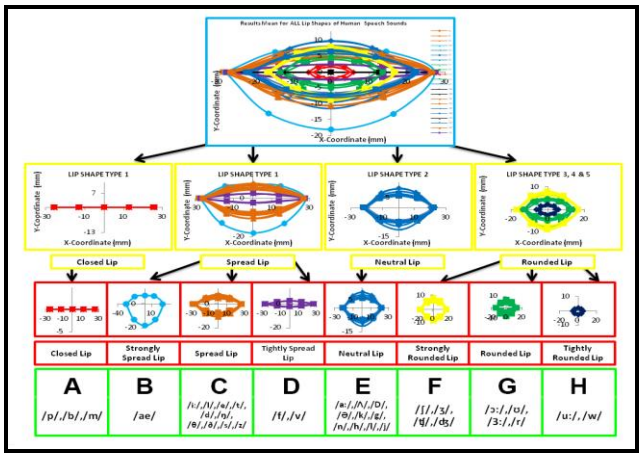


Fig. 16 Classification of Human Lip Shapes and Position

As mentioned and analyzed earlier, human lip shapes were divided into 4 groups derived from the previous studies, they are closed lip, spread lip, neutral lip and rounded lip [15,16,36]. However, results of data analysis from the graphs had proved that human lip shape can actually divided into 8 groups, which were closed lip, strongly spread lip, spread lip, tightly spread lip, neutral lip, strongly rounded lip, rounded lip and tightly rounded lip. The category of closed lip used to pronounce the consonants phoneme sounds ingroup A. The category of strongly spread lip used to pronounce the vowel /ae/ sound. The category of spread lip used to pronounce vowels and consonants sounds in group C. Moreover, the category of tightly spread lip used to pronounce consonants /f/ and /v/ sounds. The category of neutral lip used to pronounce the vowels and consonants in group E. Besides, the category of strongly rounded lip used to pronounce the consonants sounds in group F. The category of rounded lip used to pronounce the vowels and consonants sound in group G. Consequently, the category of tightly rounded lip used to pronounce the vowel /u:/ sound and consonant /w/ sound. The results findings show that some of the human lip shapes in vowels and consonants sounds have the same parameter of viseme category. Such as the categories of viseme spread lip, neutral lip, rounded lip and tightly rounded lip. The visemeof classification were fundamental parameter data set used to develop an accurate platform system for real time talking avatar in multiple languages.

Cross Lingual Transfer

Cross-lingual transfer carried out in this stage was to investigate the correct lip motion in 3 different languages: English, *Bahasa Melayu* and Mandarin. Cross-lingual transfer approach in this research was an idea on a phonetic mapping from *Bahasa Melayu* and Mandarin, the source language, to the phoneme set of the International Phonetic Alphabet (IPA) scheme based on similar phonetic properties among the phonemes.

The results on the acoustic phonetic model according to the research findings from the previous stage were continually used to transfer to the target languages in *Bahasa Melayu* and Mandarin language in Figure 17.

The results from the cross-lingual transfer method were clearly shown in the mapping model in Figure 17. The first row in the chart is the acoustic phonetic model from our research findings that showed the set of viseme classification on English IPA symbols. The second row is the phoneme sounds in *Bahasa Melayu*. The results of transformation to IPA symbols are highlighted in red colour and presented underneath each *Bahasa Melayu's* phoneme sounds. The last row in the table is the phoneme sounds in Mandarin languages. There are three categories shown in this row: JhuyinFuhao, Hanyu Pinyin and IPA symbols. Similarly, the results of transformation from JhuyinFuhao and Hanyu Pinyin to IPA symbols are shown at the bottom and highlighted in red colour.

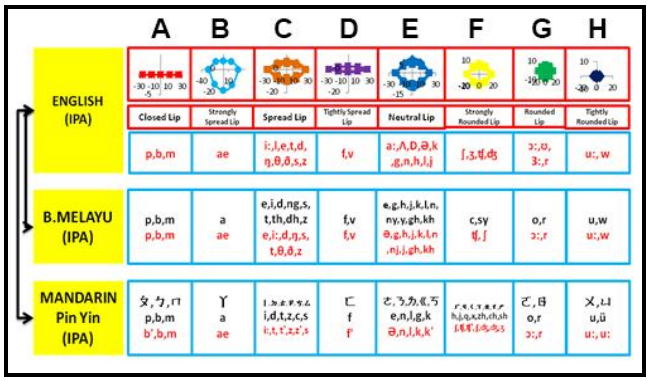


Fig. 17 Cross Lingual Transfer based to the Research Findings Acoustic Phonetic Model to *Bahasa Melayu* and Mandarin Languages in IPA Scheme

According to the result model of cross lingual transfer, the phoneme sounds of English, *Bahasa Melayu*and Mandarin in group A, are using the viseme of closed lipto pronounce the sounds. Besides,viseme of spread lip is used to produce the phonemes sound of group B. The phoneme sounds in group C are pronounced by using the viseme spread lip. The category of viseme tightly spread lip is used to produce the sounds in group D. Moreover, the phoneme sounds of English, *Bahasa Melayu*and Mandarin in group E, are using the viseme of neutral lipto pronounce the sounds.Viseme of strongly rounded lip is used to produce the phonemes sound of group F. Subsequently, the phoneme sounds in group G are pronounced by using the viseme rounded lip. Lastly, phonemes of group H were presented as /u: / and /w/ in IPA scheme, with the viseme of tightly rounded lip to produce the sounds.As a result, the IPA scheme can be applied in three languages of English, *Bahasa Melayu* and Mandarin based on the viseme lip shape classification of human speech sounds.

Viseme Classification

As a result, the modelling of the character's mouth positions was based on the result data from the measurement of active shape model. According to the result findings, the character's target mouth positions were created. The 8 visemes model's lip shapes were then imported into Autodesk MotionBuilder software to build the digital speech system platform in real time as shown in Figure 18.

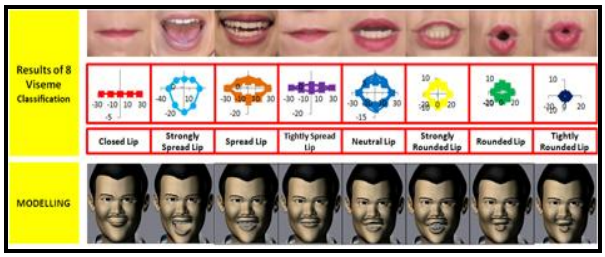


Fig. 18 The Results of 8 Viseme Classifications and Modelling

Based on the study, the automated digital speech system platform was successfully created to perform the lip synchronization animation in real time application. A user needs only to plug in a microphone device into the computer, the automated digital speech system can then performs. The system was made to detect three languages: English, Bahasa Melayu and Mandarin.

Fig. 19 Automated Digital Speech System Platform in English, Bahasa Melayu and Mandarin

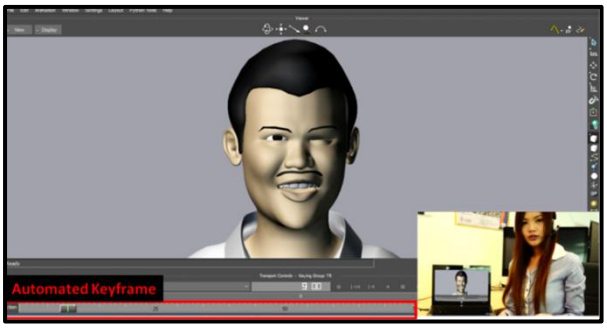


Fig. 20 Real Time Speech Driven Lip Sync Animation in English, Bahasa Melayu and Mandarin

One of the lip sync animation in live performance by using digital speech system through the input audio device is shown in Figure 19 and Figure 20. The digital speech system can also be used in offline mode by detecting the sounds through the inserted audio track file. Therefore, dubbing technique, in which the dialogue is recorded to match a pre-edited version of the animation, is no longer needed for the animation that is published in multiple languages. This method was often time consuming and difficult to achieve an accurate result in matching all the applied languages to the same mouth movement. Hence, the digital speech system is an ideal approach to produce an animation in multiple languages by recording the character's mouth movements which are driven by the live talk through the audio device.

It will automatically keyframe and record all mouth movements of the talking to match the sounds. Although the digital speech system platform applied the actual study from viseme human speech sounds, every animation movement was created based on reality first, then changed and modified to achieve the user's desired motion such as exaggeration animation. It has been show that this new technique gives the user the ability to control the realism of lip sync animation with automated key framing in three languages of English, Bahasa Melayu and Mandarin. These features of system can be applied in live performance, animation production industry, entertainment, gaming, digital marketing and media education.

Performance Validation on Lip Sync Animation

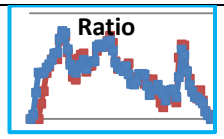
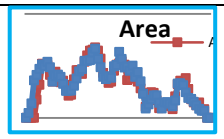
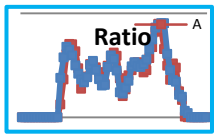
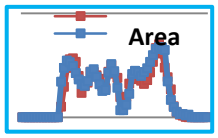
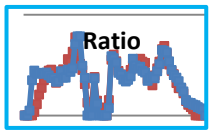
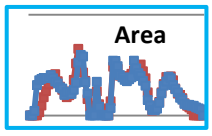
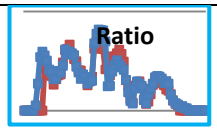
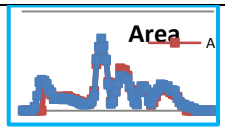
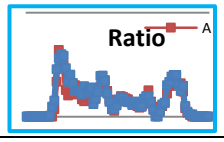
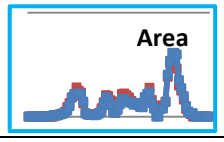
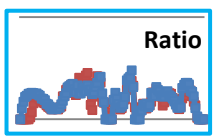
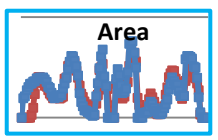
The created system was carried on to validate with different candidates in different languages in different gender of female and male. The candidates were invited to use the digital speech system and read the sentences in different languages of English, Bahasa Melayu and Mandarin. 3 male candidates and 3 female candidates were invited to use the digital speech system and read the sentences in different languages of English, Bahasa Melayu and Mandarin. The candidates included different races formed by Malay, Bumiputera and Chinese peoples. All the measurements were done by comparing model's data and real data via lip position's height (h), width (w), ratio (r) and area (a). Every scene was recorded and each frame was calculated in the AutoCAD software to gain accurate human and avatar's lip positions. The mathematical functions of percentage error and percentage accuracy formula are used in this measurement are:

$$\text{Percentage Error} = \frac{|\text{Approximate Value} - \text{Exact Value}|}{|\text{Exact Value}|} \times 100\% \quad (1)$$

$$\text{Percentage Accuracy} = 100\% - \text{Percentage Error} \quad (2)$$

Percentage error is a measure of the experimental results from the accepted value that would be most closely related to accuracy. It is gained from the difference between approximate values (model's data) and exact values (real data). Besides that, percentage accuracy was used to represent the number of times out of 100 when the avatar's lip sync animation is performed correctly. The data was then calculated in Microsoft Excel and presented in the graphs. Therefore, the data of accuracy was calculated from the measurement between the real human data graph, in which highlighted in blue colour and the avatar model's graph, in which highlighted in red colour. The accuracy data computed the percentage of the created avatar's lip sync animation in achieving realistic simulation.The comparison results are analysed in Table 1.

Table. 1 Comparison Result between Realistic Lip Movements and Model Lip Sync Animation in Different Languages

Different Gender/ Languages	Comparison Graph Ratio	Comparison Graph Area	Accuracy (%)
1. Male Speaker (English) "She is talking on the phone"			80.86
2. Male Speaker (Mandarin) *"大家早上好"* "Good Morning Everyone"			82.14
3. Male Speaker (*Bahasa Melayu*) *"SelamatPagiSemua"* "Good Morning Everyone"			77.22
4. Female Speaker (English) "She found her necklace"			79.67
5. Female Speaker (Mandarin) *"我过得很好"* "I am doing great"			80.76
6. Female Speaker (*Bahasa Melayu*) *"Diasukabacabukucerita"* "He likes to read story book"			76.71

Based on the graphs, the comparison shows that all the avatar lip sync motion paths achieved approximately the same result as in realistic human data, except some small differences between both lip motion paths as shown in Table 1. The differences occurred in the beginning of the speaking, in which the human's mouth opened faster than the avatar in the time of millisecond because the human's mouth opened before the sound came out; vice versa, the avatar's mouth opened as soon as the sound was detected in the system. In addition, some issues like the computer's operating system, sound delaying and environment were also affecting the data to be different between the real person's and the animated avatar's lip motions. However, those would not stimulate big influence in the overall results. The motion paths of modelled avatar's lip sync animation were still matched approximately in the range of accuracy with overall accuracy percentage more than 76%. The validation has successfully proved that the used of 8 viseme classification of human lip shapes can develop accurateperformance of real time avatar lip sync animation, talking in English, *Bahasa Melayu*and Mandarin.

IV. DISCUSSION

This paper proposed the 8viseme classification of human lip shapes and developed an automated digital speech system for performing lip sync animation in real time. The viseme lip shapes include closed lip, strongly spread lip, spread lip, tightly spread lip, neutral lip, strongly rounded lip, rounded lip and tightly rounded lip. The proper lip shapes had been integrated in a real time character animation that synchronized to the International Phonetic Alphabet phonemes sounds for English, Bahasa Melayu and Mandarin languages. Autodesk Motion Builder was the core computer graphics software used to develop the system and perform the tasks. Other than that, Autodesk Maya was used to do the blendshape modelling for 8 viseme classification lip shapes. The integration with plug-in component and live device in Autodesk MotionBuilder software enables the hardware detection between the 3D blendshape modeling and microphone input devices in developing the system, by using the neural network mapping to create the speech recognition function. Automated Digital Speech System was developed to make a virtual computer-generated avatar to talk automatically by using lip synchronization platform. It allows people and companies to use the application in animation production and live performances such as theater, broadcasting, education, game and live presentation according their needs to satisfy the multilingual society in Malaysia.

Fig. 21 Application of Digital Speech System in Animation Production for Doing Lip Sync Animation Automatically

Real time lip sync animation is valuable in open-ended field with tons of potential.The created platform is applicable for doing lip sync animation automatically for online and offline application as presented in Figure 21. For online application, it can be implemented in a live performance for entertainment and media education. Conversely for offline application, it can be used for making the lip sync animation in production phase, where the system can record the live performance for lip sync animation without setting keyframe values. It helps the users to save more time and reduce the workload while enable faster output.

Fig. 22 Application of Digital Speech System in Broadcasting Market

Fig. 23 Application of Digital Speech System in Entertainment or Education Presentation

For real time performance, the system is able to be implemented in broadcasting market, production environment, interactive applications and education purpose: segment hosts in television or commercials shows such as music video, variety, news and game show. It can also be used to serve a contemporary performance where the real time simulation is required, as shown in Figure 22 and Figure 23.

Fig. 24 Application of Digital Speech System in Establishment Place

Apart from that, the real time character animation derived from the system is able to attract attention of a crowd and get the advertising messages noticed. This mode of advertising is encouraged by business houses who want to make their products known to a wide range of customers. So, it is a fascinating idea to implement the system's production at established places such as restaurant, shopping mall or theme park as presented in Figure 24. On the other hand, itcan also apply for online game to convey the message and interact with others player in real time. It allows the avatar character in the game to talk and synchronized with the player. This has a great potential to integrate in VR game.

This study designed viseme lip shapes and developed automated lip sync animation in real time system. The system is a multilingual solution which accurately predicts mouth movement on real human face when a person is speaking and directs to lip sync process. The system is adjusted to the characteristics of English, *Bahasa Melayu* and Mandarin languages. So, when a person is speaking into a microphone, the computer image will animate nearly identical to the way he is speaking in live time. And because it is audio-driven, it is at a price point that makes it scalable to huge amounts of speech demands. To be highlighted, in the stage of pre-production in any film and animation feature, the real time speech function for reference to the animators and artist will shorten the pre-production duration by reducing work on animating. The used of this application enables the animators to estimate the exact time frame for them to do the lip sync and character's animates. It helps to give the overview of character's lip sync by planning the whole animation progress. It is a very critical planning which can save a lot of time, workload and budget, to bring a good blueprint to the next process in the stage of post-production.

Moreover, this research provided automated digital speech model of viseme classification mapping to match the key phoneme sounds for English, Bahasa Melayu and Mandarin languages. This is a critical part to the animator in making lip sync animation, especially in the stage of pre-production in any film and animation feature. Different categories of lip shapes in producing different phonemes sound had been analyzed very specifically. The lip sync animation can work surprisingly well while achieving a good balance between animation realism and run-time efficiency, as demonstrated in the results. Due to the approach is phoneme-based, it can handle speech input from different speakers naturally. Hence, it can be used as a guideline to the animators in animating their desired lip sync motion depending on their creativity, although the viseme categories of lip shape study in this research is according to the standard mathematical calculation in which lip shape was determined based on the standard measurement range. However, most of the animation performances are created for entertainment purpose where the lip shape may be influenced by the character's emotion, personality or exaggeration. Still, the correct basic physical form of lip shape for speaking avatar to match with the phoneme sound remains the same. All animations were created based on the reality in real world, and then were modified according to the creator's desired motion. So, the animation can reach naturalistic quality of believability to the audiences. The animation dynamic is not breaking the principles and rules in the performance. It is because physical realism is often the most important aspect in computer animation. It is an approach to make computer generated realistic-looking virtual environments.

To note, 3D character animation design in this study is not in the scope for our research. Therefore, the character can be changed by modifying the blendshape model, as long as the guideline in the viseme classification study is followed through. Once the blendshape model is created, the lip shape model can then be imported into the digital speech system for making the lip sync animation performance in real time. The system utilizes inexpensive hardware such as microphone which is affordable. This makes it possible for a wider range of people and companies to use the application for their needs. This paper provided useful information on the development of the digital speech system. The development process was explained which could be used as a reference in future research.

V. CONCLUSION

Overall, this study developed an automated lip sync animation for performing a virtual computer-generated avatar to talk automatically. The appropriate lip sync motion driven by the speech recognition system had been integrated in a real time character animation, allowing the user to control the character's speech by using an audio device. The viseme lip shapes are designed based on the phoneme sounds of English, Bahasa Melayu and Mandarin languages for automated digital speech system. Hence, this study indicates that utilization of real time synchronization on viseme-based speech analysis platform can simulate lip sync animation for multiple languages, English, Bahasa Melayu and Mandarin, and ensuring an accurate lip motion result at the same time. The human voice is possible to directly synthesizing the mouth movement from acoustic speech information in real time.

However, there is room for improvement where future research can implement in more methods to increase accuracy of the phoneme recognition from live speech, and also consider for more lip features parts in the performance to enhance the realism in lip sync animation. Despite that, real time system should be generalized towards other parts of a character such as facial gestures, emotions, and other body parts of hand, leg, and hip for future study. Ideally, it can also be enhanced by adding in some animation values based on realism, so the character animation will have their own characteristic to grab and hold the audience's attention while conveying the message to the audiences.

This research has successfully covered two fundamental studies in lip sync animation. They include the study on human lip shape to produce the sound, and accurate-timed lip sync motion. The findings revealed successful and satisfying results in this study. The achievement were suitable to be used in the production of animation films in multiple languages, especially in Malaysia, a country that has three major languages of English, Bahasa Melayu and Chinese widely spoken by the citizens here.

REFERENCES

1. G. Zoric&I. S. Pandzic, "A Real-Time Lip Sync System Using A Genetic Algorithm for Automatic Neural Network Configuration," Proc. *IEEE International Conference on Multimedia & Expo ICME*. Amsterdam, The Netherlands, 2005.
2. I. R. Ali, G. Sulong& H. Kolivand, "Realistic Lip Syncing for Virtual Character Using Common Viseme Set," *Computer and Information Science*, 2015.
3. S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins & I. Matthews, "A Deep Learning Approach for Generalized Speech Animation," *ACM Trans. Graphics*.36, 4, 2017.
4. J. Serra, J. F. Freitas, M. D. Dias&V. Orvalho, "Automatic Visual Speech Animation,"*Proc International Conference on Computational Processing of Portuguese-PROPOR*, Coimbra, Portugal. https://paginas.fe.up.pt /~prodei/dsie12/papers/paper_ 23.pdf, 2012.
5. G. Llorach, A. Evans, J. Blat & V. Hohmann, "Web-Based Live Speech-Driven Lip-Sync," *8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, 2016.
6. P. Edwards, C. Landreth, E. Fiume & K. Singh, "JALI: An Animator-centric Viseme Model for Expressive Lip Synchronization," *ACM Trans. Graphics*, 2016.
7. S. J. Cardman, "Time Management in a Real-Time Animation/Graphics Environment,"*Proc. 2nd Annual Conference on Computer Graphics and Interactive Techniques, Computer Graphics*, 9(9): 201-207, 1975.
8. G. Zoric, "*Automatic Lip Synchronization by Speech Signal Analysis*," Master Thesis (03-Ac-17/2002-z) on Faculty of Electrical Engineering and Computing. University of Zagreb. Zagreb, Croatia. http://www.fer.unizg.hr/images/50009007 /contel05.pdf, 2005.
9. L. Pardew, "*Figures, characters, and avatars: The official guide to using DAZ Studio to create beautiful art (2nd ed.)*," Boston, MA: Course Technology. ISBN: 9781435461208, 2012.
10. S. Roberts, "*Character Animation: 2D Skills for Better 3D, Second Edition*," Oxford ; Burlington, MA: Focal Press. ISBN: 9781136140945, 2012.
11. L. Reveret&I. Essa, "Visual Coding and Tracking of Speech Related Facial Motion,"*In Proc. of Workshop on Cues in Communication*. http://www.cc.gatech.edu/cpl/pubs/cvpr01/reveret-essa-spee ch.pdf, 2001.
12. M. J. Ball&J. Rahilly, "*Phonetics: The science of speech*," London: Edward Arnold. ISBN: 9781444165647, 1999.
13. D. Rogers, "*Reference Vowels*,"http://www.derek.co.uk/ra tional-language-learning/reference-vowels-description.pdf, 2004.
14. P. Roach, "*English Phonetics and Phonology Fourth Edition: A Practical Course*," Cambridge: CUP. ISBN: 9783125344976, 2010.
15. Dayalbagh Educational Institute, "*Spoken English Section 1, Lesson 2 vowels sounds*,"http://www.dei.ac.in/dei/books/files/pdf/spo kenEnglish/Sections/SpokenEnglish-Section1-TheSoundSystemOf English.pdf, 2013.
16. M. J. Ball& N. Muller, "*Phonetics for Communication Disorders*," Hove, UK: Psychology Press. ISBN: 9781317777953, 2014.
17. W. H. Pechter, "*Synchronizing Keyframe Facial Animation to Multiple Text-to-Speech Engines and Natural Voice with Fast Response Time*," PhD dissertation, Dartmouth College Hanover, NH. http://www.cs.dartmouth.edu/reports/TR2004-501.pdf, 2004.
18. R. Parent, "*Computer Animation: Algorithms and Techniques, 3rd Ed*," Waltham: Morgan Kaufmann. Newnes. ISBN: 9780124158429, 2012.
19. W. Belvins, "*Phonics from A to Z: A Practical Guide*," New York: Scholastic Professional Books. ISBN: 9780590315104, 1998.
20. N. S. A. R.Sharifah, N. Nornajmah&O. Rosdah, "*Bahasa MelayuKertas 1 & 2 Teks&RujukanLengkap STPM*," Selangor: KHL Printing Co. Sdn. Bhd. ArahPendidikanSdn Bhd. ISBN: 9789833716692, 2007.
21. A. Hassan, "*Linguistik Am, Siri PengajarandanPembelajaran Bahasa Melayu*," Pahang: PTS Profesional Publishing Sdn Bhd. ISBN: 9789833376186, 2005.
22. A. Razali, "PengenalanFonetikdanFonologi Bahasa Melayu," Fakulti-Pendidikan Dan Bahasa, Sabah Learning Centre, Open University Malaysia. http://assignment.oum.edu.my/uploadasg/ HBML1203SMP/18544/830504105141_163938_final.pdf, 2012.
23. Fonem Bahasa Melayu,"*Bahasa MelayuAkademik Pm3311 P4*," http://suhaimiibrahim.blogspot.com/2012/09/ fonem-bahasa-melayu.html.
24. S. A. MohdYusof, M. Paulraj&S. Yaacob, "Classification of Malaysian Vowels Using Formant Based Features," *Journal of ICT*, 7: 27-40, 2008.
25. W. Abraham, "*Frommer's Chinese PhraseFinder& Dictionary*," United States:Frommers. ISBN: 9780470178386, 2008.
26. Y. S. Lin&S. C. Peng, "*Acquisition Profiles of Syllable-Initial Consonants in Mandarin-Speaking Children with Cochlear Implants*,"ActaOtolaryngol. 123(9): 1046-53, 2003.
27. X. Du, "*Pinyin and Chinese Children's Phonological Awareness*," Department of Curriculum, Teaching and Learning. University of Toronto. https://tspace.library.utoronto.ca/bitstream/1807/25645/ 11/Du_Xintian_201011_MA_thesis.pdf, 2010.
28. W. V. T. Chiung, "Romanization and Language Planning in Taiwan,"*The Linguistic Association of Korean Journal*, 9(1): 15-43, 2001.
29. C. Shei, "*Understanding the Chinese Language: A Comprehensive Linguistic Introduction*," London; New York:Routledge. ISBN: 9781317662808, 2014.
30. A. Brown, "*International Phonetic Alphabet. The Encyclopedia of Applied Linguistics*,"http://dx.doi.org/10.1002/9781405198431. wbeal0565, 2012.
31. Cambridge,"The International Phonetic Alphabet (IPA) Phonemic Chart,"http://cambridgeenglishonline.com/Phonetics_Focus.
32. E. Allen&K. L. Murdock, "*Body Language: Advanced 3D Character Rigging*," Sybex. ISBN: 9781118058763, 2008.
33. T. Frank, M. Hoch&G. Trogemann, "Automated Lip-Sync for 3D-Character Animation,"*In15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*, Berlin, 1997.
34. C. G. Fisher, "Confusions among Visually Perceived Consonants," *Journal of Speech and Hearing Research*. 11(4): 796–804, 1968.
35. C. Santiago, J. C. Nascimento and J. S. Marques, "Combining an Active Shape and Motion Models for Object Segmentation in Image Sequences," *25th IEEE International Conference on Image Processing (ICIP)*, 2018.
36. P. Roach, "*English Phonetics and Phonology Fourth Edition: A Practical Course*," Cambridge: CUP. ISBN: 9783125344976, 2010.