National Conference on Emerging Trends in Computer, Electrical & Electronics (ETCEE-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406,Impact Factor:3.134

A review on selection and correction of text and speech data for Indian languages

Hiral Bharad[1], Trupti Kodinariya[2]

[1,2]*Atmiya Institute of Technology and Science,*

Kalawad road, Rajkot

[1]hir.bharad@gmail.com

[2]tmkodinariya@aits.edu.in

Abstract - Today the word is going towards hands-free interfacing with machine using speech commands and/or speech recognition. The computer understands binary or machine level language and human being communicate in their own different - different languages so to make machine capable of understanding the language behavior their respective language model is been developed. To develop the language model this paper focuses on the selection and correction of text and speech data for some of the Indian languages like Hindi, Indian English, Tamil, Telugu, and Marathi.

Keywords: hands-free interfacing, language model, speech recognition, text and speech data.

I. INTRODUCTION

A speech recognition is to make machine capable of recognizing speech of human being and work according to that or converting the speech into the text transcript. Any speech recognizer needs two models for this type of conversion: one is acoustic model and second is language model. Acoustic model is to describe voice property of spoken data and language model is to represent language behavior in terms of the phonetic structure of the words. The recognizer needs to be first train for this purpose with efficient and enough data. Here this paper focused on the data selection and correction method for some of the Indian languages like Hindi, Indian English, Tamil, Telugu and Marathi.

Any recognizer in the world up till now can't guarantee to recognize all the existing languages in the word so this paper describes method for above language. Moreover, for any particular language the recognizer is made up for the specific domain.

This paper basically is a review paper of two papers which describes method of data selection and correction process of the text and speech data of different languages: first paper [1] describes this method for the language Indian English and Hindi and second paper [2] describes the same for the Tamil, Telugu and Marathi.

II. METHOD OF WORKING

The paper from KIIT College of engineering Gurgaon and Nokia research Centre, China which describes the selection and correction method for the Hindi [1][2] and Indian English[3] has the following details:

The paper describes the method for the mobile communication that is to say that the

recognition process was for mobile communication environment.

The data for the 630 phonetically reached sentences was gathered from 1163 Hindi participants and 1405 English participants by the 13 different prompt sheets [4].

They have considered age wise distribution of the participants in which approximately more than 20% speakers have 15-21 year age 60% have age of 22-50 year and rest are above 50 year age. The distribution is shown in below figure 1.

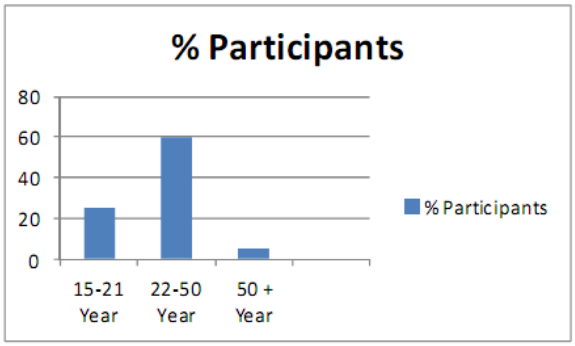

Figure 1: Age distribution of speakers [1]

The social media like Facebook and Orkut was included as a medium of data collection. The speakers have necessarily their first language as Hindi but not necessary English.

To create phonetically rich sentences words were selected from the text corpus based on their frequency of occurrence in the whole text corpus. Meaningful sentences were framed using these words which were further manually corrected based on language grammar rules.

The data correction for Hindi was based on KTRANS software which receives input from the QWERTY keyboard in roman and converts the text into Hindi using Devnagri script.

For English language they have manually corrected slang words like u, plz etc.

Then the text is given tagging like first name = FN and currency = C etc. and expansion of punctuation marks, special symbol, days, month time etc. was done.

The paper from IIIT Hyderabad and Hewlett Packard labs India which describes the selection and correction method for the Tamil, Telugu and Marathi has the following details [5]:

The data of 560 speakers for the three languages are collected based on speakers age, gender and dialects. The speakers from 18 year age to more than 60 year are selected. The age distribution is shown in below figure 2.

Language	18-30	30-40	40-50	50-60	> 60
Marathi	77	56	26	12	5
Tamil	80	28	31	16	45
Telugu	52	52	32	32	15

Figure 2: Age distribution of the speakers [2]

The selection process was done using Optimal Text Selection (OTS) algorithm which is greedy based algorithm that selects phonetically reached sentences. Moreover, Threshold based approach was implemented on OTS to select new diphone from new sentence.

Approximately half of the total data for three languages was taken from Landline and another half was selected from cellular mobile. This distribution is shown in below figure 3.

Language	Landline	Cellphone	Total
Marathi	92	84	176
Tamil	86	114	200
Telugu	108	75	183

Figure 3: Data selection for languages [2]

The correction process was done manually, from the all recordings the unnecessary utterance had been removed and transcript file was edited to match with speech file. Data was categorized by 'Good', 'With Channel distortion', 'With Background Noise' and 'Useless' whichever was appropriate.

III. COMPARISON AND CONCLUSION

The main aim of this paper is to focus method of speech and text data selection and correction so that while building efficient language model for very large vocabulary or data set for the efficient speech recognizer one has clear cut idea how to make it. The comparison is between five different Indian languages which were described in two different papers.

The paper compares the method of selection and correction so there are these two different parameters to compare these papers. The data selection method for Tamil, Telugu and Marathi [2] is better than Hindi and Indian English [1] because it has taken age wise different people's voice and two medium like cellular telephone and landline to gather data.

The correction process is better in Hindi and Indian English [1] data as they have divided data into different domains and had tagging system at the time of final expansion.

IV. FUTURE WORK

In future, one can think to combine this two approaches in such a way that to make efficient data selection one have to have age wise different people's voice and for correction of data one can choose tagging and expansion type of system as Hindi and Indian English [1] languages have.

V. REFERENCES

[1] ShwetaSinha, S.S. Agrawal, Jesper Olsen (2011) Development of Hindi mobile communication text and speech corpus, Proceedings of O-COCODSA- 2011

[2] KaruneshArora, SunitaArora, S SAgrawal, NiklasPaulsson, Khalid Choukri, "Experiences in Development of Hindi Speech Corpora based on ELDA standards" Proceedings of O-COCOSDA 2006.

[3] Project specification for "Personal Communication (PCOM) Text and Speech Data Collection For Hindi/Indian English" prepared by Nokia Research Centre, China 2008-10.

[4] ShyamAgrawal, ShwetaSinha, Pooja Singh, Jesper Olsen – "Development of Text And Speech Database For Hindi And Indian English Specific To Mobile Communication Environment" prepared by KIIT College of Engineering, Gurgaon, Nokia Research Center, China, 2011-2013.

[5] GopalakrishnaAnumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.N.V. Sitaram, S P Kishore – "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems" prepared by IIIT Hyderabad, India, Hewlett Packard Labs India, Bangalore, India.

[6] *Fundamentals of Speech Recognition*; Lawrence Rabiner&Biing-Hwang Juang

Englewood Cliffs NJ: PTR Prentice Hall 015157-2 (Signal Processing Series), c1993, ISBN 0-13-