INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Malayalam-English Code-Switched: Grapheme to Phoneme System

Sreeja Manghat[1], *Sreeram Manghat*[1], *Tanja Schultz*[2]

[1]Independent Researcher, India
[2]University of Bremen, Germany

sreejamanghat@ieee.org, sreeram9@ieee.org, tanja.schultz@uni-bremen.de

Abstract

Grapheme to phoneme conversion is an integral aspect of speech processing. Conversational speech in Malayalam – a low resource Indic language has inter-sentential, intra-sentential code-switching as well as frequent intra-word code-switching with English. Monolingual G2P systems cannot process such special intra-word code-switching scenarios. A G2P system which can handle code-switching developed based on Malayalam-English code-switch speech and text corpora is presented. Since neither Malayalam nor English are phonetic subset of each other, the overlapping phonemes for English – Malayalam are identified and analysed. Additional rules used to handle special cases of Malayalam phonemes and intra-word code-switching in the G2P system is also presented specifically.

Index Terms: G2P, low resource languages, Malayalam, code-switching

1. Introduction

A grapheme is the basic linguistic unit for most of the writing systems [1]. A phoneme is defined as the smallest sound unit of a language that discriminates between a minimal word pair [1]. The process of converting grapheme to phoneme for creating a pronunciation dictionary is called grapheme-to-phoneme conversion, often known as G2P. In traditional speech synthesis and recognition systems, G2P conversion has an inevitable role.

The word orthography refers to a set of conventions for writing a language. Phonographic systems allow rule-based grapheme to phoneme mapping. English orthography is categorized as segmental phonographic [1] with lot of exceptions from one-to-one grapheme to phoneme mapping and thus requires complex rules in G2P system. An abugida is a segmental writing system in which consonant–vowel sequences are written as a unit. Malayalam is an abugida with very close one to one grapheme to phoneme mapping with only a few exceptions [2]. The Malayalam orthography is explained later in the paper.

The first language a person learns to speak and is exposed from childhood to a certain period of time is called the native language or mother tongue [3]. Those who can use more than one language are known as multilingual speakers and studies show that multilingual speakers outnumber the monolingual speakers in the world [4]. In any community of bilingual or multilingual speakers, there is a tendency to mix language elements like phrases, sentences or lexical item of a language with other language during conversations. Code-switching is defined as the use of more than one language by a speaker within a conversation or utterance. Research works shows that there is an increase in percentage of speakers around the world with code-switching in speech [5]. India is a south Asian country with a large number of languages and its dialects spoken across the country. Kerala is a state in south India with a population of more than 35 million. Over a one-third of Keralites live in large cities and over half the population lives in urban areas. Malayalam is the official language of the state Kerala and is the native language of more than 96% people in the state. Kerala state ranks highest among Indian states in literacy rate [6]. After 10^{th} grade, the practice of English language as medium of education for 11^{th} and 12^{th} grade as well as university level courses can be considered as one of the main factors for high level of Malayalam-English code-switching among Keralites.

Malayalam-English code-switched speech contain inter-sentential code-switching, intra-sentential code-switching and intra-word code-switching. Hence monolingual G2P solutions cannot be applied to such speech corpus which contain intra-word code-switching. The problem gets further complicated when there is a clear distinction in the pronunciation, due to different dialects within Malayalam. To the best of our knowledge, there exists no G2P system for Malayalam-English which can handle such intra-word code-switching scenario.

This paper describes our implementation of a G2P system, which can handle the special cases such as intra-word code-switching for Malayalam-English speech corpus. The Malayalam-English code switched speech corpus used for the development of the implemented system contains conversation speech of 42 bilingual speakers. The primary functional test results of the implemented G2P system are also presented.

2. Related Work

There has been various works on language specific G2P systems as well as multilingual systems [7]. Though Malayalam has phonemic orthography, special rules are required in the G2P system to handle specific cases like intra-word code-switching. To the best of our knowledge there are only a very few attempts done earlier to develop a G2P system for Malayalam language in which most of them are phoneme map rather than a standard G2P system which can be utilized for speech research. Alok *et al* proposed some general rules like schwa deletion for Indic languages [8]. Sumi *et al* proposed a basic system on rule based Malayalam G2P conversion [9]. Kavya proposes FST based Malayalam phonetic analyzer which addresses some special cases by providing tags [10]. H. Jing [11] gives details on Malayalam phonology and pronunciation. Thunchath Ezhuthachan Malayalam University [12] provides a phonetic archive on Malayalam phonemes. Apart from these works, there are isolated approaches on a Malayalam transliteration and key board mapping which specify phoneme sets [13] [14].

http://dx.doi.org/10.21437/Interspeech.2020-1936

Also there has been literature work on the Malayalam schwa deletion and gemination rules. A study [7] on shared phoneme list between various Indic languages including Malayalam is done. This common phoneme set is further used in multilingual TTS systems in [15].

Though there has been some work on the monolingual Malayalam G2P, we could not find any G2P system so far for the Malayalam-English code-switching with all typical cases like intra-word code-switching handled. The system explained in this paper is part of our speech research.

3. Dataset

The Malayalam-English code-switched speech corpus used for this study contains 20 hours of speech data with 22640 utterances. Apart from that, data from social media twitter text is also obtained to identify the special cases and usage of new phrases and words which got evolved due to the high use of social media. All the speakers in the conversational speech data are fluent in Malayalam as well as English and their native language is Malayalam. It is found that our conversational code-switched Malayalam-English speech corpus has inter-sentential code-switching, intra-sentential code-switching as well as intra-word code-switching. This data was collected as a pilot dataset for speech research. Table 1 shows the code-switching statistics of the database. The speech data used for developing the system is well transcribed and annotated with proper language boundaries marked. Malayalam words are transcribed in Malayalam script and English words are transcribed in English script.

Table 1: *Types of utterances*

Number of utterances	22640
Utterances having intra-sentential code-switching	47%
Utterances having intra-word code-switching	6%
Number of intra-word code-switched words (English-Malayalam)	1200
Number of intra-word code-switched words (Malayalam-English)	120

In the database, intra-word code-switched words have the English part in English script and the Malayalam part in Malayalam script. The influence of social media results is more usage of non-vocabulary words and slang words. To make the G2P system ready to handle such words, data was also obtained from Twitter, a social media platform. We collected 20,000 words from Twitter which contains intra-word code-switching and slang words. The script obtained was in English.

The corpus contains English-Malayalam intra-word code-switched words as well as Malayalam-English intra-word code-switched words. For example, Director**ന്റെ** (Directorinte) is an English-Malayalam intra-word code-switched word in which the root word is director, an English word and the tail part or suffix is **ന്റെ** (nte) which is the Malayalam part. Where "nte" tail means "**s**" indicating the possessive noun or belongingness (**director's)**.

Kuttikals is an example of Malayalam-English intra-word code-switching where the root word "kuttikal (**കുട്ടികൾ**)" meaning children, is a Malayalam word and "s" is the English tail part in this code-switched word. Here the tail indicates just a slang way of calling the plural word kuttikal.

English-Malayalam intra-word code-switching is predominant in the dataset when compared to Malayalam-English intra-word code-switching. Some of the cases of intra-word code-switching are listed below in Table 2. Also it may be noted that none of the root words used here are borrowed words.

Table 2: *English – Malayalam Intra-word Code-switching*

Intra-word code-switched word	**English part**	**Malayalam part**	**Change at code-switching**
director**ന്റെ** (directorinte)	director	ന്റെ (nte)	r + nte -> ri + nte
point**കൾ** (pointukal)	point	കൾ (kal)	t+kal -> tu + kal
position**ലും** (positionilum)	position	ലും(lum)	n + lum -> ni + lum
film**നു** (filminu)	film	നു (nu)	m + nu -> mi + nu
generation**ഓട്** (generationood)	generation	ഓട് (ood)	No change
cinema**യുടെ** (cinemayude)	cinema	യുടെ (yude)	No change

Table 2 also explains the change in pronunciation that happens during stitching of individual parts of an intra-word code-switched word. In the majority of cases, it is seen that the final pronunciation deviates from the pronunciation derived from a mere stitching of Malayalam and English parts. This highlights the relevance of the system implemented.

4. Grapheme - Phoneme Analysis

Malayalam language consists of 15 vowels and 36 consonants. Neither Malayalam nor English phonemes are subset of each other. However, we can find overlapping phonemes between Malayalam and English. The phoneme dictionary of CMU dict is used for English language. Apart from vowels and consonants, Malayalam language has special cases like chillaksharam, chandrakkala, special symbols for consonant-consonant combinations and cases in which change in phoneme occur after conjunction of consonants. Being abugida, Malayalam writing system has inherent vowel following each consonant. For example, the consonant ധ (dha) is phonetically ധ് (dh) + അ (a) where അ (a) is the inherent vowel and ് symbol used is called chandrakkala. This is important because ധ് (dh) and ധ (dha) is used separately depending on situation.

Table 3: *Overlapping phonemes*

		Malayalam	**English**
Vowels	Total	**15**	**19**
	Overlapping phonemes	6	
		a, i, iː, u, uː, e	
Consonants	Total	**36**	**31**
	Overlapping phonemes	19	
		k, g, t͡ʃ, ɟ/dʒ, ʈ, d, p, f/pʰ, b, r, j, l, ʋ, m, ɳ, h, ŋ, s, ʃ	

The chandrakkala symbol [16] in Malayalam is a diacritic half-moon symbol (◌്) that is used to represent a consonant letter which is not followed by an inherent vowel and is known as samvrithokaaram. If the consonant letter is not followed by inherent vowel as well as any other vowel, the half-moon symbol used is known as virama. Chandrakkala (half-moon symbol diacritic) is assigned an phoneme symbol of schwa [:ə].

Table 3 shows the summary of phonemes and we can see that there is a total of 25 overlapping phonemes between both the languages which also include overlapping English long vowels.

Chillu or ചില്ലക്ഷരം (chillaksaram) [17] is a special independent form of certain consonant letters that is not followed by chandrakkala. For example, ന (na) is the Malayalam consonant and ൻ (n) is its chillu form In the G2P system, chillu is handled separately as a special case.

There are certain special cases where a vowel or consonant have different pronunciation depending upon the combination in which it is used as shown in Table 4. For example, the vowel അം (am) has different pronunciation (am and an) depending on the combination. The symbol ം is called anuswara. Also, the consonant ഫ (p^h) has two different pronunciations (f and p^h) depending on the combination.

Table 4: *Special cases of vowels and consonants*

		Malayalam Word	Transliteration
Vowel അം,ം (am)	c-v	ഗംഗ	g **an** ga
	c-v	ഭംഗി	bh **an** gi
	c-v	അംശം	am sh am
	c-v-v	കിംവാദം	ki**m** va dh **am**
Consonant ഫ (f/p^h)		വാരഫലം	Vaa ra **bha** lam
		ഫണിതം	**fa** ni th am

There are also certain special cases, where the combination of phonemes has a different pronunciation than their individual members. For example, ന്റെ is a combination Malayalam phoneme which when split is ന് + റെ (n̪ + re). The pronunciation of new combination is *nt̪e* and not *n̪ re*.

Special conjunct symbols are used when certain c-c combination occur. Certain phoneme combinations are also represented by a conjunct symbol. Malayalam contains more than 15 special conjunction symbols that are used to represent a c-c combination. For example, ങ (ng) + ങ (ng) is written as ങ്ങ(ngg). There are another three types of special symbols in a c-c combination where the second consonant is denoted by a different symbol. For example, സ് (s) + വ് (v) + അ (a) when combined is written as സ (sva). Here അ (a) is the inherent vowel. In the case of the combination പ്(p)+ര്(r)+അ(a), the combination is written as പ്ര(pra). The symbol is added before the first consonant but the pronunciation is *p ra* and not *r pa*.

5. Implementation

The implemented system has the functionalities mainly focusing on the code-switch dataset mentioned in section 3. Monolingual G2P and intra-word code-switch are the two major functionalities of the system. Finite State Transducer (FST) is used to implement the different states of G2P conversion and it provide flexibility for including more rules during further developments in the system. The overall system as shown in Figure 1 is split into three modules – word identification module, G2P module and intra-word code-switching module.

Module 1: Word Identification.

The input to this module is an utterance of speech transcription. These sentences are split into words or tokens and language identification is done. Unicode code range based language detection is used for language identification. The words identified will be of three types – English word, Malayalam word and intra-word code-switched words. It is observed that in the social media, sometimes English words are written in Malayalam or vice versa. To handle this, dictionary check and transliteration is done.

Module 2: G2P system

Module 2 handles G2P conversion of monolingual Malayalam and English words. Dictionary look up of the CMU dict is used for English G2P system. At this stage, all the new words or unidentified words in the English script are considered as intra-word code-switched words. This could have also happened due to transcription error and is given to module 3 for further processing.

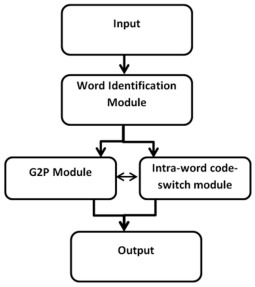

Figure 1: *Overall system implementation*

Malayalam G2P is implemented using rule based FST. Each special case mentioned in section 4 is implemented as a separate transducer. Below are some of the special cases handled by the system.

a. Chandrakkala (diacritic half-moon symbol)

For handling chandrakkala, [:ə] is added in the case of virama, and it is deleted in the case of samvrithokaaram. For example, കൂട് (kood) is ക (**k**) + ് (**:ə**) + ൂ (**u**:) + ട (**ʈ**)+ ് (**:ə**) and the phoneme is / **k u**: **ʈ :ə** / where addition as well as deletion of [:ə] happens. **C**handrakkala can be seen helpful in simplifying the cases associated with pronunciation during schwa deletion [18] in Malayalam.

b. Chillu

The system handles chillu. ർ, ൻ, ൽ, ൾ, ൺ are the chillu form of the constants റ (ɾ), ന (n̪), ല (l), ള (l̪) and ണ (ɳ) respectively. Currently, it is handled by adding schwa symbol [:ə] to the phoneme of chillu.

c. Anuswara (ം)

As explained in Table 4, anuswara have different pronunciation depending upon the position and association with other phonemes.

d. Visarga (ഃ)

The system handles visarga. For example, in ദുഃഖം (dhukkam) and ദുഖം (dhukam), the presence of visarga gives different pronunciation.

e. Combination phoneme

Few specific combination phonemes have different pronunciation when combined together. The system handles such special combinations.

Module 3: intra-word code-switching

This module works in 2 stages. First stage is the identification of type of intra-word code-switch. Second stage converts the identified intra-word code-switched word into its corresponding phoneme. Intra-word code-switching can be Malayalam-English or English-Malayalam. Those words which could not fit into either cases of intra-word code-switch are considered as unidentified words.

a. Code-switched word identification

The code-switched words in code-switching speech dataset were transcribed with root and tail in their respective script. The code-switched words taken from the social media data like twitter were mostly in single language. We have added the functionality to handle this. The module will analyze the tail word first since it is lesser in number and identifies whether it is Malayalam-English or English-Malayalam intra-word code-switched word. Once the tail word is found in a language, then

the module checks for the root word in the second language. If the root word does not belong to dictionary of the second language, then it is added to the list of out of vocabulary (OOV) words and will be further analyzed for new words. If the tail word does not belong to either of the tail word list after transliteration to Malayalam or English, the word is put into the OOV list.

b. Intra-word code switched

As shown in Table 2, the phonemes of the final intra-word code-switched words need not be the exact concatenation of its individual half words. Once the nature of intra-word code-switching is identified and the tail word recognized, rules are applied to get final phonemes. Currently we have identified 14 such rules and it is expected to extend as the dataset effort progresses with more slang concatenations getting added day by day.

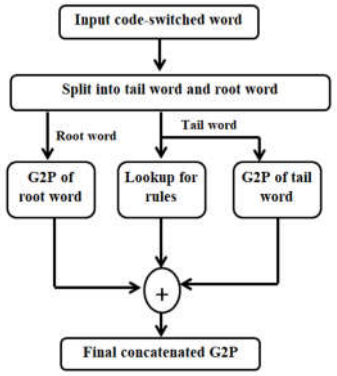

Figure 3: *Intra-word code-switching*

The algorithm for G2P conversion of intra-word code-switching is shown in Figure 3. We have added rules as shown in Table 2, for special cases of intra-word English-Malayalam code-switching.

6. Results

420 reference words were taken and the phonetic dictionary was created manually. These reference words were chosen in such a way that they cover almost all the different phonetic cases and is used to test the Malayalam G2P system. The test pronunciation dictionary is created by passing the reference word list through the monolingual G2P module of the system as batches. Cross-validation [19] is done with phoneme error rate calculated in each iteration with an 80:20 ratio.

The Phoneme Error Rate (PER) is calculated as below.

PER in % per word = 100* Edit distance / Number of elements in the reference phoneme sequence, where edit distance is the minimum number of insertions, deletions and substitutions required to convert the test sequence to reference.

The PER for Monolingual Malayalam, Malayalam-English intra-word and English-Malayalam intra-word has been calculated. A PER of 3.2% was obtained using the developed Malayalam G2P module. English-Malayalam and Malayalam-English code-switched words were taken with each of these words having unique root or tail for testing the system and results showed an average PER of 3.8% and 2.1% respectively as shown in Table 5. The special cases of code-switch occurring with names of person, places and entity is not considered when PER was calculated. The major errors contributing to PER were due to the implementation when the multiple rules came. Apart from this, the coverage of the system was calculated with the percentage of the words skipped [20].

Table 5: *PER and coverage error*

Data	PER			Coverage		
	Min	Max	Avg	Coverage Error	New words	Transcription error
Monolingual Malayalam	2.6%	6.7%	3.2%	14%	47%	53%
Malayalam-English intra-word	0.8%	4.1%	2.1%	6%	78%	22%
English-Malayalam intra-word	2.3%	5.2%	3.8%	12%	88%	12%

Coverage error = 100*number of words skipped / total number of words considered.

Upon further analysis of words skipped (considered as OOV), it was clearly seen that the coverage error is a combination of transcription error and new words. The percentage of new words and transcription error for monolingual Malayalam, Malayalam-English intra-word code-switch and English-Malayalam intra-word code-switch is as shown in table 5. Speakers tend to use many new words due to the influence of movies and social media. Presence of new words can also occur due to dialects. Once identified, OOV words are added manually to the dictionary. For example, the word “machu” (a slang word meaning friend) is new word for G2P system and such words are often transcribed as Malayalam word in the code-switched speech dataset. കട് (kad) is an example for transcription error where the required word is കാട് (kaad). Additionally, names of places and person are currently added to Malayalam dictionary. At this stage, we have not made many changes to English monolingual part apart from cmudict as the focus is mainly on Malayalam and code-switching. Further changes to the system will be required when accented English is also considered.

7. Conclusion

Traditionally, grapheme to phoneme conversion is an important tool in developing any ASR engine as well as speech synthesis systems. Malayalam is a low resourced Indian language. To the best of our knowledge, there is no G2P system developed for code-switched Malayalam-English corpus. We attempt to develop a G2P system for Malayalam-English code-switched data. The amount of intra-word code-switching is showing a trend of upward rise. It is observed that neither Malayalam phoneme set nor English phoneme set is a subset of the other. There are overlapping phonemes in both the languages. The dataset was analyzed for different special cases and rules were developed to handle these special cases during implementation. PER and coverage error is calculated and presented. This is a preliminary system developed based on a pilot dataset of Malayalam-English code-switched speech data. We expect to extend this base system as we progress with our next iterations of data collection and speech research.

8. References

[1] T. Schultz, and K. Kirchhoff, “Language Characteristics” in *Multilingual Speech Processing.* Elsevier, 2006. pp. 5-32.

[2] T. D. Peter and W. Bright, *The World's Writing Systems*. Oxford University Press, 1996.

[3] L. Bloomfield, *Language,* London: Routledge, 2005.

[4] G. R. Tucker, *A Global Perspective on Bilingualism and Bilingual Education*, CMU. Georgetown University Round Table on Languages and Linguistics, 1999.
[5] E. B. Bullock, and J. A. Toribio, *The Cambridge Handbook of Linguistic Code-switching*, Cambridge University Press, 2009.
[6] Government of India, *Census of India 2011*, *Available at: http://censusindia.gov.in* [Accessed: 10 September 2019].
[7] B. Ramani, S. L. Christina, G. A. Rachel, V. S. Solomi, M. Kumar, Nandwana, A. Prakash., S. A. Shanmugam, R. Krishnan, S. P. Kishore, K. Samudravijay, P. Vijayalakshmi, T. Nagarajan, & A. H. Murthy, "A Common Attribute based Unified HTS framework for Speech Synthesis in Indian Languages", *8th ISCA Speech Synthesis Workshop,* Spain. August 31 – September 2, 2013, Spain: Barcelona, 2013.
[8] A. Parlikar, S. Sitaram, A. Wilkinson and W. A. Black, "The Festvox Indic Frontend for Grapheme-to-Phoneme Conversion". *Proceedings of the LREC* 2016, 2016.
[9] S. S. Nair, C. R. Rachitha, and C. S. Kumar, "Rule-Based Grapheme to Phoneme Converter for Malayalam", *International Journal of Computational Linguistics and Natural Language Processing, 2013,* Vol 2 Issue 7.
[10] K Manohar, *FST based Malayalam Phonetic Analyser Available at:*https://kavyamanohar.com/post/malayalam-phonetic-analyser/ [Accessed: 10 September 2019].
[11] H. Jiang, *Malayalam: a Grammatical Sketch and a Text*. Department of Linguistics Rice University, 2010.
[12] Thuchath Ezhuthachan Malayalam University, *Malayalam Phonetic Archieve*. Available at http://www.cmltemu.in/phonetic/#/ [Accessed: 10 September 2019].
[13] GitHub, *An algorithm that transliterates Malayalam script to Roman / Latin characters (commonly 'Manglish') with reasonable phonetic fairness.* Available at: https://github.com/knadh/ml2en [Accessed: 10 September 2019]
[14] N. Nandakumar, *Parayumpole Phonetic Transliteration Tool v2.1.1.1 (TESTING).* Available at: https://nandakumar.co.in/apps/parayumpole/ [Accessed: 10 September 2019].
[15] A. Prakash, A. L. Thomas, S. Umesh and H. A. Murthy., "Building Multilingual End-to-End Speech Synthesisers for Indian Languages", in 10th ISCA Speech Synthesis Workshop, Vienna, pp 194-199, 2019.
[16] R. Sebastian, "Atomic Chillu causing spoofing", *Workshop on Problems of Malayalam encoding in Unicode,* Kerala. January 24-25, 2007, Kerala: University of Kerala, 2007.
[17] Malayalam Keyboard layout and character encoding, Report by the Kerala Bhasha Institute, May 2001.
[18] B. Narasimhan, R. Sproat, and G. Kiraz, "Schwa deletion in Hindi Text to Speech synthesis" in International Journal of Speech Technology, volume 7, pp 319-333, 2004.
[19] T. Gunasegaran, and Y. N. Cheah, Evolutionary cross validation. 8th International Conference on Information Technology (ICIT), 2017.
[20] A. Deri and K. Knight, "Grapheme-to-Phoneme Models for (Almost) Any Language", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 399–408, 2016.