Comparison of Palateral, Retroflex and Alvelor Lateral in Automatic Speech Recognition

Cini Kurian[1]

[#] Associate Professor, Al-Ameen College, Edathala, Aluva, Kerala, India

Abstract

Speaking with the machine to achieve desired task , make the modern devices easier and convenient to use. Although may interactive software applications are available, the use these applications are limited due to language barriers. Hence development of speech recognition systems in local languages will help anyone to make use of this technology. In this paper Speech Recognition performance of three important phonemes of Malayalam Language – Palateral Lateral , Retroflex lateral and Alvelor Lateral have been analyzed.

Keywords — *Malayalam , Automatic Speech Recognition.*

I. INTRODUCTION

Automatic speech recognition has tremendous potential in Indian scenario. Although literacy rate of India is above 65%, less than 6% of India's total population uses English for communication. Since the internet has become universal, common man now mainly depend the same for any sort of information and communication. Therefore it is imperative that the about 95% of our population cannot enjoy the benefits of this internet revolution. If these information is available in local languages, India could also be benefited by this technology revolution and could stand along with developed countries.

It would be a vital step in bridging the digital divide [1] between non English speaking people and others. Since there is no standard input for Indian languages, it eliminates the key board mapping of different fonts. In Indian scenario, where there are about 1670 dialects of spoken form, speech recognition technology has wider scope and application.

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.

Speech recognition system keeps elderly, physically handicapped and blind people closer to the Information technology revolution. Speech recognition benefits a lot in manufacturing and control applications where hands or eyes are otherwise occupied. It has large application for use over telephone, including automated dialing, telephone directory assistance, spoken database querying for novice users, voice dictation systems like medical transcription applications, automatic voice translation into foreign languages etc. Speech enabled applications in public areas such as; railways, airport and tourist information centers might serve customers with answers to their spoken query

II MOTIVATIONS

Peri Bhaskararao, Tokyo University of Foreign Studies, Tokyo, Japan in his paper titled "Salient phonetic features of Indian languages for Speech Technology" commended that *"Developing a speech recognizer in any language requires a through acoustic and phonetic study. However, Malayalam language, well-known for its rich and unique phonemes, no such studies have been conducted,"* [2]. This references was one of the motivations to do this research in speech recognition of these phonemes of Malayalam language.

The following issues has been identified in speech recognition research, especially in Malayalam language, which inspired us to focus the work on ASR

a) Diversity of phonetic realization: For most of literate languages, phonemes and letters in their scripts have varying degrees of correspondence[3]. Since such a relationship exists, a major part of a speech technology deals with the correlation of script letters with time-varying spectral stretches in that language. Indian languages said to have more direct correlation between their sounds and letters. Such similarity gives a false impression of similarity of text-to-sound rule across these languages. A given letter which is parallel across various languages may have different degrees of divergence in its phonetic realization in these languages.

III. LITERATURE SURVEY

.

Designing a machine that converse with human, particularly responding properly to spoken language, has intrigued engineers and scientists for centuries. Today speech technology enabled applications are commercially available for a limited but interesting range of tasks. Very useful and valuable services are provided by these technology enabled machines, by responding correctly and reliably to human voices. In order to bring us closer to the “Holy Grail” of machines that recognize and understand fluently spoken speech, many important scientific and technological advances have been took place, but still we are far from having a machine that mimics human behavior.

Speech recognition technology has become a topic of great and interest to general population, through many block buster movies of 1960's and 1970's[4]. The anthropomorphism of "HAL", a famous character in Stanley Kubrick’s movie “2001: A Space Odyssey”, made the general public aware of the potential of intelligent machines. In this movie, an intelligent computer named “HAL” spoke in a natural sounding voice and was able to recognize and understand fluently spoken speech, and respond accordingly. George Lucas, in the famous Star Wars saga, extended the abilities of intelligent machines by making them intelligent and mobile Droids like R2D2 and C3PO were able to speak naturally, recognize and understand fluent speech, move around and interact with their environment, with other droids, and with the human population.

Apple Computers in the year of 1988, created a vision of speech technology and computers for the year 2011, titled “Knowledge Navigator”, which defined the concepts of a Speech User Interface (SUI) and a Multimodal User Interface (MUI) along with the theme of intelligent voice-enabled agents. This video had a dramatic effect in the technical community and focused technology efforts, especially in the area of visual talking agents[5][6] .

Languages, on which so far automatic speech recognition systems have been developed are just a fraction of the total around 7300 languages. Chinese, English, Russian, Portuguese, Vietnamese, Japan, Spanish, Filipino, Arabic, Bangali, Tamil, Malayalam, Sinhala and Hindi are prominent among them[7].

When the research tries to develop certain recognition system it requires certain previously stored data i.e. database for respective recognition system. There are various speech databases available for European Language but very less for Indian Language. Various speech database developed in different Indian Languages for speech recognition technology are also being discussed.

IV. THEORATICAL FRAME WORK OF THE METHODOLOGIES USED

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words, independent of the device used to record the speech (i.e., the transducer or microphone), the speaker, or the environment.

It is assumed that the speaker decides what to say and then embeds the concept in a sentence, W, which is a sequence of words (possibly with pauses and other acoustic events such as uh’s, um’s,er’s, etc.) The speech production mechanisms then produce a speech waveform, $s(n)$, which embodies the words of W as well as the extraneous sounds and pauses in the spoken input. A automatic speech recognizer attempts to decode the speech, $s(n)$, into the best estimate of the sentence, $\widehat{W}$, using a two-step process, as shown in Figure 1[8].

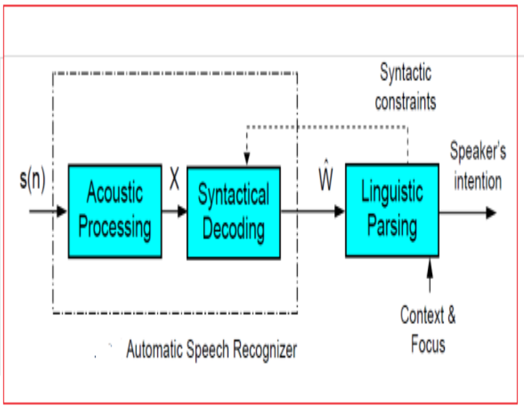

Figure 1 - ASR decoder from speech to sentences

The first step in the process is to convert the speech signal, $s(n)$, into a sequence of spectral feature vectors, X, where the feature vectors are measured every 10 ms (or so) throughout the duration of the speech signal. The second step in the process is to use a syntactic decoder to generate every possible valid sentence (as a sequence of orthographic representations) in the task language, and to evaluate the score (i.e., the a posteriori probability of the word string given the realized acoustic signal as measured by the feature vector) for each such string, choosing as the recognized string, $\widehat{W}$, the one with the highest score. This is the so-called maximum a posteriori probability (MAP) decision principle, originally suggested by Bayes[9,10,11][.

Mathematically, we seek to find the string $\widehat{W}$ that maximizes the a posteriori probability of that string, given the measured feature vector O, i.e.,

$$\widehat{W} = arg_{w}\ max\ P(W|O) \qquad (1)$$

Using Bayes Law, we can rewrite this expression as[12].

$$\widehat{W} = arg_w \max \frac{P(O|W)P(W)}{P(O)} \qquad (2)$$

Thus, calculation of the a posteriori probability is decomposed into two main components, one that defines the *a priori* probability of a word sequence *W*, *P*(*W*), and the other the likelihood of the word string *W* in producing the measured feature vector, *P*(*O*|*W*). (We disregard the denominator term, *P(O)* , since it is independent of the unknown *W)* . The former is referred to as the Acoustic Model, *P*(*O*|*W*)., and the latter the Language Model, *P*(*W*). These quantities are not given directly, but instead are usually estimated or inferred from a set of training data that have been labelled by a knowledge source, i.e., a human expert. The decoding equation is then rewritten as [13]

$$\widehat{W} = arg_w = arg_w max\ P(O|W)P(W) \quad (3)$$

We explicitly write the sequence of feature vectors (the acoustic observations) as:

$$O = o_1 , o_2 , o_1 ,\ldots\ldots o_N \qquad (4)$$

where the speech signal duration is *N* frames (or *N* times 10 msec. when the frame shift is 10 msec). Similarly we explicitly write the optimally decoded word sequence as:

$$\widehat{W} = w_1\, w_2 \quad w_3 \ \ldots\ldots w_M \qquad (5)$$

where there are *M* words in the decoded string. The above decoding equation defines the fundamental statistical approach to the problem of automatic speech recognition.

Probabilities for word sequences are generated as a product of the acoustic and language model probabilities. The process of combining these two probability scores and sorting through all plausible hypotheses to select the one with the maximum probability, or likelihood score, is called decoding or search..

V . DATA BASE DESIGN

i) Palatel Lateral_ഴ

The Palatal lateral phonemes occurs in American English, Irish English, Western Countries dialects, Mandarin Chinese, Pashto, a few Brazilian Portuguese dialects and some languages in India such as Tamil and Malayalam, as well as several Australian Aboriginal and indigenous South American languages .

Minimal pairs for the study of this phoneme has been designed in three categories as shown below. We have 13 minimal pairs in 3 categories. Palateral lateral vs. retroflex lateral has 3 pairs, Palateral lateral vs alvelor lateral has 4 pairs and Retroflex lateral vs. alvelor lateral category has 6 pairs.

a) Palateral lateral vs retroflex lateral

- കഴം , കളം (/kazham/ - pool , / kal'am / - yard)
- അഴി , അളി (/azhi/ - destroyed , / al'i / - beetle)
- കോഴ , കോ ള (/ ko'zha/ - bribe , /ko'l'a / - a drink)

b) Palateral lateral vs alvelor lateral

- കോഴ , കോല (/ko'zha/ - bribe, ko'la - verandah)
- വ ഴി , വ ലി (/vazhi/ -way , / vali - pull)
- തൊഴി , തൊ ലി -(/tozhi/- kick , /toli / - skin)
- കഴ , കല - (/kazha/- a long stick , / kala / - male deer)

c) **Retroflex lateral vs alvelor lateral**

- ക ലി , കളി (/kali/ - irritation , / kal'i/ - play)
- വാല് , വാള് (/vaalu'/ -tail , /vaal'u'/ - sword)
- കല , കള (/ kala/ -male deer , /kal'a/ - sweet but indistinct)
- വല , വ ള (/vala/ -net , /val'a/ - bangle)
- നില , നി ള (/nila/ - position , /nil'a / - river)
- നാ ല് , നാള് (/naalu'/ - four, / naal'u' / - day)
- കോല , കോള (/ko'la/- verandah , /ko'l'a / - name of a drink)

VI. SPEECH RECOGNTION PERFORMANCE

Retroflex Lateral vs. Palatel Lateral vs. alveolar lateral :

Speech recognition performance have been carried out with the above referred six tokens. Test data includes total of 30 tokens spoken by five speakers. The result has been reported with confusion matrix as shown in table 1. It is clear from the table that 10% of /la/ confuses with /l'a / and 20% of /l'a/ confuses with /zha/ .

Table 1: Confusion matrix - speech recognition performance of /la/ vs /l'a/ vs /zha/

	la	l'a	zha	total
la	9	1	0	10
l'a	0	8	2	10
zha	0	0	10	10
total	9	9	12	30

The phoneme /la/ confuses with /l'a/ in 10% and the phoneme /l'a/ confuses with /zha/ in 20% . Hence it can be concluded that the phonetic nature of /zha/ which is a very peculiar phoneme of Malayalam language need more investigation and deep analysis from the linguistic point of view to assess its exact nature. Hence this study opens an area for future researchers.

REFERENCE

[1] 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.

[2] Bhaskararao P., "Salient phonetic features of Indian languages", Sadhana, 36(5), pp. 587-599, Oct. 2011.

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

[4] B. H. Juang & Lawrence R. Rabiner. Automatic Speech Recognition -- A Brief History of the Technology Development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara.

[5] 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.

[6] Furui, S., "50 Years of Progress in Speech and Speaker Recognition Research Identification", In ECTI Transformations on Computer and Information Technology, vol. 1, no. 2, 2003

[7] *Wiqas Ghai* , Navdeep Singh *"Literature Review on Automatic Speech Recognition"* International Journal of Computer Applications (0975 – 8887) Volume 41– March 2012.

[8] S. Young (1999). Acoustic Modelling for Large Vocabulary Continuous Speech Recognition. Computational Models of Speech Pattern Processing: Proc NATO Advance Study Institute. K. Ponting, Springer-Verlag: 18-38

[9] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.

[10]] L.,R Rabiner, "A tutorial on Hidden Markov model and selected application in speech recognition" , Pro.IEEE,7(2):257-286, February 1998

[11] Hwang, M. Y. (1993). Sub-phonetic acoustic modeling for speaker-independentcontinuous speech recognition. Ph.D. Thesis. Carnegie Mellon University.

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

[13] L.,R Rabiner, "A tutorial on Hidden Markov model and selected application in speech recognition" , Pro.IEEE,7(2):257-286, February 1998