



## Normalization of code-switched text for speech synthesis

Sreeram Manghat<sup>1</sup>, Sreeja Manghat<sup>1</sup>, Tanja Schultz<sup>2</sup>

<sup>1</sup>IEEE Graduate Member

<sup>2</sup>Cognitive Systems Lab, University Bremen, Germany

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

### Abstract

In multilingual communities, code-switching is a common phenomenon. Due to the increase in usage of social media, high level of code-switching is present in social media text as well. These code-switched social media texts are often seen written in monolingual script. Text normalization techniques of the conventional Text-to-Speech (TTS) and machine translation systems may not be able to handle such code-switched texts. Malayalam is a low resource Indic language. Conversational Malayalam contains high level of inter-sentential, intra-sentential as well as intra-word code-switching with English. This paper specifies the techniques for handling Malayalam-English code-switched text data. Evaluation results of experiments conducted on Malayalam-English code-switched data is also presented.

**Index Terms:** code-switching, speech corpus, database development, low resource languages, Malayalam

### 1. Introduction

Speakers who can use more than one language are called multilingual speakers and these multilingual speakers outnumber monolingual speakers [1]. Code-switching is the use of more than one language by a speaker within a conversation or utterance. With the increase in number of bilingual speakers and use of social media, people tend to use multiple languages in speech [2] as well as social media texts. In social media, these texts may not have standard dictionary words. Short forms or microtexts, abbreviations, symbols and words with vowels or letters missing are some of the common cases seen in social media texts.

Number of languages in India is the second highest in the world [3]. The advancements in internet and increased use of social media and TV motivate the use of English language in daily life. Malayalam is the native language of the south Indian state Kerala. Malayalam is a Dravidian language and a low resource Indic language [4]. English being the medium of education at the school and college level, there is a large of number of people native to Kerala who tend to have very high level of code-switching with English. Malayalam-English code-switched speech corpus contains inter-sentential, intra-sentential and intra-word code switching [5]. This is reflected in the social media as well.

Speech synthesis systems or Text-to-Speech (TTS) systems and machine translation system (MT) accept normalized text with some exceptions in text normalization where when there is need to preserve true context or emotion, still maintaining the syntactic elements. Most of the TTS systems are monolingual and the input text is expected to be in single language and written in the standard form. The number of

social media users is increasing day by day and they tend to use more than one language in social media and write both the languages in the same script or different scripts. Also the users tend to abbreviate English words and terms by using phonetic of numbers and letters. For example, people tend to write sentences such as 'How r u' (How are you), 'c u l8r' (see you later) and these may not be found in standard English dictionary and is commonly used in social media texting. This is also called as microtext [6]. The modern TTS systems have to be equipped to accept code-switched input texts as well these social media specific style of texting.

To the best of our knowledge, there exist no normalization modules that can handle Malayalam-English code-switched input social media text. This paper describes our two stage approach used for text normalization developed as part of our ongoing speech research. The analysis is done on the social media data collected and results are presented.

### 2. Related Works

Due to the increase in number of bilingual speakers [2], analysis of code-switched text has been a topic of research in the recent years. Text normalization is an important input stage requirement for any TTS or machine translation systems. Over the past years, various text normalization techniques have been proposed by different authors on non-Indic and Indic languages.

Code switching allows ease of communication with a larger set of phrases and emotions [7]. Romanization is a common form of code-mixing. Romanization refers to the transliteration to the roman script. Romanization makes formal rules not applicable to such data and increase the complexity in natural language processing tasks [8]. In Hindi-English code-mixing and normalized by segregation [7], they created a manually annotated corpus and used it for POS tagging. The results show that the corpus contains 40% of Hindi words written in Romanized script.

English social media text was normalized by using distributional representation [9]. They substituted spelling mistakes with their corresponding normal form.

The shared task on code-switching [10] to analyze the CSTM data used 4 language pairs and the task was to identify which language the word belongs to. In the shared task one of the approaches was to use various lexical and character-based features [11] for word level language identification.

With the development of deep learning algorithms, various researches started using deep learning models for text normalization [12]. There was research on machine learning based text normalization in particular for TTS systems [13]. Neural network based systems showed capability to handle noisy text as well [14]. They achieved a F1 score of 81.75%.

### 3. Code-switched text dataset

In Malayalam-English code-switched social media texts, Malayalam words are often written in English script. Also, there is a large use of English-Malayalam intra-word code-switched words in these social media texts. The text normalization module for Malayalam-English code-switched text should be able to handle all such cases.

#### 3.1. Malayalam language

Malayalam is phonemic language having direct phoneme to grapheme correspondence. Malayalam language consists of 15 vowels and 36 consonants. Neither Malayalam nor English phonemes are subset of each other. There are 25 overlapping [15] phonemes between Malayalam and English. Being abugida, Malayalam writing system has inherent vowel following each consonant. For example, the consonant ഓ (dha) is phonetically ഓ̣ (dh) + ഓ (a) where ഓ (a) is the inherent vowel and ̣ symbol used is called chandrakkala. This is important because ഓ̣ (dh) and ഓ (dha) is used separately depending on situation.

#### 3.2. Dataset

Our dataset for text normalization include social media text from twitter, collected using python script and twitter API. The data from twitter contain inter-sentential, intra-sentential and intra-word code-switching of Malayalam with English. In the data collected from twitter as shown in Table 1, 2150 words are in Malayalam script and 20,000 words in English script. The dataset also contains intra-word code-switching and slang words.

Table 1. Social media text data

Words in Malayalam script	2150
English words	11500
Malayalam words in English script	7580
Intra-word code-switched words	420
Numbers and special symbols	500

The different types of words in the dataset is as detailed below.

- English Dictionary words:** Words written in English script and that are found in the English dictionary. These words considered to be written without any spelling mistakes.
- English words with spelling mistake:** English words that do not have the standard spelling.

Table 2. Agglutination in Malayalam

Malayalam word	English transliteration
പോയി	Poyi
പോയിരുന്നു	Poyirunnu
പോയിക്കൊണ്ടിരുന്നു	Poyikondirinnu
പോയിക്കൊണ്ടിരിക്കുകയാണു	Poyikondirikukayanu
പോയിക്കൊണ്ടിരിക്കുകയാണിരുന്നു	Poyikondirikukayyirunnu
പോയിട്ടുണ്ടാകും	Poyitundakum

- Malayalam words:** Words written in Malayalam script. These words are considered to be written without any errors.
- Malayalam words in English script:** Malayalam words are often written in English script. There words are written either in the standard transliterated form or with letters missing in the standard form. In the latter case, the pronunciation is preserved. Malayalam words have high level of agglutination and vowel missing in this make text normalization more challenging. Table 2 shows an example

of agglutination in Malayalam. In few Malayalam words written in English script, it is seen that there is a sequence of repeating same character. For example, standard Malayalam word 'poyi' is seen written as 'poyiii'.

- Intra-word code-switch:** A large number of intra-word code-switching is seen in English-Malayalam code-switched text. This has to be handled separately. There is change in pronunciation at the point of intra-word code-switching and it has to be handled separately. 'Hospitalil poyi' is an example of English-Malayalam intra-word code-switching where 'hospitalil' means 'to hospital'. There are 14 such suffix [15] which is commonly stitched to the English word to get the corresponding English-Malayalam code-switched word.
- Microtext:** There is a new trend in typing among the social media users, where they write words in short length in order to reduce the message length and for the ease of texting in social media. For example, writing ttyl (talk to you later), fb (facebook), mob (mobile), asap (as soon as possible) etc.
- Numbers and Special characters:** Special symbols are commonly used to represent currency. Numbers are written in different form in different situation like date, phone number, currency, numeric values etc. The pronunciation for number in each of this different situation is different.
- Out of vocabulary words:** Many nouns are not part of vocabulary. These will be considered as OOV and it has to match with maximum probability score for language mapping. Also due to the influence of movies and tv, the list of new words used in conversation is very common. For example, the word 'machu' in Malayalam is a non-standard word used to refer a friend.

### 4. Text normalization strategy

Our novel approach uses a two stage strategy to complete the text normalization of Malayalam-English code-switched text, as shown in figure 1.

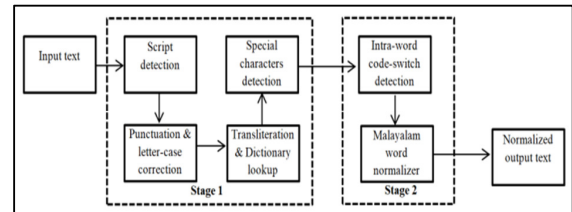


Figure 1. Text normalization module

#### 4.1. Stage one – English words and special characters

The first stage of text normalization include Unicode based language detection, letter-case correction, punctuation correction and English dictionary lookup. The input text to stage one can be in English script, Malayalam script or mixed script of English and Malayalam. From Table 1, it is seen that most social media users have a tendency to write Malayalam words or intra-word code-switched words in English script thus writing the entire text in English. Unicode range based language identification is used for this purpose and tags are given. This step only detects the script in which words are written. Once the script language is identified, punctuation correction and letter case correction for English script is done. After this step, words tagged as English is given to the dictionary lookup step of stage one. These

words can be English words (with or without spelling error) or Malayalam words written in English script. A standard English spell check module is combined to the dictionary lookup module. The words in Malayalam script is directly transliterated to do dictionary check with English so as to identify English words written in Malayalam and remaining Malayalam words are given to stage two of text normalization. Also, a small additional dictionary is manually created for very commonly used microtexts such as *ttyl*, *asap* etc. All other words in English script are marked as OOV words and are also given to stage two of text normalization. The number module of stage one identifies and normalizes numbers and special symbols.

#### 4.2. Stage two – Malayalam and intra-word code-switch

There are two types of inputs to the stage two. One is the Malayalam words written in Malayalam script. Other is the OOV words of stage one that are written in English script.

##### Case 1: Words in Malayalam script

The words written in Malayalam script can be Malayalam word or intra-word code-switched words written in Malayalam. It is observed from the dataset that, the tendency to write short words or microtexts are not seen when a user writes completely in Malayalam script.

##### Case 2: Words in English script

The input words in English script to this stage can be Malayalam words written in English script, Malayalam-English intra-word code-switched words, and OOV words.

##### Step 1: Detection of intra-word code-switched word

English-Malayalam intra-word code-switching is predominant than Malayalam-English intra-word code-switched words [15]. It is observed that the suffix part of the code-switched word is limited in number and thus rule based look up is used to detect the presence of suffix part. For example, the word 'directorinte' is an intra-word code-switched word where 'director' is an English word and 'nte' is the Malayalam suffix. The algorithm start by checking words in English script is for possible suffix part from the list. This is done by analyzing the last 2 characters, followed by the next last 2 characters in the word. Once this suffix part is found, the word is split into English part and the Malayalam part after suffix lookup. Standard English dictionary lookup is done for the English part of the word. There is change in pronunciation at the point of stitching in such cases. There are rules which explain the change in the pronunciation after the stitching of English word with its Malayalam suffix. These rules are applied and the normalized word for the intra-word code-switched word is obtained.

##### Step 2: Malayalam words in English script

The words that are remaining after step 1 are the words written in English script that can be Malayalam words or OOV words. Malayalam words in English script can be with or without spelling errors. From the dataset, it is seen that many tend to write it short by omitting vowels in a way that the pronunciation of the shorten version is similar to the standard version. For example, they tend to write 'marnu' instead of 'marannu'. Another situation is where words are written with repetitive vowels. This special case is handled at the moment for TTS by vowel reduction. Also, in Malayalam we have words like *manam* and *maanam* where these two are standard words.

Our approach uses phonetic based spelling correction for Malayalam. In Malayalam, consonants will not be written

without its inherent vowel unless it is at the end of the word or specific situations. The steps in our approach are as shown below.

- a. Split the word into sub units by doing similarity check with phonemes.
- b. Check if the sub unit is part of a CCV combination.
- c. Check if vowel is present after consonant and how many vowels present before next consonant.
- d. Set vowel length as 2 and if more than two is present, reduce it to two and go to step 5.
- e. Take each sub unit without vowel and add inherent vowel as per phonetic rules of Malayalam
- f. Check for matching Malayalam phonemes and output normalized text.

This stage works by assuming that the word written in English script is a Malayalam word. The words are first split into sub units. The phonetic set of Malayalam vowels (V) and consonants (C) is a standard set where every consonant have a fixed possible CV, CVV, CCV combination. Each sub unit is from one consonant to the next consonant. For example, words like *manam* and *maanam* will have 'ma' and 'maa' as their first sub unit respectively. The nature of consonant is different at the end of a word where the consonant will not have an inherent vowel and this is considered as well during the splitting process. Next part is the spelling correction part where the number of vowels is standardized for a misspelled word and also inherent vowels are added to the words written in short form, as explained previously. Once these steps are completed, we get the word in a form where we have each sub unit in the standard form and this can be considered as the normalized text.

## 5. Implementation

### 5.1. Stage one

The first stage uses unicode range based approach for language detection and detection of special characters. This is given to the next step for punctuation correction and letter case correction of English script.

All the words written in Malayalam are directly transliterated to English to detect the presence of English words written in Malayalam script. Words with numbers written in between the string are converted by replacing the number with its corresponding standard word. For example, 'l8r' gets replaced by 'leight'r'. These transliterated words are passed through a spell checker combined with dictionary to filter out standard English dictionary words. The text tagged with 'S' belong to numbers and special characters. This text is checked for the length of string and presence of special characters between numbers and basic rule based approach is used to identify the required output for normalized text.

### 5.2. Stage two

A Finite State Transducer (FST) based approach is used for implementing stage two. The input to stage two is Malayalam words or intra-word code-switched words written in English script. In stage two, intra-word code-switched words are detected first as shown in figure 2. After that, FST will correct the misspelled Malayalam words as showed in stage 2 of figure 1. At the end of this stage, Malayalam words are in Malayalam script.

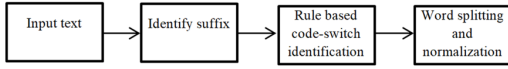


Figure 2. Code-switch text normalization module

## 6. Experimental Results

Preliminary experiments are done on the social media text obtained. To the best of our knowledge, there exist no tools that can be used to compare our text normalization for Malayalam-English code-switched data. Testing of the text normalization module is done using two approaches. In the first approach, the normalized text is matched with a reference for similarity check. In the second approach, the normalized text is given to a TTS system. The test for intelligibility is done and Mean Opinion Score (MOS) for naturalness is obtained for the synthesized speech.

### 6.1. Similarity check

To the best of our knowledge, there exists no data set to be considered as gold standard. From our database of 22000 words, English dictionary words are filtered out using python script.

Table 3. Test Accuracy

Input word type	Accuracy
English words	85%
Malayalam words	100%
Malayalam words written in English script	81%
Intra-word code-switched words	87%
Special characters	75%

Out of the remaining words, 1200 words are selected in random. This is split into 4 batches of 300 words each. Words can be Malayalam words, intra-word code-switched words or English words that aren't found in dictionary. These words are manually normalized to create a reference set. The same 1200 words are passed through the text normalization module and is compared with the reference set generated manually. Two words are considered same or similar only when the output generated by the module is 100% matching with the one obtained after manual normalization. This gives a measure of accuracy. The results of comparison with reference are shown in table 3.

### 6.2. Synthesis

In the second approach, the normalized text was given to our tacotron 2 based TTS system. To the best of our knowledge, there exist no TTS system that can handle Malayalam-English code-switched input. The test for intelligibility and Mean Opinion Score was analyzed by giving it to 30 listeners. The text normalization module developed was used in our deep learning based TTS system and is labeled TTS1 as shown in Table 3. To compare our TTS with existing systems, we used two other two monolingual Malayalam TTS systems available [16] [17] and labeled them as TTS2 and TTS3 respectively. Since TTS2 and TTS3 can handle only monolingual Malayalam input, all the sentences were transliterated to Malayalam for comparison purposes. Both TTS2 and TTS3 can accept Unicode input as well as transliterated input. The input text to all the three TTS systems is same and contains 30 sentences. The input given was without normalization. All the listeners were in the age group 18 to 40 and had Malayalam as

their native language. Everyone had college level education and was fluent in English. 18 listeners were male and 12 were females.

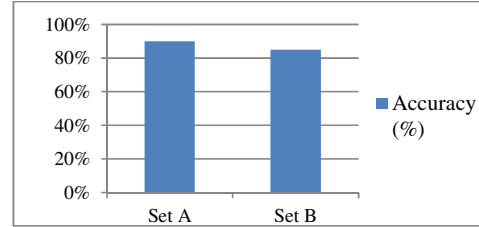


Figure 3. Test for intelligibility

The synthesized output was analyzed in two aspects. In one method, the intelligibility of the output was done by giving it to 30 speakers and asked them to transcribe the text. Inter-transcriber agreement was made during transcription. 20 sentences were given to each transcriber. 10 sentences were taken in random from the synthesis output. Accuracy was measured by comparing the transcribed list with the standard reference created manually. The total number of words transcribed by each transcriber was considered for measurement of accuracy. This score is considered as a measure of intelligibility. Figure 3 shows the average accuracy when same and different sentences were given. Set A is the common sentences and Set B is the different sentence set. Total average accuracy of 87.5% was obtained.

Table 4. MOS score

	Minimum Score	Maximum Score	Mean Score	Standard Deviation
TTS1	3.6	4.1	3.82	0.19
TTS2	3.3	3.9	3.6	0.21
TTS3	3.1	3.5	3.25	0.13

In the second method, MOS score based analysis was done to understand the naturalness of the output. Table 4 shows the MOS score. A mean opinion score of 3.82 was obtained for the TTS output based on our developed text normalization module. MOS score was obtained from 30 speakers and 20 sentences were given to each speaker. 10 sentences were same and remaining was randomly selected from synthesis output. The listeners used headphones and were at their home during testing. The results show that our TTS perform better than other two TTS systems when non-normalized transliterated code-switched input data was given.

## 7. Conclusion

The number of multilingual speakers is increasing day by day and this is reflected in social media text as well. Social media users often write code-switched text in same script and with spelling errors. Text normalization is an important step for speech synthesis systems and machine translation systems. Malayalam is a low resource Indic language with high level of code-switching with English. Our approach handles text normalization of social media text with Malayalam-English code-switching. It can handle Malayalam words and English words written in either of the scripts. Our approach detects intra-word code-switched words and Malayalam words with spelling error. The normalized text output is tested with the reference generated and TTS system. Results shows that our TTS with text normalization have higher accuracy compared to other two.

## 8. References

- [1] G. R. Tucker, *A Global Perspective on Bilingualism and Bilingual Education*, CMU, 1999.
- [2] B. E. Bullock and A. J. Toribio, "The Cambridge Handbook of Linguistic Code-switching", Cambridge University Press, 2009.
- [3] C. Moseley, *Encyclopedia of the World's Endangered Languages*, Routledge, 2008.
- [4] "Census of India 2011", Available at: <http://censusindia.gov.in> [Accessed: 21 September 2020].
- [5] S. Manghat, S. Manghat and T. Schultz, "Malayalam-English Code-Switched: Speech Corpus Development and Analysis", In *proceedings of the First Workshop on Speech Technologies for Code-Switching in Multilingual Communities (WSTCSMC 2020)*, 2020.
- [6] K. D. Rosa and J. Ellen, "Text classification methodologies applied to micro-text in military chat," in *Proc. Eight International Conference on Machine Learning and Applications*, Miami, 2009, pp. 710–714.
- [7] S. Sharma, P. Srinivas, R. C. Balabantaray, "Text normalization of code mix and sentiment analysis, in *proceedings of Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1468–1473. IEEE (2015)
- [8] Y. Vyas, S. Gella, J. Sharma, K. Bali and M. Choudhury, "Pos tagging of English hindi code-mixed social media content", in *proceedings of EMNLP*. vol. 14, pp. 974–979, 2014.
- [9] V. K. R. Sridhar, "Unsupervised text normalization using distributed representations of words and phrases", in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015.
- [10] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, et al., "Overview for the first shared task on language identification in code-switched data, in *Proceedings of the First Workshop on Computational Approaches to Code Switching*. pp. 62–72, 2014.
- [11] G. Chittaranjan, Y. Vyas, K. Bali and M. Choudhury, "Word-level language identification using crf: Code-switching shared task report of msr india system", in *Proceedings of The First Workshop on Computational Approaches to Code Switching* pp. 73–79, 2014.
- [12] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment Treebank", in *proceedings of the 2013 conference on empirical methods in natural language processing*. pp. 1631–1642, 2013.
- [13] R. Sproat and K. Hall, "Applications of maximum entropy rankers to problems in spoken language processing", in *proceedings of Interspeech*, pages 761–764, 2014.
- [14] T. Baldwin, Y. B. Kim, M. C. Mameffe, A. Ritter, B. Han, and W. Xu, "Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition", in *proceedings of WNUT*, 2015.
- [15] S. Manghat, S. Manghat and T. Schultz, "Malayalam-English Code-Switched: Grapheme to Phoneme System", in *proceedings of Interspeech 2020*, 2020.
- [16] "TDIL TTS", Available at: <https://www.iitm.ac.in/donlab/hts/> [Accessed: 21 August 2021].
- [17] "Google based TTS", Available at: <https://www.googletexttospeech.com/p/malayalam-text-to-speech-mp3-downloader.html> [Accessed: 21 August 2021].