Sustaining Linguistic Diversity Through Human Language Technology : A Case Study for Hindi

Shweta Sinha, Shyam S Agrawal
KIIT College of Engineering, Gurugram, India
meshweta_sinha@rediffmail.com, ss_agrawal@gmail.com

Abstract

Language is a mean to communicate ideas, knowledge and express our cultural identity. To protect the legacy of our cultural heritage, language diversity needs to be sustained. Human language technology (HLT) can offer a lot to reduce the rate of language extinction. The focus of this paper is towards digital preservation of under-resourced languages. The discussion is apropos to the Indian languages; that almost all are under-resourced. The linguistic diversity of India is highlighted and its fate in this digital era is analyzed. This paper discusses the digital representation of the language and discusses HLT as a step towards preserving languages. Platform for online collection of speech is explained for gathering speech samples in three Indian languages; Hindi, Punjabi and Manipuri. The meta-data highlights the dialectal diversity of the speakers. These diversities have been analyzed acoustically for the Hindi speakers.

Keywords: Language Diversity, Hindi Language , Digital Preservation

1. Introduction

The United Nations Educational, Scientific and Cultural Organization defines that, cultural heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from the past generations, maintained in the present and vouchsafed for the benefit of future generations. The preservation of cultural heritage is not only concerned with safeguarding physical aspects of tradition but, is equally responsible for lesser physical aspect like languages, customs and beliefs. Ever since the existence of civilization the demand for communication exists. Language evolved as a sophisticated medium through which one can express thoughts that influences our society. With time, each culture evolved its own language and huge literary base for their language. In today's globalized world languages are disappearing at a very fast pace. On an average, every two weeks a language dies [1]. Language diversity is the basis of our rich cultural heritage and diversity. Of late, the loss of language diversity has grasped the attention of UNESCO also[2], as with the loss of any language, the memories and experiences of the culture are also lost. It is often observed that the positive impact of language on social, political and economical strata of the society influence the acceptance of the language in the society. Also, the colonial legacy on any country can burden the speakers and the native language with the use of exogenous language in formal and official domain. All these leads to cultural assimilation and usually results in the loss of suppressed language in years to come.

In this digital era, for the linguistic preservation and cultural redemption technology development and digital representation has become the sine qua non. Out of approximately six thousand languages of the world merely a fraction is digitally represented and efforts have to be made to reduce the exacerbation of digital divide. Hence, technology is essentially required for all and every language of the world in order to slow down its extinction. Human language technology development can offer a lot for reinvigorating and documenting any language. Till date these technological developments have been confined to the developed languages only. Technology development for any language can make life easy for its users and raise their interest for its use. Automatic speech recognition, speech to text synthesis and translation system based applications for any language can help in the growth of a language and facilitate access to textual and audio contents of the language.

Attention towards HLT is much needed for preserving under-resourced languages or other languages of the developing countries. This paper focuses on the efforts towards the preservation of some of the low resourced Indian languages, Hindi being the pivot of discussion. The paper describes unity in language diversity in context to India, section 3 presents the status of digital representation of Indian languages. Technology for sustaining the language diversity is explained in section 4. Technology development for Hindi language as a case is presented in section 5, and section 6 concludes the paper.

2. Unity in Language Diversity of India

India is a land of varied hues of culture, religion, race and languages. These variations account for the existence of different ethnic groups residing within the sanctum of one single nation. People of India speaks a large number of languages that can be divided into four families as: the Indo-European, Dravidian, Austro-Asiatic, and the Sino-Tibetan Family[10]. 73% of the Indian population speak one of the languages of Aryan group; a subgroup of Indo-European family[10]. Table 1 presents the language and speaker population of major Indian languages. Dravidian languages are spoken by 20% of the population and merely a small population speaks the languages from other two language families.

Sl. No	Name of Languages	Language Family	Speaking Population (millions)
1	Hindi	Indo-Aryan	422
2	Bengali	Indo-Aryan	83
3	Tamil	Dravidian	60.7
4	Marathi	Indo-Aryan	71.9
5	Telugu	Dravidian	74
6	Urdu	Indo-Aryan	51
7	Oriya	Indo-Aryan	33
8	Gujrati	Indo-Aryan	46
9	Punjabi	Indo-Aryan	29
10	Malyalam	Dravidian	33

Table1. A microcosm of linguistic diversity of India

In total, there are 122 major languages(spoken by more than 10K population), around 1600 distinct dialects along with 13 different scripts for writing. Out of all, Sanskrit is the most ancient language and is considered as the mother of most of the Indo-Aryan languages. It is the only language that transcended the region and boundaries of North and South India. Hindi the major language of India and has evolved from Sanskrit. Apart from this the country has developed highly sophisticated languages that mark India as a unique subcontinent that foster multiple cultures. The proverb **kos kos pər bədəlɛ pa**:**ni**:, **cha**:**r kos pər bədəlɛ va**:**Ni**:.explains that the Indian population thrive to the diversity. This can be translated as: every mile, the taste of water changes; and every four miles, the dialect changes". The unity in cultural and linguistic diversity of the country is very aptly conveyed through this proverb.

3. Digital Representation of Indian Languages

India, a land of multitudes has 30 languages that are spoken by more than a million native speakers. Many languages of the country have died in last few decades. The loss may be due to existence of fewer numbers of their native speakers, non-existence of any documentary evidence for the language, or response to new domains or media in that language. According to UNESCO's "Atlas of the world's languages in danger (2009)" [3] most of the Indian languages are vulnerable, i.e. they are mainly spoken inside the house and restricted in a particular domain. India has the largest number of endangered languages in the world[3]. Figure 1 represents the statistics of the languages of India. The categories mentioned are mainly based on the intergenerational transmission of the languages.

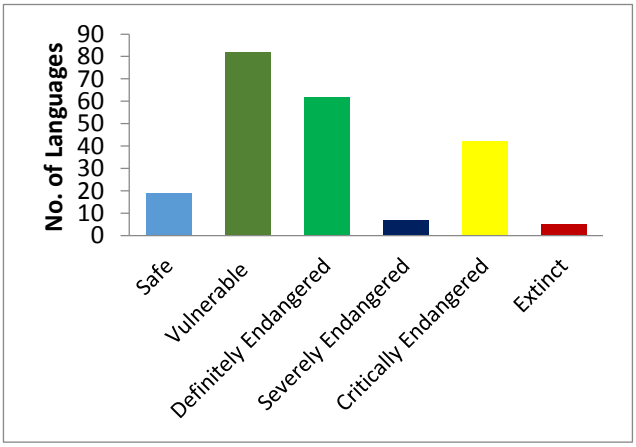

Figure 1. State of Indian Language Existence

One way to safe guard these languages is to provide more and more web resources in these languages along with technology development for human interaction. The internet user base has grown many folds in last few years. Survey by KPMG and Google show that there are 234 million Indian language internet users as compared to 175 million English internet users[4]. And the user base is expected to rise at an alarming 18% rate[4] which will generate demand for digital resources. Predicted internet user base by 2021 is represented in Figure 2 (Source: [4]). Three Indian languages, Hindi, Punjabi, and Bangla are among world's top 10 most widely spoken languages [5]. But, none of these find their place in the top ten languages on the web [6].In general the users of the web consider local language digital content to be more reliable.

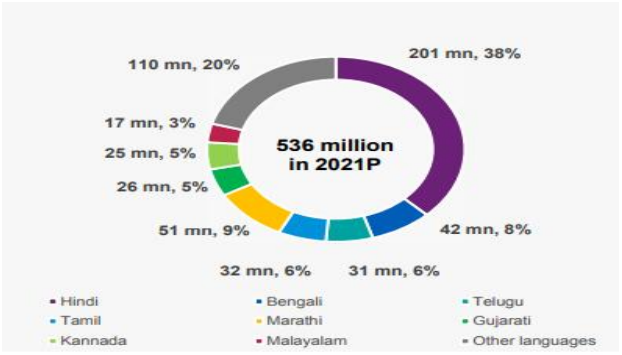

Figure 2. Internet users by native language

Limited language support either in terms of technology or content may force them to move towards English or other developed country's language. The under-representation of the Indian languages on the internet may leave them behind in the race of technological development for natural language processing techniques. Also the digital growth of the spoken languages by the means of language based applications will undoubtedly provide inclusive growth to the society and transform India into a digitally empowered country.

4. Technology Development : Efforts Towards Preservation of Language in Digital Era

Lack of digital data for the languages on the web categorizes them as under-represented language. Technology development is essentially required to revive, maintain, preserve and disseminate our traditional languages and in turn protect them from dying. Interface design for these languages will help human to communicate with computer and connect to the globalized world. Automatic speech recognition and text to speech are few of the HL techniques that have the prospects to completely alter the user's perspective for a language. Automatic language translation is the way to remove the human-to-human communication barrier. Applications based on these can make the life easier for the native language users and also document the language.

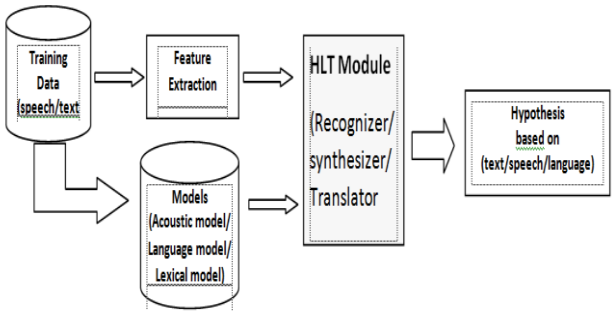

Figure 3. Architecture of HLT based system

5. Towards Technology building for Hindi : A Case Study

Till date HLTs are mainly concerned with the languages with large resources only. Indian languages with very limited digital resources lie far behind in this technological race. In a quest to protect our linguistic heritage from extinction we have initiated a small step towards development of HLT based system; to start with, ASR system for under-resourced languages. Due to lack of annotated text and speech data, nothing much has been achieved in the development of speech based applications for Indian languages. Preparing and gathering language resource is the biggest hurdle in the development of any HLT based system. High-quality ASR performance for the Indian languages is achievable only if real time data for these languages exists. For the growth of language technologies in context to Indian languages, it is essential to have corpus for speech as well as text representing pronunciation lexicon, language dictionaries etc.

Till date the efforts put forward has produced substantial data for Hindi and few other languages. The samples collected for Hindi has been acoustically analyzed for finding the dialectal diversities.

5.1 Data collection for under-resourced languages

The development of any of the HLT based application necessitates the collection of data for the concerned language to be fed as input to the system. Collection of data for the under-represented languages is a complex and tedious task. The situation is more critical when it comes to collecting recorded speech data. Data acquisition in terms of speech sample is very time and labour intensive task. Almost all the languages of India is an under-resourced language. Challenges for collecting spoken samples for these under-represented languages are many. Native speakers of under-represented languages either reside in rural areas or are distributed across a large geographical area. Portability is one of the major concerns for resource generation of these languages. Throughput, stability, latency and cost are the huge prohibitive factor for large-scale data collection. To collect speech samples from native speakers of languages residing in far-off locations we have developed a client-server based multi-lingual online speech collection system [7].

To start with, the data collection has been initiated for three Indian languages: Hindi, Punjabi and Manipuri. Hindi is the most commonly used inter-communication language. Even though this language has a large user base, due to variabilities in the use of this language its detail study with respect to articulation /pronunciation is essential.

The dialects of Hindi are categorized as the Eastern and the Western dialect. Punjabi is the 10th most widely spoken language of the world.It is spoken by people of Punjab region of Pakistan and India. Dialectal diversity exists for this language too. Manipuri, also known as Meitei is the predominant language of the Southeastern Himalayan state of Manipur. Apart from Manipur it is also spoken by people of Assam, Tripura, Bangladesh and Myanmar. This language belongs to Tibeto-Burman family. Approximately, 3 million people in the world speak this language.

Several recording specifications were set for the collection of samples. The collected database and its specifications for these languages have been summarized in Table 2.

Specifications	Indian Languages		
	Hindi	Punjabi	Manipuri
Speaker Registration	100	50	50
Male Speakers	68	27	25
Female Speakers	32	23	25
Dialects Covered	4	2	1
Sentences/Speaker	300	300	300
Isolated Utterances	200	200	200

Table 2. Corpus specification for Indian languages

5.2 Hindi Speech Sample Analysis

The major dialects of Eastern Hindi are Awadhi, Bagheli Bhojpuri and Chhattisgarhi and those of the Western Hindi dialects are Braj Bhasha, Haryanvi, Bundeli, Kannauji and Khari boli. Huge dialectal diversity exists among these varieties. Although there are about 18 classified dialects of Hindi [8]. The sample collected are predominantly from the speakers of these four dialects: Bhojpuri(BP), Bagheli (BG), Khariboli (KB) and Haryanvi(HR). Dialect influences individuals speaking style[9]. Insight into the phonological differences among the dialects can outline the factors that affect the acoustic properties.

5.2.1 Acoustic analysis of vowels in Hindi Dialects

Vowels are more often distorted than consonants in accented speech [9]. Hindi language has 10 vowels, that are categorized as 3 short vowels (/ə/, /I/, /ʊ/) and 7 long vowels (/ɑ:/, /i:/, /u:/, /e:/, /ɛ:/, /o/, /ɔ:/). To measure the dialectal influence on acoustic characteristics of these vowels duration, fundamental frequency, formants and intensity of these vowels were analyzed in reference to Hindi dialects.

The first and second formant analysis for the Hindi vowels w.r.t the four dialects outline that the second formant values for Bhojpuri dialect speakers are higher for back vowels (/ɑ:/, /ʊ/, /u:/, /o/, /ɔ:/) , for Bagheli speakers F2 is higher for all but /ɑ:/ . Haryanvi speakers and Khari Boli speakers have an approximately same value of second formant except for /ɑ:/ , where it is higher as compared to Khari boli speakers. It can be further observed that for the front vowels (/i:/, /I/, /e:/, /ɛ:/) F2 for Bhojpuri speakers are low compared to Khari boli speakers. F1 for Haryanvi speakers are high for close front vowel (/I/, /i:/). F2 value for all front vowels except

for open front vowel (/ɛ:/) is high for Bagheli speakers as compared to speakers of Khari boli dialect.

It has been further observed that prosodic features are more influenced due to dialectal influence. Duration is the highly affected prosodic feature that has been studied. Table 3 summarizes the findings of the acoustic analysis of Hindi samples. Also, to obtain the significance of dialects on the features ANOVA test was conducted on the feature parameters extracted from the samples.

Acoustic Feature	Vowels Influenced	Discrimination of Dialect
F0	Long Vowels (/ɑ:/ /i:/ /u:/ /ɛ:/ ɔ:/)	Significant for all dialect pair
F0	Short Vowels (/I/, /ʊ/, /e:/)	Not significant for any pair of dialect
F1	Back Vowels (/ɑ:/, /ʊ/, /u:/, /ɔ:/)	Significant for BG-KB dialect pair
F1	Front Vowels (/i:/, /I/, /e:/, /ɛ:/)	Significant for KB-BP dialect pair
F1	Close Front Vowels (/I/, /i:/)	Significant for KB-BP and KB-HR dialect pair
F2	Back Vowels (/ʊ/,/o/,/ɔ:/)	Significant for KB-BP dialect pair
F2	Front Vowels (/i:/, /I/, /e:/)	Significant for KB-BG pair
F3	All vowels except(/i:/,/I/,/ʊ/)	Significant for all dialect pair under study
Average Duration	All vowels at different word positions influenced due to dialect.	Significant for all dialect pair; exceptions: BP-BG not important for /ɔ:/; KB-BP not significant for /u:/
Intensity	No major distinction due to dialect	Influence selective in nature. Not significant for most of the vowels in almost all dialect pair under study.

Table 3. Summary of acoustic analysis of Hindi speech samples

5.2.2 Model Creation and automatic recognition of Speech

Based on the above analysis the features that are able to distinguish the utterances can be identified. The steps for the development of ASR requires Feature extraction, acoustic and language model and decoding techniques. These are the future work that has to be executed for the collected data.

Feature extraction techniques for the extraction of speech features have to be employed further. Using these features acoustic model need to be built. These can be Gaussian mixture models or Hidden Markov models . Recognition techniques need to be devised for the identification of utterances.

6. Conclusion

Language diversity in the world signifies the richness of our cultural heritage, To protect our culture and heritage the diversity of tongue has to be protected. Languages of developing countries are under-represented and low resourced languages of the world and are always in danger of getting lost in the coming days. The paper presents the status of Indian languages on the internet. Applications in the area of HLT has been shown as a way to protect the dying or the under resourced languages. A case study of Hindi is discussed in this paper to highlight our efforts towards sustaining the language diversity in this digital era.

7. References

[1] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100.

[2] Unesco:http://www.unesco.org/new/en/indigenous-peoples/cultural-and-linguistic-diversity/[Accessed on: 24-12-17]

[3] Moseley, Christopher (ed.). 2010. Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version: http:// www.unesco.org/ culture/en/endangeredlanguages/atlas

[4]https://assets.kpmg.com/content/dam/kpmg/in/pdf/ 2017/04/Indian-languages-Defining-Indias-Internet.pdf

[5] S. Arora, K. K. Arora, M. K. Roy, S. S. Agrawal, and B. Murthy, (2016). Collaborative speech data acquisition for under resourced languages through crowdsourcing," Procedia Computer Science, vol. 81, pp.37-44.

[6]Online : ttp://www.internetworldstats.com/stats7.htm/, [Accessed 5-March-2017].

[7] S Sinha, S Sharan, S S Agrawal,(2017). O-MARC: A multilingual online speech data acquisition for Indian languages, Oriental-COCOSDA , Nov 1-3, 2017, held at Seoul, S Korea.

[8] Arora, K. Arora, S. Agrawal, S. S. Paulsson, N. and Choukri, K. (2006). Experiences in Development of Hindi Speech Corpora based on ELDA standards, Oriental-COCOSDA 2006 held at Penang, Malaysia.

[9]. John C Wells. (1982) Accents of English, volume 1. CambridgeUniversity Press.

[10]Smriti Chand, http://www.yourarticlelibrary.com/ language/indian-languages-classification-of-indian-languages/19813