Automatic Generation of Pronunciation Lexicon for Malayalam- A Hybrid Approach

Presented by

Sajini T
Scientist C, Language Technology Section
C-DAC - Thiruvananthapuram

Authors
R Ravindra Kumar
K G Sulochana
Jose Stephen

www.cdactvm.in

Contents

- Introduction
- Pronunciation rules in Malayalam and its classification
- Creation of Exception patterns and list
- Implementation of pronunciation lexicon
- Implementation details
- Conclusion
- References

Introduction

- Pronunciation: defined as

  A way of speaking a word, especially a way that is accepted or generally understood or A graphic representation of the way a word is spoken, using phonetic symbols

  Quality of ASR and TTS systems depends on pronunciation
- Pronunciation lexicon maps the orthographic representation of a word to its pronunciation
- Core component of ASR and TTS system
  - Defines the set of valid phoneme sequences, key component in defining the search space of a speech recognizer
  - Provides correct pronunciation for a word in Text to speech

Introduction – contd..

- Creation of pronunciation lexicon is tedious task
  - The existence of foreign words (or words with exceptional pronunciation), and presence of valid multiple pronunciation makes the creation of pronunciation lexicon difficult, even for phonetic languages.

- For large vocabulary recognizers and unlimited vocabulary TTS manual approach is not a feasible option and hence automating the process is a must

Pronunciation rules in Malayalam and its classification

We can classify words in Malayalam in to 3 types

- Type 1: Phonetic words – pronunciation in correspondence to the orthographic representation
  - amma a m m a
- Type 2: Pronunciation which is different from its orthographic representation
  - nanaykkuka n a n# a y k k u k a

Pronunciation rules in Malayalam and its classification – contd..

- Type 3: pronunciation different from respective orthographic representation and have multiple valid pronunciations

  - ennaal — e n n aa l 🔊
  - ennaal(2) — e n# n# aa l 🔊
    - Pronunciation different and depends on the content

  - bulb — b u l b 🔊
  - bulb(2) — b u l b u’ 🔊
    - Add /u’/ sound for some words

Pronunciation rules in Malayalam and its classification – contd..

- Rules are not sufficient for generating pronunciation lexicon
- The pronunciation lexicon is generated using rules and by handling exception

Pronunciation rules in Malayalam and its classification – contd..

- Pronunciation rules are formulated from the analysis of speech corpus
- Rules are classified into 2
  - Group 1
  - Group 2

Pronunciation rules in Malayalam and its classification – contd..

- Group 1
  - Rules depending on the position and the neighbouring characters
  - Example /JA/ will be pronounced as /JE/ at word initial
  - jalam j e l a m
- Group 2
  - Rules applied irrespective of position and neighbouring character
  - /RA/ + /RA/ -> /TTA/
  - parram p a tt a m

Pronunciation rules in Malayalam and its classification – contd..

Exceptions

- Major exception is in the pronunciation of <NNA>
- Dental /NA/ alveolar /NA/ and its geminations have same orthographic representation
- /PHA/ sound in foreign words is different from the /PHA/ sound in malayalam words
- Some pronounce Malayalam /PHA/ as English /PHA/
- /RA/ is pronunced as /RRA/

Pronunciation rules in Malayalam and its classification – contd..

Rules for exception

Case of /NA/

- Dental/nasal /NA/ will occur only at word initials and not with any conjunct combinations
- Rules for /NA/ gemination formulated from corpus analysis

Creation of exception patterns and list

- Pronunciation for exception words is generated using by
  - Creation of exception patterns and its substitution
  - Creation of exception list, with word and its pronunciation

Advantage of using exception pattern

- Exception pattern & exception list reduce the search space
- Words which are not covered by the exception patterns are added in exception list

Creation of exception patterns and list – contd..

- Analysis on approximately 0.35 million words was done to formulate exception patterns and words

- Source of corpus *Online newspapers*

- Exception pattern reduce the search space and lexicon creation time

- Analysis inference
  - Majority ~87% of /NNA/ are dental
  - Using identified patterns, majority of alveolar /NNA/ words were covered
  - Remaining exceptions were common nouns and & foreign words

Creation of exception patterns and list- contd..

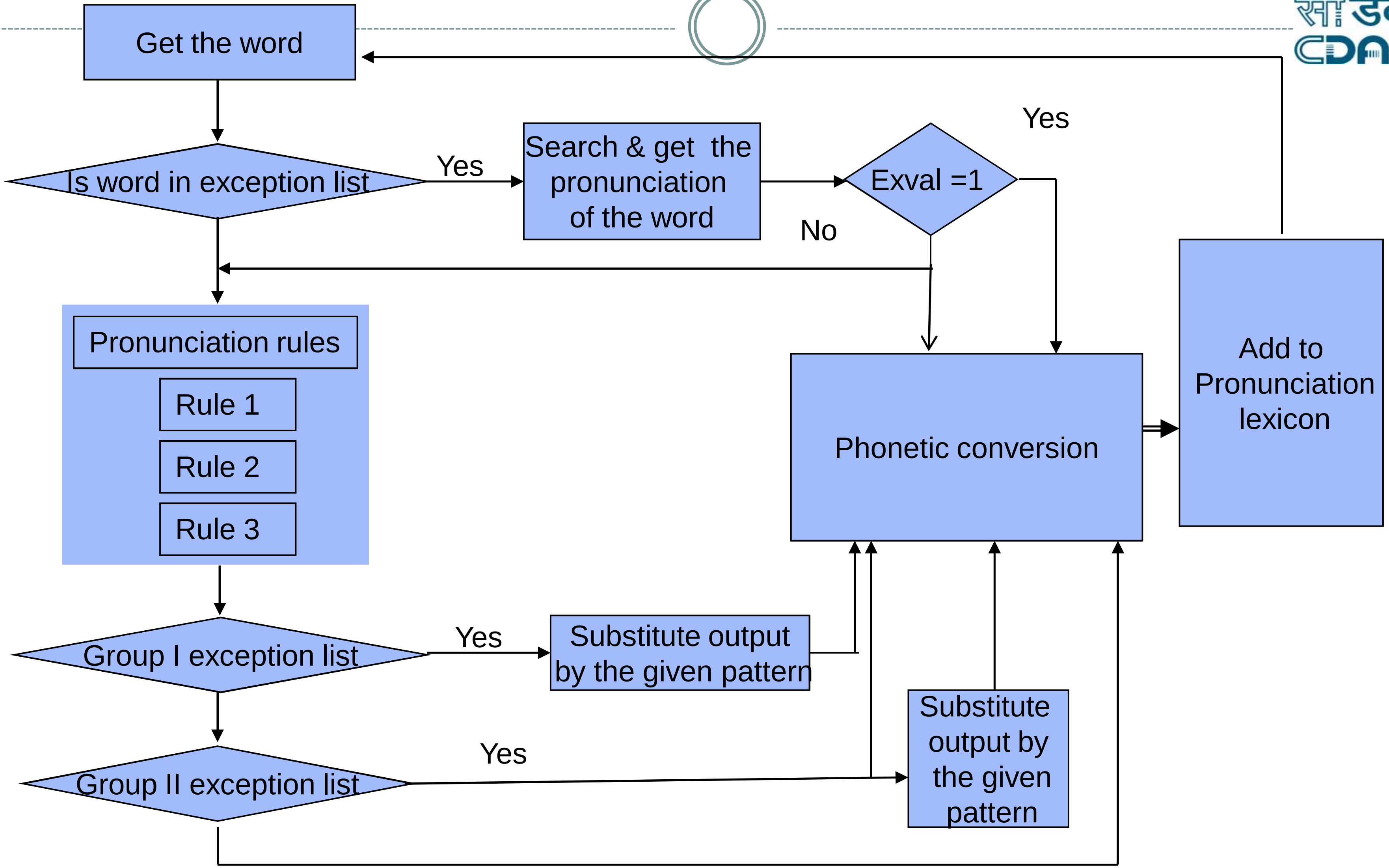
- Frequency of foreign words are high but its count is less (~300)
- From the text corpus selected 250 phonetically rich sentences and recorded by 20 speakers
- Inference on speech corpus analysis
  - Words containing bilabial aspirated unvoiced <PHA> has 2 valid pronunciations
  - Majority add a short u sound at the end of consonant ending foreign words
  - Multiple valid pronunciation exist for certain words containing specific patterns

utsavam	u t s a v a m
utsavam(2)	u l s a v a m
utpannam	u t p a n n a m
utpannam(2)	u l p a n n a m

Creation of exception patterns and list- contd..

- Based on the inferences patterns for exception words were extracted
- Patterns are classified into two
  - Patterns for alveolar /NNA/ - eg /NNAM/, /TANNE/
  - Pattern for /PH/
- These patterns are stored along with the rule file in the following format
- <inpattern><TAB><subpattern>
- Exception words which cannot be identified by patterns are stored in exception file in the format
  - <words>(<Exval>)<TAB><pronunciation>

Implementation of Pronunciation lexicon



Get the word
Is word in exception list
Yes
Search & get the pronunciation of the word
Exval =1
Yes
No
Pronunciation rules
Rule 1
Rule 2
Rule 3
Phonetic conversion
Add to Pronunciation lexicon
Group I exception list
Yes
Substitute output by the given pattern
Group II exception list
Yes
Substitute output by the given pattern

www.cdactvm.in

Implementation details : -

- Rules are separated from the program
- Easy updating, greater flexibility
  - Requires modification in rule file only and mapping file
- Independent of any phonetic notations
  - Rule file & mapping file-to any notation
  - Rule as implemented Unicode standard
  - All rules and exception patterns in single file
  - Order of applying rule
    - Rule 1- Rule 2- Rule 3
    - Then exception pattern group 1- group 2

Implementation details – contd..

Format of the rule file

- Rule 1
  - <inpattern><TAB><subpattern>
- Rule 2
  - <inpattern><TAB><subpattern>
- Rule 3
  - <inpattern><TAB><subrules><subpattern>
- Group 1
  - <inpattern><TAB><subpattern>
- Group 2
  - <inpattern><TAB><subpattern>

Conclusion : -

- ✓ More accurate method than rule based
- ✓ Improved accuracy of ASR
- ✓ Naturalness in TTS

www.cdactvm.in

References : -

1. Carnegie Mellon University, Pittsburgh, PA. Automatic Generation of Pronunciation Dictionaries

2. Dr. V R Proabodhachandran Nayar, Swanavikjanam, Malayalam for Beginners

3. S. Preema, Manu Joseph Department of Linguistic University of kerala. Malayalam Frequency Count (Study report)

Question and Answer time!

Thank you