Automatic Phoneme Identification for Malay Dialects

Yen-Min Jasmina Khaw[1], Tien-Ping Tan[1] and Bali Ranaivo-Malançon [2]
[1]*School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia.*
[2]*Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia.*
jasminakhaw87@hotmail.com

***Abstract*—In many languages such as English, French, German, and Mandarin, there is a documented way of how words are pronounced. The pronunciation of a word is determined by the sequence of phonemes or some speech sounds. Each language or dialect might have different phoneme set. However, there is often a lack of phonological study for a dialect. The number of phonemes is unknown for some of the dialects or languages without a written form. In this work, we propose an approach to identify the phonemes for a dialect from the dialect text transcript and speech corpus, leveraging on existing resources from standard language and multilingual resources. Our study was carried out on Malay dialects. The result shows that the accuracy of the phoneme identification approach is high when we compare the results against previous works in the area.**

***Index Terms*—Phoneme Identification; Malay Dialect; Multilingual; Text Transcript.**

I. INTRODUCTION

In many languages such as English, French, German, and Mandarin, there is a documented way of how words are pronounced. The pronunciation of words can usually be found in a dictionary. The pronunciation of a word is determined by the sequence of phonemes, typically described using IPA (International Phonetic Alphabets) symbols. Other types of speech units may also be used, for example in pinyin is in Mandarin for describing the pronunciation of Mandarin characters. Each language or dialect might have different phoneme set. For example, there are thirty-six phonemes in Malay [1] and forty-four phonemes in English [2]. Determine the phonemes of a language is crucial. It is the first step in many speech processing application such as speech synthesis and automatic speech recognition. The process is often analysed manually by a linguist. After the phonemes are identified, letter to sound or grapheme-to-phoneme (G2P) conversion is a routine that maps the spelling of words to a string of phonetic symbols representing the pronunciation [3]. For example, the word 'ibu' in Standard Malay (English: mother) is converted to pronunciation of /i b u/ where grapheme 'i' is converted to phoneme 'i', grapheme 'b' to phoneme /b/ and grapheme 'u' to phoneme /u/ correspondingly [4].

Speech processing for non-written languages such as dialects is not well studied. There are a few works in speech translation [5, 6, 7], speech syntheses [8, 9] and speech recognition [10, 11]. Pronunciation dictionaries are used to train speech processing systems by describing the pronunciation of words in manageable units such as phonemes [12]. Since most Malay dialects do not have a pronunciation dictionary, by finding and applying G2P conversion rules is one of the ways to develop a Malay dialect pronunciation dictionary. Building pronunciation dictionary for a dialect requires finding out the vocabularies and pronunciations in the language. However, the phonology and phonetics of Malay dialects are not well studied. The phonemes or elementary speech units in a language must be determined before a pronunciation dictionary can be developed. For analysing the phonology of a language or dialect, perception test, acoustic phonetic analysis, and speech processing techniques can be used. Some acoustic analysis [13] can be carried out for analysing the recorded speech sound. It can be done manually, but this approach is time-consuming. Therefore, an automatic way to determine the number of phonemes used in dialects will be very useful.

The Malay dialects in Malaysia that can be grouped according to the geographical distribution. Malay dialects in Peninsular Malaysia are classified into seven groups, the North-Western group comprising Kedah, Perlis, Penang and North Perak dialects; the North-Eastern group, that is, the Kelantan dialect; the Eastern group, that is the Terengganu dialect; the Southern group comprising Johor, Melaka, Selangor and Perak (Southern); the Negeri Sembilan group; the Pahang dialect as a group by itself and not as a member of Southern group and the Perak dialect, the latter of which covers the area of Central Perak . Each group may be further classified to different subdialects according to different areas. For example, Malay dialects spoken in Perak (northern state of Malaysia) can be classified according to five areas. The northern part of Perak speaks Petani and Kedah dialect; the southern part speaks Selangor dialect; slightly to the east part speaks Rawa dialect, while the area around the middle of Perak around Parit and Kuala Kangsar speaks Perak dialect . Different variety of Malay is also very prominent in East Malaysia, Borneo, such as Sarawak Malay dialect.

In this paper, we proposed an approach to identify the phonemes for a dialect from the dialect text transcript and speech corpus by leveraging on standard language resource (e.g. Standard Malay) and multilingual resources. The approach will not pinpoint the actual phonemes in the IPA, but determine the number of unique phonemes and their occurrence in a text. For many speech processing applications such as speech synthesis and automatic speech recognition, this is already sufficient, since they do not need to know the actual type of phonemes in the IPA. At present, two Malay dialects, Kelantan and Sarawak have been collected and analysed. The reason for analysing these two dialects because they are very distinctive compared to Standard Malay. Non-native dialect speakers have difficulty understand the language even if they are native Malay speakers. The paper is organised as following. In section II, the literature review is discussed. Malay dialect read speech and its transcripts in

normalised and unnormalised form acquisition are described in section III. Section VI explains automatic phoneme identification for dialects. In section V, it presents experiments on automatic phoneme identification using Kelantan and Sarawak dialect. The analysis is discussed in section VI. Finally, section VII contains the conclusion and future work.

II. LITERATURE REVIEW

There are several previous studies on phoneme identification approaches such as perception test, acoustic-phonetic and multilingual phone identification.

A. Perception Test

Perception test is an experimental procedure to find which aspects of the signal are used by listeners in decoding speech either to find out more about the signal. Perception test is easy to be carried out. Perception test requires native listeners to listen to some sample of sounds, which differs only in a speech sound . It is typically asking listeners to identify a word or to discriminate between pairs of words. It often uses synthetic or manipulated speech signals to get control over the exact sound. If the listener can distinguish the speech sound, then the speech sound is a phoneme of the language. For example, a phoneme perception test that designed for the phonemes /s/ and /ʃ/ has been developed to investigate whether detection and recognition tasks can measure individual differences in phoneme audibility and recognition for various hearing instrument settings . However, phoneme-level testing is not always easy to use word intelligibility to find out about specific cues or contrasts. It influences of higher linguistic levels which are knowledge of possible words such as frequency of possible words and a likelihood of words in context. In some situations, better to focus on individual phonemes.

B. Acoustic Phonetics

Acoustic phonetics is the study of the acoustic characteristics of speech, including an analysis and description of speech regarding its physical properties, such as frequency, intensity, and duration. Spectrogram is used to study the acoustic features of the spoken signal . Figure 1 shows an example of the spectrogram.

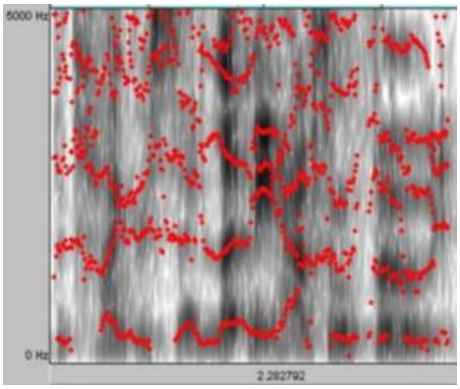


Figure 1: An example of the spectrogram

However, using acoustic analysis can be quite challenging. It requires experts to be carried out. Thus, in a situation when this is not possible, phoneme identification approaches can be useful to predict the phoneme set.

C. Multilingual Phoneme Identification

Multilingual phoneme models can be used for identifying phonemes of unknown languages. Kohler (1996) presents the work to exploit the acoustic-phonetic similarities between several languages. The OGI Multi-Language Telephone Speech Corpus was used where the languages of American English, German and Spanish were selected from the corpus which covers 11 languages in all. A statistical distance measure was introduced to determine the similarities of sounds. Besides, a new approach of multilingual phoneme modeling was introduced. The introduced acoustic-phonetic modeling considers language dependent as well as language independent properties using a density clustering algorithm. The multilingual phonemes modelling technique, which can be used for a variety of language, reduces the number of phoneme-based units in a multilingual speech recognition system. It includes only partial overlap of acoustic region yields improvement of 2%. However, the recognition rates are lower than in the language dependent case.

Anderson, et al. (1994) presents a method to identify poly-phonmes and mono-phonemes for four European languages . Ten acoustically-similar speech sounds were identified across the four languages British-English, Danish, German, and Italian. These sounds that constitute a substantial proportion of the phonemes of each language are designated as (language independent) poly-phonemes and may serve as a multilingual training base for labeling and recognition systems. The remaining sounds of each language, which do not fulfill the similarity conditions, are dubbed mono-phonemes. The speech sounds across languages can be usefully compared along the similarity scale ranging from quasi-equivalence to maximally dissimilar. The similarity measure is based on the work on the identification of poly- phonemes utilised the algorithm of Houstgast to transform the confusion matrices into symmetric similarity matrices. They are merely an extension of, not categorically different from the traditional variation scale with intra-personal equivalence at one end, and dialectally different speakers at the other. Both ends of the similarity scale can be exploited functionally. At the level of quasi-equivalence, poly-phonemes were identified across four languages which increased the average recognition score when used in place of language-specific models. In the maximally dissimilar range, the mono- phonemes offered a basis for language identification.

Gokcen & Gokcen (1997) presents the work toward a universal base for automatic speech recognition . A multilingual phoneme and model set were built. The phoneme set of the system consists of six different languages: US English, Brazilian Portuguese, French, German, Japanese, and Spanish. It contains 61 phoneme symbols, compared to 208 phoneme symbols for the six separate languages. The models were built based on three languages and tested them using two other languages (for which there were no models). The recognition engine uses a continuous density HMM approach . The algorithm accomplishes spectral analysis of a speech signal, making models of the subwords, pattern matching, and performance of post processing validity tests. Features in speech are automatically extracted and compared with previously established reference patterns. A good recognition accuracy was achieved. It has been shown that a single phoneme set and model set based on a few languages, can sufficiently represent other related languages such that new languages can be incorporated into a speech recognition

system with a significantly reduced development time and cost.

Kienappel et al. (2000) present a method to use speech data from multiple languages to enhance the performance of a flexible vocabulary command word recognizer which is trained using a small amount of speech data of the target language . Multilingual phoneme units from the 182 phonemes of base languages French, German, Italian, Portuguese and Spanish are clustered using the phoneme clustering algorithm. It yields a set of 94 multilingual phoneme units. Phoneme-level mapping from the target languages English and Danish to the multilingual phoneme units was then conducted through phoneme distance based mapping and phone confusion based mapping. Next, context-dependent (CD) modeling of multilingual phonemes was carried by capturing the multilingual co-articulation effects using decision graphs for triphone clustering. Finally, it was the step of cross-language adaptation to the new target language using maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR). This method can significantly improve the task-independent phoneme recognition in a new target language with limited training material. The performance was evaluated against the knowledge-based approach of mapping identical SAMPA phoneme symbols. The method achieves significant improvement of recognition performance in the target languages Danish and English by cross-language transfer of multilingual models trained on French, German, Italian, Portuguese and Spanish speech.

III. Malay Dialect Read Speech and Its Transcripts in Normalized and Unnormalized Form Acquisition

In this study, Malay dialect read speech was used instead of conversation style speech as the speaking rate of the conversation is very fast, and it might affect the accuracy of the approach. For preparing the read speech corpus, sentences were selected from Standard Malay text corpus such that the sentences were rich with different varieties of context dependent grapheme. The sentences were then translated using dialect Malay sentences by native speakers. The recording was then carried out. At this point, the transcribed read speech transcript was unnormalised, since the words were written based on the native speaker writing norm. The normalised read speech transcript was prepared based on the alignment of sentences in Standard Malay and Malay dialect. The words in the parallel sentences (e.g. unnormalised and normalised Kelantan sentence) were then aligned using an alignment algorithm. Example below shows the normalised and unnormalised text:

Kelantan dialect (unnormalised): teh adek tawa hebey keh tok letok gulo.

Standard Malay: teh adik rasa tawar kerana terlupa letak gula.

Kelantan dialect (normalised): teh adik tawar hebey keh tak letak gula.

IV. Automatic Phoneme Identification for Dialects

Our approach of automatic phoneme identification approach takes advantage of normalised and unnormalised similar words in the transcripts and multilingual resources in determining the phonemes in a dialect. There are two assumptions made.

Assumption 1: If two aligned context-dependent graphemes from normalised Malay dialect word (Standard Malay word) and unnormalised Malay dialect word are of the same grapheme type, we assume the grapheme is mapped to the same phoneme of the normalised grapheme. If normalised grapheme is a Standard Malay grapheme, then unnormalised grapheme has the same phoneme as the Standard Malay phoneme. Else if the normalised grapheme is not a Standard Malay grapheme, then the unnormalised grapheme is mapped to an unknown unique phoneme that does not exist in Standard Malay.

Figure 2 shows an example of aligned grapheme of same grapheme type. The word ‘kad’ (English: card) in the normalised transcript is aligned to ‘kad’ in the unnormalised transcript of Kelantan dialect at grapheme level where ‘k’ is aligned to ‘k’; ‘a’ to ‘a’ and ‘d’ to ‘d’ correspondingly. This mean that if the grapheme ‘d’ from the normalized transcript is mapped to phoneme /d/, then the grapheme ‘d’ from unnormalised transcript also map to /d/.

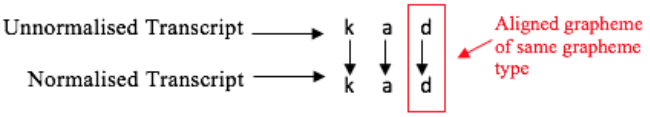


Figure 2: Aligned grapheme of same grapheme type

Assumption 2: If two aligned context-dependent of unnormalised and normalised graphemes are of different types, the unnormalised grapheme either map to a phoneme that associated with the normalised grapheme, or the unnormalised grapheme is mapped to a unique phoneme in Malay dialect. In another word, the unnormalised grapheme used by native speaker might either indicate the phoneme associated with the normalised grapheme or a unique phoneme.

Figure 3 shows an example of aligned grapheme of different grapheme type. The word ‘masa’ (English: time) in the normalised transcript is aligned to ‘maso’ in the unnormalised transcript of Kelantan dialect at grapheme level. The two aligned graphemes, ‘a’ from normalised Malay dialect word and ‘o’ from unnormalised Malay dialect are different. This means that the grapheme ‘o’ in the unnormalised word might either map to a unique dialect phoneme or the phoneme that grapheme ‘o’ from the normalised word is associated.

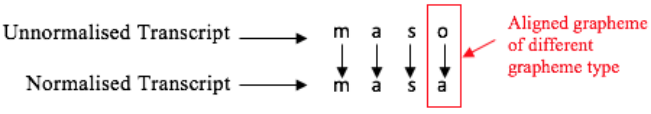


Figure 3: Aligned grapheme of different grapheme type

Figure 4 shows an example of context-dependent aligned grapheme between unnormalised Malay dialect words and normalised Malay dialect words. The word ‘lapar’ (English: hungry) in normalised transcript is aligned to ‘lapa’ in unnormalised transcript of Kelantan dialect at grapheme level where ‘l’ is aligned to ‘l’; ‘a’ to ‘a’; ‘p’ to ‘p’ and ‘a’ to ‘a+r’ correspondingly. For final ‘r’ in the normalised transcript, it is deleted in the unnormalised transcript. The context dependent grapheme ‘a+r’ in the normalised transcript that

aligned to grapheme 'a' in unnormalised transcript might map to a unique phoneme.

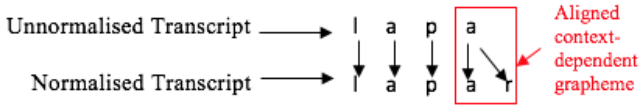


Figure 4: Aligned context-dependent grapheme

For the assumption 2, additional verification is needed. First, multilingual phoneme recognizer is used for predicting the phoneme sequences of the read speech. A grapheme-phoneme confusion matrix is then created by aligning the phoneme sequences from the multilingual phoneme recognition system against the corresponding grapheme sequences of the normalised word and unnormalised word. We assume that if the type of top two phonemes from the grapheme-phoneme confusion matrix for normalised graphemes that mapped to a unnormalised grapheme are the same, the unnormalised grapheme is mapped to phoneme as Standard Malay phoneme that associated with the normalised grapheme of same grapheme type. Otherwise, the unique phonemes in Malay dialect that are possibly different from the Standard Malay are determined using paired sample T-test. Figure 5 illustrates automatic phoneme identification for Malay dialect.

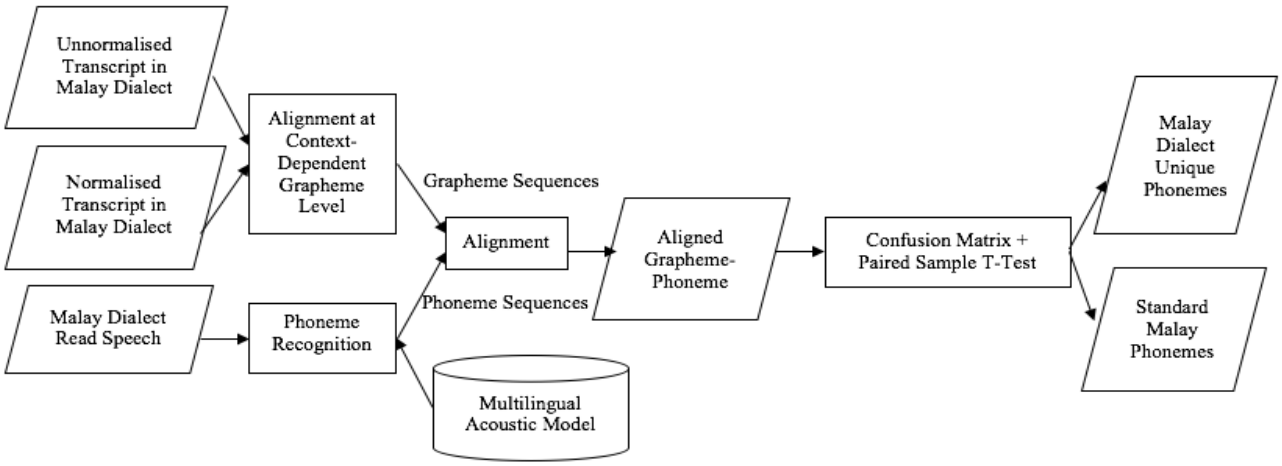
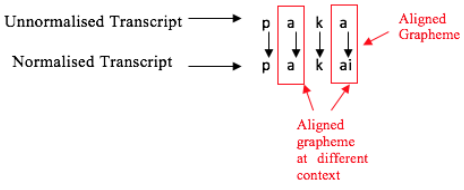


Figure 5: Automatic phoneme identification for Malay dialect

A. *Graphemes Mapping*

The word pairs of normalised and the unnormalised text are first extracted and aligned at grapheme level using Levenshtein distance. For example, the word 'pakai' in normalised transcript (English: wear) is aligned to 'paka' in unnormalised transcript of Kelantan dialect at grapheme level where 'p' is aligned to 'p'; 'a' to 'a'; 'k' to 'k' and 'ai' to 'a' correspondingly. From this example, the final grapheme 'ai' in the normalised transcript that aligned to final grapheme 'a' in the unnormalised transcript are of different grapheme type which might map to a unique phoneme. Besides, each grapheme in the unnormalised transcript of Malay dialect might be aligned to some different types of graphemes in normalised transcript such as from the example shown in Figure 6, 'a' is aligned to 'a' and 'ai' each at different context.

All the possibility of aligned unnormalised graphemes and normalised graphemes are listed to further investigate the possible unique phonemes observed from unnormalised graphemes of Malay dialect that align to different grapheme types of normalised graphemes as stated in assumption 2.



Unnormalized Graphemes	Normalized Graphemes (Type 1)	Normalized Graphemes (Type 2)
p	p	
a	a	ai
k	k	

Figure 6: Aligned grapheme of unnormalised and normalised transcripts

B. *Graphemes-Phoneme Confusion Matrix and Paired Sample T-test*

The subsequent test is carried out to determine the phonemes in some contexts where there is an alignment of different unnormalised grapheme and normalised grapheme, as stated in assumption 2. A multilingual phoneme recognizer is used to decode Malay dialect utterances to the possible phoneme sequences. For example, the French phoneme recognizer produces a phoneme sequence for Kelantan dialect utterances using French phonemes set. A grapheme-phoneme confusion matrix is created by aligning the phoneme sequences from the multilingual phoneme recognition system against the corresponding grapheme sequences through time alignment [28]. The time for graphemes can be obtained by forced aligning the Malay dialect utterances using an automatic speech recognizer, Sphinx3 from CMU [29]. Figure 7 shows an example of time alignment between grapheme sequences and phoneme sequences.

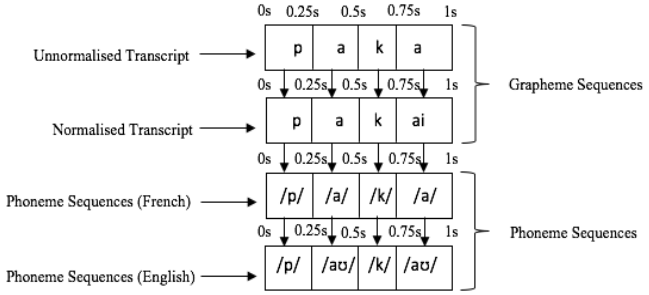


Figure 7: Time alignment between grapheme sequences and phoneme sequences

Finally, the grapheme-phoneme confusion matrix of unnormalised graphemes align to normalised graphemes of

same grapheme types was compared with normalised grapheme of different grapheme types including context-dependent graphemes to find out the unique phonemes in Malay dialect that are possibly different from Standard Malay through paired sample T-test. Based on the assumption, if the type of top two phonemes from the grapheme-phoneme confusion matrix for normalised graphemes that aligned to a unnormalised grapheme are the same, the unnormalised grapheme is mapped to phoneme as Standard Malay phoneme that associated with the normalised grapheme of same grapheme type. Otherwise, pair sample T-test is conducted.

For example, in Figure 7, the unnormalised grapheme 'a' is aligned with the normalised grapheme 'a' at 0.25s. At 0.75s, another grapheme 'a' is aligned with the grapheme 'ai'. When the unnormalised grapheme 'a' aligns with normalised grapheme 'a', both are of the same type, which is 'a'. Thus, we predict the unnormalised grapheme is mapped to a phoneme in Standard Malay associated with normalized grapheme 'a'. When unnormalised grapheme 'a' aligns to normalised grapheme 'ai', since the two graphemes are of different type, we predict the unnormalised grapheme 'a' is either mapped to a phoneme that associates with the normalised grapheme 'a', or a unique phoneme in Kelantan dialect. Grapheme-phoneme confusion of unnormalised grapheme 'a' align to normalised grapheme 'ai' is calculated. It will compare with the grapheme-phoneme confusion of unnormalised grapheme 'a' align to normalised grapheme 'a' based on the assumption made. Table 1 shows the example of confusion of unnormalised grapheme 'a' align to normalised grapheme 'ai' and unnormalised grapheme 'a' align to normalised grapheme 'a'.

Table 1
Example of Grapheme-Phoneme Confusion of Unnormalised Grapheme 'a' Align to Normalised Grapheme 'ai' and Unnormalised Grapheme 'a' Align to Normalised Grapheme 'a'

No.	Unnormalised Grapheme / Normalised Grapheme	French Phoneme-English Phoneme (Top 1)	Confusion	French Phoneme-English Phoneme (Top 2)	Confusion
1.	Unnormalised Grapheme 'a'				
a.	Same Grapheme Type				
	a/a	/a/-/aʊ/	0.0861	/a/-<sil>	0.0734
b	Different Grapheme Type				
	a/ai	/a/-/aʊ/	0.0755	/a/-<sil>	0.0649

From the example in Table 1, the type of top two phonemes from the grapheme-phoneme confusion for normalised graphemes 'a' that aligned to a unnormalised grapheme 'a', and normalised graphemes 'ai' that aligned to a unnormalised grapheme 'a' are the same. Therefore, unnormalised grapheme 'a' is mapped to phoneme as Standard Malay phoneme. However, if the type of top two phonemes from the grapheme-phoneme confusion for normalised graphemes 'a' that aligned to a unnormalised grapheme 'a' and normalised graphemes 'ai' that aligned to a unnormalised grapheme 'a' are of different type, pair sample T-test will be conducted. It is to determine if unnormalised grapheme 'a' align to normalised grapheme 'ai' is mapped to unique phoneme in Kelantan dialect or to a Standard Malay phoneme. The paired sample T-test calculates the difference between unnormalised graphemes that aligns to normalised graphemes of same grapheme types and unnormalised graphemes that aligns to normalised graphemes of different grapheme types including context-dependent graphemes to report if the differences are statistically significant.

After identifying the number of phonemes in Malay dialects, the pronunciation dictionaries for Malay dialects can be developed.

V. Experiments on Phonemes Identification for Malay Dialects

This section discusses experiments on our proposed phoneme identification. The approach requires transcripts in normalised and unnormalised form, and multilingual phoneme recognizer to determine the number of unique phonemes in a dialect. There are 2209 sentences from Kelantan and 1100 sentences from Sarawak dialect were acquired. The read speech was recorded in a soundproof room at Universiti Sains Malaysia (USM), Penang and Universiti Malaysia Sarawak (UNIMAS), Sarawak. As for multilingual phoneme recognizer, French and English phoneme recognition systems were used to decode Malay dialect utterances.

A. Graphemes Mapping

Words were extracted from the normalised and unnormalised sentences of Kelantan dialect and Sarawak dialect. Then, alignment between graphemes from normalised and unnormalised words was carried out. There are two assumptions made.

i. Kelantan Dialect

We align the grapheme of unnormalised word and normalised word using Levenshtein distance. It is to find out the grapheme type of alignment between unnormalised graphemes and normalised graphemes. There are 7490 words in the transcript of Kelantan dialect. The aligned unnormalised graphemes and normalised graphemes can be of different grapheme types at different context. Table 2 shows the grapheme type mapping of normalised graphemes for each unnormalised grapheme of Kelantan dialect.

From Table 1, there are forty-three unnormalised graphemes. Each of them aligns to a normalised grapheme of same grapheme type in Kelantan dialect, for example unnormalised grapheme 'b' aligns to normalised grapheme 'b', and unnormalised grapheme 'g' is aligns to normalised grapheme 'g'. Next, there are seven unnormalised graphemes that align to different grapheme types of normalised graphemes including context-dependent graphemes. For example, unnormalised grapheme 'h' aligns to a normalised grapheme 's' and unnormalised grapheme 'gh' aligns to a normalised grapheme 'r'. Besides, unnormalised grapheme 'a' aligns to normalised grapheme 'a+l' where there is a grapheme 'l' at the right context of vowel 'a' of normalised Kelantan dialect words compared to unnormalised Kelantan dialect words.

Table 2
Grapheme Type Alignment of Normalised Graphemes for Each Unnormalised Grapheme of Kelantan Dialect

No.	Unnormalised Graphemes	Normalised Graphemes			
		Grapheme Type 1	Grapheme Type 2	Grapheme Type 3	Grapheme Type 4
1.	a	a	ai	au	a+l or a+r
2.	b	b			
3.	c	c			
4.	d	d			
5.	e	e	a+m or a+n or a+ng	e+m or e+n or e+ng	e+l or e+r
6.	f	f			
7.	g	g			
8.	gh	r	gh		
9.	h	h	s		
10.	i	i	i+m or i+n or i+ng	i+l or i+r	
11.	j	j			
12.	k	k			
13.	kh	kh			
14.	l	l			
15.	m	m			
16.	n	n			
17.	ng	ng			
18.	ny	ny			
19.	o	o	a	o+m or o+n or o+ng	o+l or o+r
20.	p	p			
21.	q	q			
22.	r	r			
23.	s	s			
24.	sy	sy			
25.	t	t			
26.	u	u	u+m or i+n or u+ng	u+l or u+r	
27.	v	v			
28.	w	w			
29.	y	y			
30.	z	z			
31.	pp	pp			
32.	bb	bb			
33.	tt	tt			
34.	dd	dd			
35.	kk	kk			
36.	gg	gg			
37.	ss	ss			
38.	cc	cc			
39.	jj	jj			
40.	ll	ll			
41.	mm	mm			
42.	nn	nn			
43.	ww	ww			

We summarise Table 2 to the following three tables. Based on assumption 1, the list of unnormalised Kelantan dialect graphemes that align to the normalised graphemes of the same grapheme types is shown in Table 3.

Table 3
Unnormalised Graphemes Align to Normalised Graphemes of The Same Grapheme Types in Kelantan Dialect

Unnormalised and Normalised Graphemes of Same Grapheme Types				
Graphemes			Additional Unique Graphemes	
a	j	q	pp	mm
b	k	r	bb	nn
c	kh	s	tt	ww
d	l	sy	dd	
e	m	t	kk	
f	n	u	gg	
g	ng	v	ss	
gh	ny	w	cc	
h	o	y	jj	
i	p	z	ll	

For these graphemes except additional unique graphemes in Kelantan dialect, we know the actual Standard Malay phonemes that they will map to. For the unique graphemes, we assume they will map to unique phonemes in Kelantan dialect.

From Table 2, based on assumption 2, the list of unnormalised Kelantan dialect graphemes that align to the normalised graphemes of the different grapheme types are shown in Table 4.

Table 4
Unnormalised Graphemes Align to Normalised Graphemes of The Different Grapheme Types in Kelantan Dialect

No.	Unnormalised Graphemes	Normalised Graphemes			
			Grapheme Type 1		Grapheme Type 2
1.	a	1a)	ai	1b)	au
2.	gh	2a)	r		
3.	h	3a)	s		
4.	o	4a)	a		

For these graphemes, they might be either mapped to the phonemes as Standard Malay phonemes or unique phonemes in Kelantan dialect where the actual phonemes are unknown.

Besides, we also do context-dependent grapheme check for graphemes where the right context is of different grapheme types between the aligned normalised graphemes and unnormalised graphemes. The list of the aligned context-dependent graphemes from normalised words and

unnormalised words in Kelantan dialect with different grapheme types at the right context are shown in Table 5.

Table 5
Unnormalised Graphemes Align to Normalised Graphemes with Different Grapheme Types at the Right Context in Kelantan Dialect

No.	Unnormalised Graphemes	Normalised Graphemes					
		Grapheme Type 1		Grapheme Type 2		Grapheme Type 3	
1.	a	1a)	a+l or a+r				
2.	e	2a)	a+m or a+n or a+ng	2b)	e+m or e+n or e+ng	2c)	e+l or e+r
3.	i	3a)	i+m or i+n or i+ng	3b)	i+l or i+r		
4.	o	4a)	o+m or o+n or o+ng	4b)	o+l or o+r		
5.	u	5a)	u+m or u+n or u+ng	5b)	u+l or u+r		

From the table, we can see that there is a grapheme 'l' or 'r' at the right context of vowels 'a', 'e', 'i', 'o' and 'u' of normalised Kelantan dialect words compared to unnormalised Kelantan dialect words. Besides, a nasal consonant can be found in the right context of vowels 'e', 'i', 'o' and 'u' of normalised Kelantan dialect words compared to unnormalised Kelantan dialect words. For these graphemes, they might be either mapped to phonemes as Standard Malay phonemes or unique phonemes in Kelantan dialect. The unnormalised graphemes from assumption 2 will be further investigated using multilingual phoneme recognizer later.

ii. Sarawak Dialect

We align the grapheme between unnormalised word and normalised word using Levenshtein distance. It is to find out the grapheme type of alignment between unnormalised graphemes and normalised graphemes. There are 4183 words in the transcript of Sarawak dialect. The aligned unnormalised graphemes and normalised graphemes can be of different grapheme types at different context. Table 6 shows the grapheme types alignment of normalised graphemes for each unnormalised grapheme of Sarawak dialect.

From Table 5, there are thirty unnormalised graphemes that aligned to each normalised grapheme of same grapheme type in Sarawak dialect, for instance, unnormalised grapheme 'c' is aligned to normalised grapheme 'c' and unnormalised grapheme 'w' is aligned to normalised grapheme 'w'. Next, there are six unnormalised graphemes that aligned to different grapheme types of normalised graphemes including context-dependent graphemes. For example, unnormalised grapheme 'e' aligned to a normalised grapheme 'ai' and unnormalised grapheme 'gh' aligned to a normalised grapheme 'r' of different grapheme type. Besides, unnormalised grapheme 'a' was aligned to normalised grapheme 'a+deleted k' where there is a deleted grapheme 'k' at the right context of normalised Sarawak dialect words compared to unnormalised Sarawak dialect words.

We summarise Table 6 to the following three tables. Based on assumption 1, the list of unnormalised Sarawak dialect graphemes that aligned to the normalised graphemes of the same grapheme types is shown in Table 7.

Table 6
Grapheme Types Alignment of Normalised Graphemes for Each Unnormalised Grapheme of Sarawak Dialect

No.	Unnormalised Graphemes	Normalised Graphemes		
		Grapheme Type 1	Grapheme Type 2	Grapheme Type 3
1.	a	a	a+deleted k	
2.	b	b		
3.	c	c		
4.	d	d		
5.	e	e	ai	e+deleted k
6.	f	f		
7.	g	g		
8.	gh	r	gh	
9.	h	h		
10.	i	i	i+deleted k	
11.	j	j		
12.	k	k		
13.	kh	kh		
14.	l	l		
15.	m	m		
16.	n	n		
17.	ng	ng		
18.	ny	ny		
19.	o	o	au	o+deleted k
20.	p	p		
21.	q	q		
22.	r	r		
23.	s	s		
24.	sy	sy		
25.	t	t		
26.	u	u	u+deleted k	
27.	v	v		
28.	w	w		
29.	y	y		
30.	z	z		

Table 7
Unnormalised Graphemes Aligned to Normalised Graphemes of The Same Grapheme Types in Sarawak Dialect

Unnormalised and Normalised Graphemes of Same Grapheme Types				
a	g	kh	o	t
b	gh	l	p	u
c	h	m	q	v
d	i	n	r	w
e	j	ng	s	y
f	k	ny	sy	z

For these graphemes, we know the actual phonemes that they will map to phonemes as Standard Malay phonemes.

From Table 6, based on assumption 2, the list of unnormalised Sarawak dialect graphemes that aligned to the normalised graphemes of the different grapheme types are shown in Table 8.

Table 8
Unnormalised Graphemes Aligned to Normalised Graphemes of The Different Grapheme Types in Sarawak Dialect

No.	Unnormalised Graphemes	Normalised Graphemes
1.	e	ai
2.	gh	r
3.	o	au

For these graphemes, they might be either mapped to the phoneme as Standard Malay phoneme or unique phonemes in Sarawak dialect where the actual phonemes are unknown.

Besides, we also do context-dependent grapheme check for graphemes where the right context is of different grapheme types between the aligned normalised graphemes and unnormalised graphemes. The list of the aligned context-

dependent graphemes from normalised words and unnormalised words in Sarawak dialect with different grapheme types at the right context are shown in Table 9.

Table 9
Unnormalised Graphemes Aligned to Normalised Graphemes with Different Grapheme Types at The Right Context in Sarawak Dialect

No.	Unnormalised Graphemes	Grapheme Type 1	
1.	a	1a)	a+deleted k
2.	e	2a)	e+deleted k
3.	i	3a)	i+deleted k
4.	o	4a)	o+deleted k
5.	u	5a)	u+deleted k

For these graphemes, they might be either mapped to phonemes as Standard Malay phonemes or unique phonemes in Malay dialect. From the table, there is a deletion of grapheme 'k' at the right context of vowels 'a', 'e', 'i', 'o' and 'u' in normalised Sarawak dialect words compared to unnormalised Sarawak dialect words. The unnormalised graphemes from assumption 2 will be further investigated using multilingual phoneme recognizer later. From unnormalised Sarawak transcription, no additional grapheme was found.

B. Grapheme-Phoneme Confusion Matrix and Paired Sample T-test

A further test was performed to determine phoneme mapping for unnormalised graphemes that align to different grapheme types of normalised graphemes including context-dependent aligned graphemes as stated in assumption 2. There are forty-three unnormalised graphemes of Kelantan dialect and thirty unnormalised graphemes of Sarawak dialect. First, we need to determine if the two aligned graphemes including context-dependent graphemes from normalised Malay dialect word and unnormalised Malay dialect word of different grapheme types in Kelantan dialect and Sarawak dialect are corresponding to unique phonemes or Standard Malay phonemes. French and English phoneme recognition systems were used to decode 2209 Kelantan dialect utterances and 1100 Sarawak dialect utterances. Then, the produced phoneme sequences of the read speech in Kelantan dialect and Sarawak dialect from multilingual phoneme recognizer were aligned at phoneme level against the pairs of normalised and unnormalised transcripts at grapheme level using time alignment to calculate the grapheme-phoneme confusion matrix. Finally, the grapheme-phoneme confusion matrix of unnormalised graphemes aligned to normalised graphemes of the same grapheme types was compared with normalised grapheme of the different grapheme types to find out the unique phonemes in Malay dialect that are possibly different from Standard Malay through paired sample T-test. We assume that if the type of top two phonemes from the grapheme-phoneme confusion matrix for normalised graphemes that aligned to a unnormalised grapheme are the same, the unnormalised grapheme is mapped to phoneme as Standard Malay phoneme that associated with the normalised grapheme of same grapheme type. Otherwise, pair sample T-test is conducted.

VI. ANALYSIS

For the same grapheme type of two aligned graphemes from normalised Malay dialect word (Standard Malay word) and unnormalised Malay dialect word, they were mapped to phonemes as Standard Malay phonemes based on assumption 1. There were seven of unnormalised graphemes that aligned to normalised graphemes of the different grapheme types in Kelantan dialect as in Table 1. However, there is only one unique phoneme found. For example, the unnormalised grapheme 'o' aligned to normalised grapheme 'a' was mapped to a unique phoneme, /P_{vowel1}/ in Kelantan dialect. The types of unique phonemes for these graphemes mapped were unknown. For the rest of the aligned graphemes from unnormalised and normalised Malay dialect word with different grapheme types in Kelantan dialect, they were found to be similar to Standard Malay phoneme as in assumption 2.

Table 10 shows the list of graphemes to phonemes mapping of Kelantan dialect for additional unique graphemes in Kelantan dialect. There are thirteen of additional graphemes in Kelantan dialect. For these graphemes, we assume they were mapped to unique phonemes in Kelantan dialect. However, the actual types of unique phonemes mapped were unknown. Therefore, these unique phonemes were represented using symbols.

Table 10
List of graphemes to phonemes mapping for additional unique graphemes in Kelantan dialect

No.	Graphemes	Phoneme (Symbols)
1.	pp	/$P_{consonant1}$/
2.	bb	/$P_{consonant2}$/
3.	tt	/$P_{consonant3}$/
4.	dd	/$P_{consonant4}$/
5.	kk	/$P_{consonant5}$/
6.	gg	/$P_{consonant6}$/
7.	ss	/$P_{consonant7}$/
8.	cc	/$P_{consonant8}$/
9.	jj	/$P_{consonant9}$/
10.	ll	/$P_{consonant10}$/
11.	mm	/$P_{consonant11}$/
12.	nn	/$P_{consonant12}$/
13.	ww	/$P_{consonant13}$/

In context-dependent grapheme check for graphemes where the right context is different between the aligned normalised graphemes and unnormalised graphemes, there are ten unnormalised graphemes with different grapheme types at the right context compared to normalised graphemes found in Kelantan dialect. The experimental results do not tell us the type of phonemes they are. Five unique phonemes of the vowels were identified as listed in Table 11.

Table 11
List of graphemes to unique phonemes mapping of Kelantan dialect with different grapheme types at the right context between unnormalised graphemes and normalised graphemes

No.	Unnormalised Grapheme / Normalised Grapheme	Phoneme (Symbol)
1.	e/a+m or a+n or a+ng	/P_{vowel2}/
2.	e/e+m or e+n or e+ng	/P_{vowel3}/
3.	i/i +m or i+n or i+ng	/P_{vowel4}/
4.	o/o+m or o+n or o+ng	/P_{vowel5}/
5.	u/u+m or u+n or u+ng	/P_{vowel6}/

By comparing our result with the work by Abdul (2006), the phoneme /P_{vowel1}/ that we found is in fact /ɔ/ in the IPA chart, the unnormalised grapheme 'e' that aligned to

normalised grapheme of ‘a+m or a+n or a+ng’ is mapped to phoneme, /P_{vowel2}/. /P_{vowel2}/ is in fact /ɛ̃/ in the IPA chart. The unnormalised grapheme ‘e’ that aligned to normalised grapheme of ‘e+m or e+n or e+ng’ is mapped to phoneme, /P_{vowel3}/, where this phoneme is /ẽ/ in the IPA chart. The phoneme /P_{vowel4}/, /P_{vowel5}/, and /P_{vowel6}/ are /ĩ/, /õ/ and /ũ/ respectively in the IPA chart. For the additional graphemes found in Kelantan dialect, there are mapped to the geminate consonants each.

For Sarawak dialect, the number of phonemes used was the same as Standard Malay. There was no additional unique grapheme found in Sarawak dialect which means it does not lead to any unique phoneme. Therefore, no unique phoneme was found in Sarawak dialect. Table 12 shows the number of phonemes found in Kelantan dialect and Sarawak dialect. There are twelve vowels and thirty-nine consonants found in Kelantan dialect. For Sarawak dialect, there are six vowels, twenty-five consonants and three diphthongs are discovered.

Table 12
Number of Phonemes in Kelantan Dialect and Sarawak Dialect

Dialects	Number of Phonemes		
	Vowels	Consonants	Diphthongs
Kelantan Dialect	12	39	0
Sarawak Dialect	6	25	3

VII. CONCLUSIONS AND FUTURE WORK

An automatic phoneme identification approach has been proposed for under-resourced Malay dialect. The approach uses normalised transcript and unnormalised transcript to identify the possible phonemes in a dialect. Our proposed phoneme identification approach can be applied to other dialects that have similar writing with standard language, especially under-resourced languages such as Terengganu dialect and Perak dialect. The approach determines the phonemes, but not the actual types of it in the IPA. The accuracy of the phoneme found is high by comparing our result with the previous work of Kelantan dialect and the works of Sarawak dialect . For future works, the experiments on phoneme identification for other Malay dialects such as Perak dialect and Kedah dialect can be evaluated.

ACKNOWLEDGMENT

This work is supported by FRGS grant 203.PKOMP.6711536.

REFERENCES

[1] Y. M. Maris, The Malay Sound System, Malaysia: Siri Teks Fajar Bakti, 1979.

[2] D. Reithaug, Orchestrating Success in Reading, Canada: Stirling Head Enterprises, 2002.

[3] G. Norkevičius, G. Raškinis and A. Kazlauskienė, "Knowledge-Based Grapheme-to-Phoneme Conversion of Lithuanian Words," in *SPECOM 2005, 10th International Conference Speech and Compute*, Greece, 2005.

[4] T. P. Tan and B. Ranaivo-Malancon, "Malay Grapheme to Phoneme Tool for Automatic Speech Recognition," in *Third International Workshop on Malay and Indonesian Language Engineering*, Singapore, 2009.

[5] S. Stuker and A. Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in *in SLTU*, Hanoi, 2008.

[6] S. Stuker, L. Besacier and A. Waibel, "Human Translations Guided Language Discovery for ASR Systems," in *in Interspeech*, Brighton, 2009.

[7] L. Besacier, B. Zhou and Y. Gao, "Towards Speech Transla- tion of Non Written Languages," in *in SLT*, Aruba, 2006.

[8] S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar and A. W. Black, "Text to Speech in New Languages without a Standardized Orthography," in *in Speech Synthesis Workshop*, 2013.

[9] S. Sitaram, S. Palkar, Y. Chen, A. Parlikar and A. W. Black, "Bootstrapping Text-to-Speech for Speech Processing in Languages Without an Orthography," in *in ICASSP*, Canada, 2013.

[10] F. Stahlberg, T. Schlippe, S. Vogel and T. Schultz, "Word Segmentation through Cross-Lingual Word-to-Phoneme Alignment," in *in SLT*, USA, 2012.

[11] F. Stahlberg, T. Schlippe, S. Vogel and T. Schultz, "Pronunciation Extraction from Phoneme Sequences through Cross- Lingual Word-to-Phoneme Alignment," in *in SLSP*, Tarragona, 2013.

[12] O. Martirosian and M. Davel, "Error Analysis of a Public Do- main Pronunciation Dictionary," in *in PRASA*, 2007.

[13] N. Rezaei and A. Salehi, "An Introduction to Speech Sciences (Acoustic Analysis of Speech)," *Iranian Rehabilitation Journal*, vol. 4, no. 4, pp. 5-14, 2006.

[14] J. T. Colins, "Malay Dialect Research in Malaysia: the Issue of Perspective," *Bijdragen tot de Taal-, Land- en Volkenkunde*, pp. 235-264, 1989.

[15] H. O. Asmah, Aspek Bahasa dan Kajiannya, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1991.

[16] Z. B. Ahmad, The Phonology & Morphology of the Perak Dialect, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1991.

[17] P. Ladefoged, Vowels and Consonants: An Introduction to the Sound of Languages, United Kingdom: Black Well Publishing, 2000.

[18] N. Schmitt, A. Winkler, M. Boretzki and I. Holube, "A Phoneme Perception Test Method for High-Frequency Hearing Aid Fitting," *Journal of the American Academy of Audiology fast track*, vol. 27, p. 367–379, 2016.

[19] R. D. Kent and C. Read, The Acoustic Analysis of Speech, Canada: Singular/Thomson Learning, 2002.

[20] N. S. Kenneth, Acoustic Phonetics (Current Studies in Linguistics), Cambridge, MA: MIT., 2000.

[21] X. D. Huang, A. Acero and H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, New Jersey: Prentice Hall PTR, 2001.

[22] A. Cole, Y. K. Muthusamy and B. T. Oshika, "The OGI Multi-language Telephone Speech Corpus," in *In Proc ICSLP'92*, Banff, 1992.

[23] O. Andersen, P. Dalsgaard and W. Barry, "Data-Driven Identification of Poly- and Mono-phonemes for four European Languages," in *Proceedings of EUROSPEECH'93*, Berlin, 1993.

[24] A. J. Bosman, Speech perception by the hearing impaired, Netherlands: Doctorial thesis, University of Utrecht, 1989.

[25] S. Gokcen and J. M. Gokcen, "A Multilingual Phoneme and Model Set: Toward a Universal Base for Automatic Speech Recognition," in *Automatic Speech Recognition and Understanding, Proceedings., IEEE Workshop on*, 1997.

[26] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, 1989.

[27] A. K. Kienappel, D. Geller and R. Bippus, "Cross-Language Transfer Of Multilingual Phoneme Models," in *ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium Paris*, France, 2000.

[28] T. P. Tan, Automatic Speech Recognition for Non-Native Speakers, France: Universit´e Joseph-Fourier - Grenoble I, 2008.

[29] M. K. Ravishankar, "Sphinx3 Decoders: Online," 2006. [Online]. Available: http://cmusphinx.sourceforge.net/sphinx3/doc/s3_overview.html.. [Accessed 4 April 2017].

[30] T. P. Tan and L. Besacier, "Improving Pronunciation Modeling for Non-native Speech Recognitio," in *in Proc. Interspeech*, Brisbane, 2008.

[31] H. M. Abdul, Sintaksis Dialek Kelantan, Kuala Lumpur: Dewan Bahasa dan Pustaka, 2006.

[32] H. O. Asmah, The Phonological Diversity of the Malay Dialects, Kuala Lumpur: Bahagian Pembinaan dan Pengembangan Bahasa, Dewan Bahasa dan Pustaka, 1977.

[33] H. O. Asmah, Susur Galur Bahasa Melayu, Malaysia: Dewan Bahasa dan Pustaka, Kementerian Pendidikan, 1988.