VISEME SET AND ASCII MAPPING CODES PREPARATION FOR STANDARD MALAY AUDIOVISUAL SPEECH SYNTHESIS

Siti Salwa Salleh, Rahmita Wirza Rahmat, Ramlan Mahmud and Fatimah Ahmad
Faculty Computer Science and Information Technology
Universiti Putra Malaysia, Serdang, Selangor.

Abstract: This paper presents the preparation aspects of identifying ASCII mapping codes for Standard Malay (SM) Phonetics alphabet and visemes set in developing a Standard Malay (SM) audiovisual speech synthesis (AVSS). We describe our effort on the examination of existing visemes set which can be used in presenting SM mouth shapes during speech event. The results of this analysis will be employ in building up a talking head, a computer generated model of speaking faces which integrate a speech synthesis module to the visual representation of facial animation. The visemes and phonemes will be coupled tightly in the AVSS to provide naturalness of the speaking talking head. The study starts by identifying the SM phonemes and dissect the similarity and differences between SM and American-English phoneme. The purpose is to identify the additional phonemes present in SM but not in the American-English. The result would lead to the establishment on the viability to use and extend the DECTAlk visemes set in SM ausdiovisual speech synthesis. To obtain the SM visemes set, a video recording were done. The video recording captures an event of a subject who pronounced a collection of words which consist of every phoneme in SM. In addition to allow the manipulation of the phonetic alphabet in the ASCII computer system, our effort also proposed the setting up of SM phonetics alphabet mapping codes. The mapping procedure followed the guidelines which suggested by the Speech Assessment Methods (SAM) community.

Keywords: Audiovisual speech synthesis, talking head, phoneme, viseme, Standard Malay

INTRODUCTION

Audiovisual speech synthesis is an application that integrates speech synthesis and a visualization of facial animation. The facial animation is synchronized with the segments of speech sound accurately to generate a high intelligibility of the speech perceived and to enhance naturalness of the talking head.

Speech comprises a mixture of audio frequencies, and every speech sound belongs to one of the two main classes known as vowels (V) and consonants (C.). Vowels and consonants belong to basic linguistics units known as phonemes [9]. Each phoneme correspond to a single sound system of the language. For example, the consonant phoneme are /p/,/b/ and /t/ and the vowel phoneme are /a/,/e/ and /u/.

In audiovisual speech synthesis application the visible mouth shapes and motion is known as visemes. Therefore this application will intensively use an integration of phonemes and visemes as the basic units of visible articulatory event.

Many audiovisual speech synthesis have been developed progressively [1, 2, 3, 4, 7] but most of them synthesized an English (American or British English) speech sound.

Not many Malay Language speech synthesis have been studied and developed. El-Imam [6] in his work have developed unrestricted vocabulary Text-to-Speech conversion system for segmental synthesis of Standard Malay speech. His synthesized system uses a modified version of a synthesis method that was previously used to synthesize Arabic. He used that method due to the resemblance of Standard Malay (SM) to Arabic phonetic structure.

The current lack of SM speech synthesis motivates this research. In this paper we will discuss on the characteristics of SM and its phonetics properties, presenting a brief comparison between American-English phonemes and SM phonemes, followed by a discussion on the need to recodes SM phonetics standard into ASCII code, a description of additional visemes for SM and finally concludes with what we perceive as research directions.

DISCUSSIONS

Standard Malay Phoneme

The general study on the characteristics of speech sounds is called phonetics which falls under the pronunciation aspects of language. The diagram below depicted a division of language study structure. Phonetics is important aspects in developing a speech synthesis application, therefore it is necessary for us to understand the phonetics of the Malay language in order to produce a SM audio visual speech synthesis.

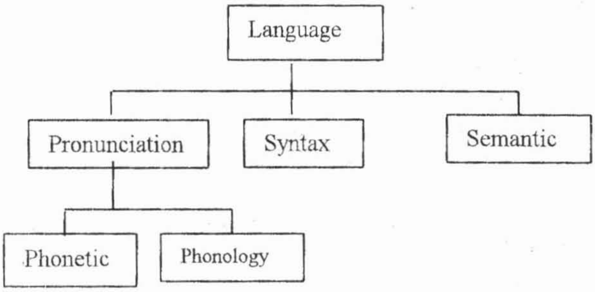


Figure 1: Language study structure.
Adapted from Yusuf.S,1998 [15]

The Malay Language a member of the Western Branch of the great Malayo-Polynesian (Austronesian) phylum of language [8]. Malay is the most widely spoken language of Malaysia. The term 'standard' Malay is a term to designate a variety of language which is basically accepted by members of the speech community to be the norm or the prestige dialect which is used in formal situation [10]. Presently, SM is the national and official language of Malaysia, being spoken by most population of the whole country.

There are thirty six phonemes in Standard Malay phonetics. Nineteen primary consonant which are native consonant sounds are /p/, /b/, /t/, /d/, /k/, /g/, /?/, /m/, /n/, /ɲ/, /ɳ/, /ts/, /dz/, /s/, /h/, /r/, /l/, /w/, and /j/. Another eight secondary consonant are /f/,/v/,/δ/, /ð/, /z/, /ʃ/, /x/ and /θ/ that occur in SM are borrowed from other languages, predominantly English and Arabic [6]. Table 1 shows the underlying characterization of the consonant phonemes of SM.

Table 1: The standard Malay consonant phoneme (adapted from El-Imam,2000)

Place & Manner of articulation	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Palato-alveolar	Palatal	Velar	Uvular	Glotal
Oral Stop	/p/ /b/			/t/ /d/				/k/ /g/		/?/
Nasal (stop)	/m/			/n/			/ɲ/	/ɳ/		
Affricate						/ts/ /dz/				
Fricative		/f/ /v/	/θ//δ/	/s/ /z/		/ʃ/			/x/ /θ/	
Lateral				/l/						
Approximant		[w/			/r/		/j/	/w/	/h/	

Table 2: Standard Malay vowel system (adapted from El-Imam,2000)

Tongue position/height	Front	Central	Back
High or closed	i		
High-mid or half-closed	e	ð	
Low or open		a	
High or closed			u (rounded)
High-mid or half closed			o (rounded)

There are six pure (vocal baku) [5] vocal in SM, /a/, /e/, /e/, /i/, /o/ and /u/. Another two pure vocal, phoneme /E/ and /O/ are not vocal "baku", but it can be used to pronounce word which is borrowed from English. Three diphthong phoneme in SM are /ai/,/au/ and /oi/. For example the dipthong appear in the words 'pantai' where the it appear in a final position in a handful of words.

The most frequently occurring structure in simple SM words is (C)VCV(C) [Payne,1970] where normally the SM stems begin with a consonant. The SM syllables structure can be in CV and CVC, and "vowel-initial stems" is also exist [6].

Comparison to English Phoneme

The American English (AE) consists of twenty four consonants, eight vowels and seven dipthongs [9]. Refer to Appendix 1 for the list of all phonemes in AE. SM differ from AE in the presence of /ɣ/, /x/,/q/,/ɲ/,/ʔ/and the absence of consonant /Θ/, /ð)/, vowel /a/, /ɒ/ and dipthong /ɪ:/, /eɪ/, /u:/ and /eʊ/. SM has a much simpler syllabic structure compared to AE. Words in AE can be in a form of consonant-vowel clusters (CV), vowel-consonant clusters (VC) , and vowel-consonant-vowel clusters (VCV), which is also similar to SM.

Similarity in SM and AE exists in initial and final consonantal cluster of type consonant-consonant cluster (CC) and consonant-consonant-consonant (CCC) cluster that appear as word affixes are also exist. This advent is more obvious in words borrowed from English. For example the prefix /pl/ in a word of "plastik" and and /str/ in word "strategi".

SM Phonetics Mapping Codes

The International Phonetic Alphabet (IPA) is a standardized alphabet for phonetic notation: a comprehensive set of symbols and diacritical marks used to transcribe the speech sounds of all languages in a uniform fashion. It contain a standardized set of symbols for use in transcribing any of the world's languages.

The IPA is widely used in linguistics and dictionaries to indicate the pronunciation of words, and as a basis for creating new writing systems for previously unwritten languages. In new language acquisition aspects the IPA is used in some foreign language text books and phrase books to transcribe the sounds of languages which are written with non-latin alphabets. It is also used by non-native speakers of English when learning to speak English.

It is remarkable that speech sounds in IPA are presented in Greek symbols, which prevent manipulation by computer programs normally applicable where ASCII code is used. As such, they need to be recoded, as applied to all phonetics symbols which are not able to be transmitted or manipulated in ASCII file. The Speech Assessment Methods (SAM) consortium has drawn up a Speech Assessment Methods Phonetics Alphabet (SAMPA); a standard code to map IPA symbol to ASCII in 1988-1991. Since then, many SAMPA notations for other language (Thai, Arabic, French etc) have been established. Nevertheless, until now a mapping the IPA onto ASCII code for the SM phoneme have yet to set up. It is our objective in this paper to present a proposition on a mapping standard codes for SM using SAMPA guidelines transcription.

SAMPA basically consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII codes in the range 33..127, the 7-bit printable ASCII characters. The IPA lower-case alphabet symbols are naturally remain the same in the recoded mapping. These lower case Greek letters a..z, ASCII 97..122. The new codes covered those phonetics symbols within range 37..126. Appendix 2 show a proposed new mapping codes from SM phonetics alphabets into ASCII codes.

Viseme for SM Phoneme

Viseme is a visual presentation of mouth shapes and motions during speech event. Phoneme need to be mapped to correct viseme in audiovisual speech synthesis. Most of the speech sounds (thirty one phonemes) are already in existence in English and in SM. Except four phonemes; /ɣ/, /x/, /q/, /ɲ/ exist in SM but not in English. The benchmark viseme dataset use in acquainting SM visemes are viseme set used in the DECTalk system (Waters & Levergood,1994 [11, 12]).

DECTalk viseme sets were derived from the lip-reading format. In English lip-reading is based on the observation of forty-five phonemes and associated visemes [13]. DECTalk has established forty-five visemes images.

The Visemes Capture Method

In order to elicit the mouth shapes, two subjects utter SM words and sentences. One hundred and eight words were uttered, where it come from three words from each SM phoneme. To make the subject produce a natural speech event, thirty six simple sentences were also uttered. The sentences contain simple words and they are meaningful. Words uttered were VCV, VC , CC and CCC cluster.

The video camera used was Sony DCR-TVR38 DV Camcorder at a 30fps capture speed. The distance between subject and the camera was about 2 meters. Small microphone was also pinned in order to minimize noise in the speech sound acquired. Lights used were a 500 watt lights and defused with a tracing paper to make a proper illumination during the video recording.

Visemes Data Analysis

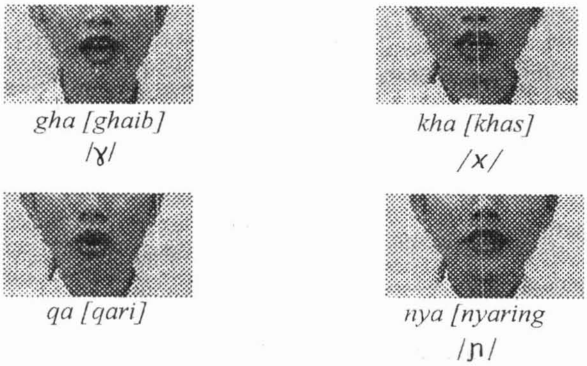
The video was imported into the AVI file format. The AVI file is then analyzed by using the Adobe Premiere to get the frame by frame viseme for each phoneme to be acquired. The viseme were obtain at the video rate of 2 fps, 5fps,10fps and manual click to each frame and the visibility were take into consideration.

Next, one single image for each viseme is identified and extracted from the corpus sequence. This is done manually by searching through the recorded frames.

All visemes were compared to DECTalk viseme image using certain measurement. The visemes were measured base on lips parameters used by Fromkin [cited in King, 2000]. The parameters identified are:

i. Width of lip opening.
ii. Height of lip opening.
iii. Area of lip opening.
iv. Distance between outer-most points of lips.
v. Protusion of upper lip.
vi. Protusion of lower lip.
vii. Distance between upper and lower front teeth.

80% of the viseme capture are similar to DECTalk viseme set. Therefore, it is agreeable that all viseme recorded in our session were the same with the DECTalk visemes by regardless to the individual speaking style . To complete the SM visemes dataset , we added four new mouth shapes which do not exist in the DECTalk viseme sets. The visemes are for the phoneme /ɣ/, /x/ ,/q/ and /ɲ/ The visemes are depicted in the figure 2 below.



gha [ghaib]
/ɣ/

kha [khas]
/x/

qa [qari]

nya [nyaring
/ɲ/

Figure 2: Four additional visemes in SM

CONCLUSION

We have discussed on the difference between SM and AE phoneme and add four visemes in order to be used in the development of the SM audiovisual speech synthesis. Since the speech synthesis receive input and manipulate data in the ASCII codes environment, all the SM phonetics alphabet were recodes to the ASCII representation. This new mapping coding were done by following the guidelines stated by Speech Assessment Method (SAM) Consortium.

The viseme datasets from DECTalk and four new added visemes acquired will be used in phoneme mapping for audiovisual SM speech synthesis . A model-based facial animation will be developed where the mouth shapes and motion are tightly couple to the phonemes to generate natural speech event.

REFERENCES AND BIBLIOGRAPHY

1. Badin, P., Borel, P., Bailly, G., REveret, L., BAciu, M., Segebarth, C. 2000. Towards an Audiovisual an virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images, Proceeding of the 5th Speech Production Seminar, Germany: Kloster Seeon, pp.261-264

2. Beskow. 2003. Talking Heads - Models and Applications for Multimodal Speech Synthesis, PhD thesis, KTH Royal Institute of Technology, Stockholm.

3. Beskow,J. 1996. Talking Head – Communication, articulation and animation, Proceedings of Fonetik- 1996: 53-56

4. Cohen, M.M and Massaro, D.W.1993. Modeling coarticulation in synthetic visual speech. In D. Thalman and N. Magnenat –Thalman (Eds), Models and Techniques in Computer Animation. Springer-Verlag: Tokyo, pp. 141-155

5. I. Dahaman. 1994. Pedoman Sebutan Baku Bahasa Malaysia, Dewan Bahasa dan Pustaka, Kuala Lumpur.

6. Y. El-Imam , Z.M. Don. 2000. Text to Speech Conversion of Standard Malay. International Journal of Speech Technology. 3(2):129-146. June 2000. Kluwer Academic Publishers, Dordrecht.

7. Ezzat, T and Poggio, T.1999. Visual Speech Synthesis by morphing visemes, CBCL Paper NO. 173.

8. Farid. 1980. Aspect of Malay Phonology and Morphology, Publisher: Universiti Kebangsaan Malaysia.

9. P.Roach. 2000. English Phonetics and Phonology, Cambridge University Press, 3rd Edition.

10. T.B.Seong.1994. The Sound System of Malay Revisited, Dewan Bahasa dan Pustaka, Kuala Lumpur.

11. K. Waters and T. Levergood. 1994. An Automatic Lip-Synchronization Algorithm for Synthetic Faces, ACM.

12. K. Waters and T. Levergood. 1994. Decface: An automatic lip-synchronization algorithm for synthetics faces. Technical Report CRL 93.

13. Walther, E.F. 1982. Lipreading Chicago: Nelson-Hall

14. Wells, J.C.- Barry, W.- Grice, M.- Fourcin, A.- Gibbon, d. 1992. Standard Computer-Compatible Transcription. SAM Stage Report Sen.3 SAM UCL-037, 28 February 1992. In SAM (1992) ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardization. Final Report. Year Three: 1.III.91-28.II.1992. London: University College London.

15. S.Yusuf,M.A. 1998. *Fonetik dan Fonologi*, PT Gramedia Pustaka Utama, Jakarta.

16. The Concise Oxford Dictionary, Ninth Edition, Clarendon Press, Oxford, 1995.

APPENDIX

COMPARISON BETWEEN STANDARD MALAY AND AMERICAN-ENGLISH PHONEME

Group of sound	Sound	American-English Word	Malay Word
Obstruent	p b t d k g tʃ (c) dʒ (j) ?	Pat /pat/ Bat /bat/ Tip /tɪp/ Dip /dɪp/ Pick /pɪk/ Big /bɪg/ Chuck /tʃʌk/ Jug /dʒʌg/ -	paku baku tara dada kaya gaya cara jaga ambil
Continuant	s z ʃ (š) ʒ (ž) θ ð(δ) f v h ɣ x q	Sip /sɪp/ Zip /zɪp/ Mesh /mɛʃ / Measure/mɛʒə / Thigh /θʌɪ/ Thy / ðʌɪ/ Fat /fat/ Vat /vat/ Heaven /ˈh ɛv(ə)n/ - - -	sisip zirafah syarat xenon - - fizikal van hasil ghaib khusus qari
Nasal	m n ɲ ŋ	Sum /sʌm/ Sun /sʌn/ - Sung /sɒŋ/	mamak mana nyamuk nganga
Liquid	l r	Leer /lɪ ə / Rear /rɪ ə /	lupa rupa
Semi Vocal	w y	Wet /wɛt/ Yet /yɛt/	awan yakin
Vocal	ʌ (a) ə ɛ (e) ɪ (i) ɒ (o) ʊ (u) a (æ) ɒ (ʌ)	But /bʌt/ Suppose /səˈpə ʊz/ Set /sɛt/ Pit /pɪt/ Dog /dɒg/ Put /pʊt/ Cat /kat/ Pot /pɒt/	parang perang perang giling golong gulung - -
Dipthong	ʌI(ai) (ay) aʊ (au) (aw) ɔɪ (oi) (oy]) ɪ: (iy) eɪ (ey) u: (uw) əʊ (ow)	Cry /kr ʌ ɪ/ Cow /kaʊ/ Boy /bɔɪ/ Heat /hɪ:t/ Say /seɪ/ Lose /lu:z/ Grow /greʊ/	pantai pulau sepoi - - - -

PROPOSED MAPPING CODE FOR STANDARD MALAY PHONETICS ALFHABET TO ASCII

Group of sound	Sound	SAMPA Notation	ASCII (Decimal)
Obstruent	p	p	112
	b	b	98
	t	t	116
	d	d	100
	k	k	107
	g	g	103
	tʃ (c)	c	99
	dʒ (j)	j	106
Continuant	s	s	115
	z	z	122
	ʃ (š)	S	83
	ʒ (ž)	Z	90
	f	f	102
	v	v	118
	h	h	104
	ɣ	G	71
	x	x	120
	q	q	113
Nasal	m	m	109
	n	n	110
	ɲ	J	74
	ŋ	N	78
Liquid	l	l	108
	r	r	114
Semi Vocal	W	w	119
	y	y	121
Vocal	ʌ (a)	V	86
	ə	o	111
	ɛ (e)	E	69
	ɪ (i)	I	73
	ɒ (o)	A	65
	ʊ (u)	U	85
Dipthong	ʌI (ai) (ay)	aI	97 and 73
	aʊ(au) (aw)	aU	97 and 85
	ɔɪ (oi) (oy)	OI	79 and 73

SM phonetics ASCII recodes extended from Wells [Wells et, al,1992]