JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

# Automatic Speech Recognition of Alveolar Rhotic and Retroflex Rhotic Phonemes of Malayalam Language

**Cini Kurian**

Department of Computer Science , Al-Ameen College, Edathala, Aluva ,Kerala

DOI: https://doi.org/10.26438/ijcse/v7i5.12831286 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

***Abstract**:* Development of speech recognition systems in local languages will help anyone to make use of the technological advancement of the speech recognition . In India, speech recognition systems have been developed for many indigenous languages, however very less work has been done in Malayalam Language. Malayalam language is famous for its unique phonemes. Hence one of the main objectives of this work is to explore the Alveolar and Retroflex phonemes of Malayalam language which has unique phonetic realizations.

***Keywords— Automatic Speech Recognition , Malayalam , Phonome***

## I. Introduction

Human beings are comfortable with speaking directly with computers rather than depending on primitive interfaces such as keyboards and pointing devices. The primitive interfaces like keyboard and pointing devices require certain amount of skill for effective usage.. Moreover current computer interface assumes a certain level of literacy from the user. It expects the user to have certain level of proficiency in English apart from typing skill. Speech interface [1 ] helps to resolve these issues. The tantalizing applications of Speech interface have motivated research in automatic speech recognition(ASR) since 1950's. Automatic speech recognition has tremendous potential in Indian scenario

Malayalam is one among the 22 languages spoken in India with about 38 million speakers. It belongs to the Dravidian family of languages and is one of the four major languages of this family with a rich literary tradition. The majority of Malayalam speakers live in Kerala, one of the southern states of India and in the union territory of Lakshadweep. The language has 37 consonants and 16 vowels. There are different spoken forms in Malayalam although the literary dialect throughout Kerala is almost uniform.

Many efforts have been made in other Indian languages however, Malayalam is in its infancy stage in speech recognition research. Since speech technology highly depends on each phoneme , similar sounding phones of a language and unique phonemes of a language have to be explored independently. Hence in this work speech recognition performance of Alvelor Rhotic and Retroflex Rhotic Phonmes have been studied.

## II.VARIOUS DIMENSIONS OF SPEECH RECOGNITION

A speech recognition system's accuracy depends on the condition under which it is evaluated. Under narrow conditions almost any system can attain human-like accurcy, but it is much harder to attain good accuracy under normal environment. Hence the accuracy of any system can vary along with the following dimensions.

- Vocabulary size and confusability

Generally, it is easy to classify a small set of words but as the vocabulary size increases, classification becomes a complex issue. For example, the 10 digit "zero " to "nine" can be recognized essentially perfectly [2], but vocabulary sizes of 200, 5000 and 100000 may have error rates of 3% 7% and 45% [3,4,5] . It is hard to classify the words of a vocabulary which contains confusing words although it is small in size .

- Speaker dependence vs. Speaker independence

Speaker dependent speech recognition system is dependent on knowledge of the speaker's particular voice characteristics. This system learns the characteristics of the speaker's voice through voice training (or enrolment). This type of system must be trained on a specific user before being able to recognize what has been said. This type of system works well if there is only one user speaking to the system and it cannot be used for general purpose. Speaker independent systems are generally able to recognize speech from a variety of contexts, speakers etc. This type of system is used in all general purpose recognizers. A complete speaker independent system, is hard to achieve since it needs rigorous training from all type of speakers (age wise and gender wise), all dialects etc. Error rates are typically 3 to 5 times higher for speaker independent system than for speaker dependent ones [6].

- Isolated, Connected or Continuous speech recognition

Isolated and connected speech recognition is relatively easy because word boundaries are detectable and the words tend to be clearly pronounced. Continuous speech is more difficult because word boundaries are unclear and their pronunciations are more corrupted by co-articulation. In a typical evaluation, the word error rate for isolated and continuous speech were 3% and 9% respectively [7]

- Read vs. spontaneous speech

Speech can be either read from prepared scripts, or that is uttered spontaneously. Spontaneous speech is difficult to recognize, because it tends to be peppered with disinfluencies like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laugher; and moreover, the vocabulary is essentially unlimited. On the other hand, read speech can be handled more easily since it does not contain unexpected words.

- Adverse conditions

A system's performance can also be degraded by a range of adverse conditions [8]. These include environmental noise (eg. noise in a car or a factory); acoustic distortions (e.g. echoes, room acoustic); different microphone (e.g. close-speaking, unidirectional or telephone); limited frequency bandwidth (in telephone transmission); and altered speaking manner ( shouting, whining, speaking quickly, etc.)

## II. SYSTEM DESIGN

Speech recogntion is a special case of pattern recogntion. There are two phases in supervised pattern recognition, viz.training and testing. The process of extraction of features relevant for classification is common for both phases. During the training phase, the parameters of the classification models are estimated using a large number of class exemplars ( training data). During the testing or recognition phase, the features of a test patterns ( test speech data) are matched with the trained model of each and every class. The test pattern belong to that class whose model matches best with the test patterrn . The goal of speech recognition is to generate the optimal word sequence subject to linsuistic costraints. The sentence is composed of linquistic units such as words, syllables, phonemes. In speech recognition a sentence model is assumed to be a sequence of models of such smaller units. The acoustic evidence provided by the acoustic models of such units is combined with the rules of constructing valid and meaningful sentences in the language to hypothesis the sentence. Therefore, in the case of speech recognition, the patteren matching stage can be viewed as taking place in two domains : acoustic and symbolic. In the acoustic domain , a feature vector corresponding to a small segment of test speech ( called a frame of speech ) is matched with the acoustic model of each and every class. The segment is assigned the label of the class with the highest matching score. This process of label assignment is repeated for every features vector in the feature vector sequence computed from the test data.The resultant sequence of labels are processed in conjunction with the language model to yield the recognized sentences.

### Context-Independent Models

The initial prototype of each context-independent monophone is represented as a hidden Markov model (HMM) of 3 emitting states with left-to-right topology with one Gaussian component per state and no skip transitions between states, as can be seen in figure 1

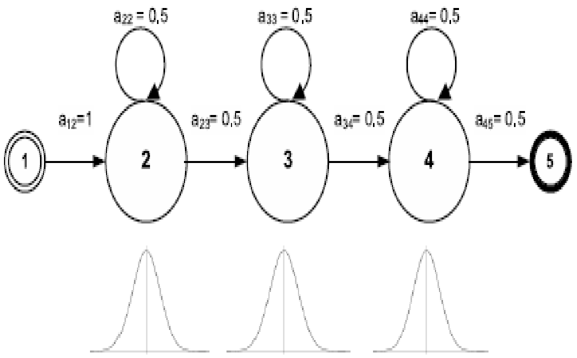


Figure 1

The HMMs are initialized with the flat-start scheme. Then, the parameters of the models are re-estimated in 2 consecutive runs of the Baum-Welch algorithm using the monophone transcription of the training data. Then 2 more iterations of the Baum-Welch algorithm are run. As the pronunciation dictionary contains some words with multiple pronunciations, a new transcription is generated; that matches best with the acoustic evidence by running the Viterbi algorithm over the training data [9] (Jurafsky and Martin, 2008; Young et al.,2002).

We then increased the number of Gaussian components up to the desired number. To increase the number of Gaussian components, the component with the largest mixture weight is cloned, the weight is divided by 2 and the means are perturbed by a small fraction of the standard deviation (typically +/- 0.2σ). The resulting HMMs are then re-estimated with 4–8 consecutive runs of the Baum-Welch algorithm. This is repeatedly done until we have estimated the models with the required number of mixtures.

### Context-Dependent Models

Since context-independent model do not capture phonetic context, their phonetic discrimination ability is poor.

Therefore, in order to achieve good phonetic discrimination, it is better to use triphones where every phone has a distinct HMM model for every unique pair of left and right neighbours.

The single-Gaussian monophone models trained; as described in the previous section are used to generate triphone prototypes. The transition probability matrix is tied across all triphones of a phone. The resulting triphone model parameters are re-estimated with the Baum-Welch algorithm with a triphone list and triphone transcriptions.

When triphones are used, usually training data becomes insufficient as there are too many models whose parameters must be estimated, hence it is necessary to reduce the number of parameters in an HMM. Diagonal covariance assumption and parameter tying are commonly used methods to reduce the number of parameters that ought to be estimated. Tying [10] is a method where two or more states that represents similar acoustic data are clustered together to create tied states. When states are tied, all the data which would have been used to estimate each individual untied parameter are effectively pooled, leading to more robust estimates for the parameters of the tied state[11]. Then decision tree based clustering [12] is used to identify the states that can be tied together. Once we have single-Gaussian,tied-state triphones, the next step is to increase the number of Gaussian mixture components till we arrive an optimal value for GMM..

## III. SYSTEM DEVELOPMENT.

Speech recognition technology is highly dependent on the spoken language. Phonetic and acoustic features of each phoneme in a language is the prime factor for speech recognition technology. . In this study we concentrate on the phonemes such as Alvelor rhotic and and Retroflex rhotic of Malayalam language. For speech recognition , we have used the famous statistical classifier the Hidden markov model and the speech database has 32 selected minimal pairs spoken by 25 speakers. Malayalam has 52 consonant phonemes, encompassing 7 places of articulation and 6 manners of articulation

### ii) Data Base

For conducting speech recognition performance of unique phonemes, we have conducted a special procedure. Initially, the words are recorded with carrier words like " njaan /the word / enn'u paranji' ( i spoke the word /word/). This is to nullify the domination of language model in the speech recognition performance. Then the minimal pair counterpart also recorded with the same carrier words. Then we analyses the speech recognition performance .

### ii) Alveolar rhotic

**A**lveolor rhotic is a unique characteristic of Malayalam language. There are so many paradoxical opinions over the position of alveolor rhotic and retroflex rhotic in phonetic chart among linguists. Since it's phonetic and acoustic character is under confusion among researchers. We have selected minimal pairs with these two phonemes and checked it's speech recognition performance.

### **iii )** Design of Database

We have selected 10 minimal pairs with retroflex rhotic as detailed below. Duration, acoustic properties and phonetic characteristic of a phoneme vary with the succeeding sound. Hence to make a better analysis of the target phoneme, we have designed three categories of minimal pairs of database. First category of phoneme has vowel 'a' succeeding it. Second category has vowel 'i' as succeeding sound and the third category has vowel 'u' as succeeding sound.

a) Words following vowel 'a'

- പാ റ , പാ ര (/paara/ - rock , / paar’a / -iron lever )
- പു റ , പു ര ( /pura/ - exteranl , / pur’a / - hut)
- ക റ , ക ര ( / kara-stain , / kar’a/ - land )

b) Words following vowel ‘i’

- ക റി , ക രി ( /Kari/ - curry , / kar’i./ - coal)

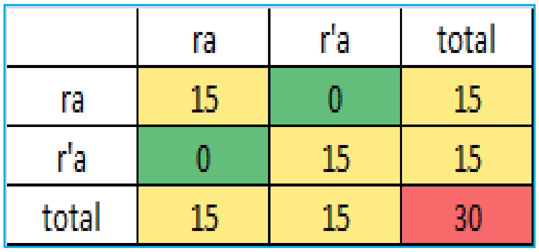
c) Words following vowel u'

- ആ റ്, ആ ര് ( /*aaru’*/ - a number or river , / aar’u’/ - who)

## IV .RESULT ANALYSIS AND DISCUSSION

The six target token as mentioned above have been spoken by 25 speakers. The speech data of 20 speakers were used for training and the remaining 5 speaker data was used for testing. Hence there are 15 tokens of /ra/ and 15 tokens of /r'a/ for testing . The speech recognition performance of these were analyzed and represented as confusion matrix in table 1.1

Table 1.1 Confusion matrix of speech recognition performance of /ra/ vs /r'a/



| | ra | r'a | total |
|---|---|---|---|
| ra | 15 | 0 | 15 |
| r'a | 0 | 15 | 15 |
| total | 15 | 15 | 30 |

In speech recognition performance /ra/ and /r'a/ do not have any confusion for recognition. In phonetic chart these two phonemes (Alvelor rhotic and retroflex rhotic ) has been placed adjacent to each other - which has been challenged by many researchers [13,14 ] . Our analysis also in favour of these - i.e. they both have no common features and cannot not be placed closer in phonetic chart.

## References

[1] Sorin Dusan and Larry R. Rabiner, "On integrating insights from human speech perception into automatic speech recognition," in Proceedings of INTERSPEECH 2005, Lisbon, 2005.

[2] HILL, D. R. (1971). Man-machine interaction using speech. In Advances in Computers, 11. Eds F. L. Alt, M. Rubinoff & M. C. Yovitts, pp. 165-230. New York: Academic Press.

[3] Balaji. V., K. Rajamohan, R. Rajasekarapandy, S. Senthilkumaran,"Towards a knowledge system for sustainable food security: The information village experiment in Pondicherry," in IT Experience in India : Bridging the Digital Divide, Kenneth Keniston and Deepak Kumar, eds., New Delhi, Sage,2004.

[4] G. Doddington, (1989), "Phonetically Sensitive Discriminants for Improved Speech Rec.", Proc. IEEE Int Conf. Acoustics. Speech and Sig. Proc., ICASSP-89, pp. 556-559, Glasgow, Scot- land.

[5]Itakura F (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoustics Speech Signal Process ASSP 23:52–72

[6] Miyatake, M. Sawai, H., & Shikano, K. (1990). Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990.

[7] Kimura, S. (1990). 100,000-Word Recognition Using Acoustic-Segment Networks. In Proc.IEEE International Conference on Acoustics, Speech, and Signal Processing.

[8]K.-F. Lee, Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system, Ph.D. Thesis, Carnegie Mellon University, 1988.

[9] Jurafsky, Daniel, and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

[10] Bahl, L. R. *et al.* (1978). Automatic Recognition of Continuously Spoken Sentences from a Finite State Grammar. *In Proc ICASSP*, pp. 418-421.

[11] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D.,Valtchev, V. and Woodland, P. (2002). T*he HTK Book (for HTK Version 3.2).*Microsoft Corporation and Cambridge University Engineering Department,England.

[12] Young (1996). "Large Vocabulary Continuous Speech Recognition." IEEE Signal Processing Magazine 13(5): 45-57

13] Punnoose, R. (2010). *An Auditory and Acoustic Study of Liquids in Malayalam.* Ph.D. Thesis, Newcastle University, Newcastle, UK

[14] J. Holmes (1988). *Speech synthesis and recognition*. Van Nostrand Reinhold (UK) Co. Ltd., Wokingham.