Development of Text And Speech Database For Hindi And Indian English Specific To Mobile Communication Environment

[1] Shyam Agrawal, [2] Shweta Sinha, [3] Pooja Singh, [4]Jesper Olsen

[1,2,3]KIIT College of Engineering, Gurgaon, [4]Nokia Research Center, China
ss_agrawal@hotmail.com, meshweta_7@rediffmail.com, poojasingh5881@gmail.com, jesper.olsen@nokia.com

Abstract

This paper describes the method and experiences of text and speech data collection in mobile communication in Indian English Hindi. The primary data collection is done in the form of large number of messages as part of Personal communication among natives of Hindi language and Indian speakers of English.

To gather the versatility of mobile communication database among Hindi and English, 12 domains were identified for collection of text corpus from speaking population belonging to deferent age groups, sex and dialects. The text obtained in raw form based on slangs and unconventional grammar were cleaned using on language grammar rules and then tagged and expanded to explain context specific meaning of the words. Texts of 1163 participants from Hindi speaking regions and 1405 English users were taken for creating 13 prompt sheets; containing 630 phonetically rich sentences created using a special software. Each prompt sheet was recorded by at least 7 users simultaneously in three channels and recorded by a total of 100 speakers and annotated. The work is a step forward in the direction of development of standards for mobile text and speech data collection for Indian languages.

Keywords - Speech data base, Text analysis, mobile communication, Hindi and Indian English Speech, multi-lingual speech processing.

1. Introduction

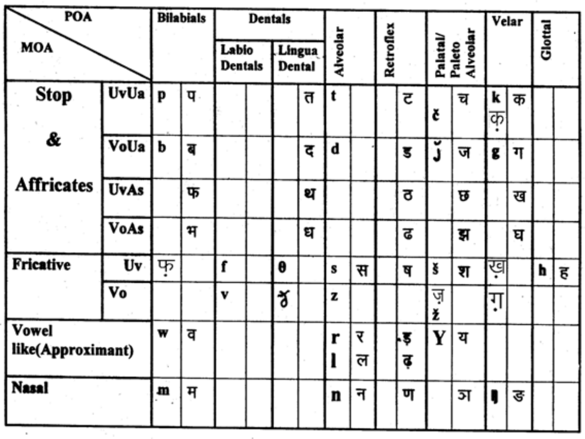
Mobile communication has significantly increased the pace of communication leading to faster message transmission in lesser time. The Multilingual and Multimodal communication with user friendly devices are in increasing demand. Communication in ones own language is becoming the necessity which should take place in the form of speech as well as in the form of written text particularly short messages. These applications help in breaking language barriers among the people of different parts of world. For all these applications it is essential to develop speech corpora to enable the researchers to study the acoustic and linguistic properties of speech and to develop models for speech recognition and synthesis. The situation in mobile communication is not same as in normal communication. With this objective in mind we have worked on speech and text collection for Hindi and Indian English for mobile communication applications. So far, very little or no effort has been made for development of corpora in these languages related to the short messages used by people for personal communication. These messages possess very different characteristics and uses slang based grammar. The understanding and recognition of these words would certainly enhance the usage of mobile phones by providing quality information at reduced cost. The language used for communication is often a cluster of many languages. Indian communication text is very often a mixture of the main language with other languages. For Example in Hindi, words from English, Urdu and Arabic are often used and vice versa for example. These databases will help in developing new language models for speech recognition, synthesis, translation ,language Identification purpose in mobile communication environment and hence a step towards breaking the language barrier using mobile phones, Tablets & PDAs through speech.

KIIT College of engineering with the support of Nokia Research Centre, China has undertaken a project for the development of a comprehensive text and speech corpus for Hindi and Indian English in the personal communication domain i.e. messages composed on mobile phones and other mobile devices. The details of development and experiences of these databases for Hindi have been described in detail in another paper[1]. The details of database collected for Indian English and its comparison with Hindi have been described in the present paper. A total of more than 100,000 messages in 12 different domains from persons of different age groups were collected from the mobile phone users speaking these languages. More than 1200 participants from different Hindi speaking regions and English speaking persons in India were recruited to contribute messages. The text was obtained in raw form which includes slang, unconventional grammar and spellings as is common in mobile messaging (SMS). The textual information after cleaning and expansion was processed to create phonetically rich sentences. These sentences were recorded by 100 speakers for Hindi from different demographic profiles (age & sex) and dialects. For recording of Indian English 100 speakers in total were selected from every part of India. These recordings have been done through 3 channels simultaneously. The work is also a step forward in the direction of development of standards for mobile text and speech data collection for Indian languages. This text and speech database is available for further development of Hindi and Indian English speech recognition, speech synthesis and speaker recognition in the mobile domain. The problem executed and the experiments gained during this work have been described.

2. Text data collection and corpus design

2.1 Phonetic Lexicon of Hindi and English

The text database of Hindi and English is collection of words taken from messages used in mobile communication. The speech database is the collection of utterances, recorded from the above collected text. To read the written text and understand the spoken utterances knowledge of phonemes and their sounds are very important. Table 1 gives a brief outline of phonetic lexicon of Hindi and Indian English consonants similarly the phonetic lexicon of Hindi and Indian English vowels is shown in figure 2.

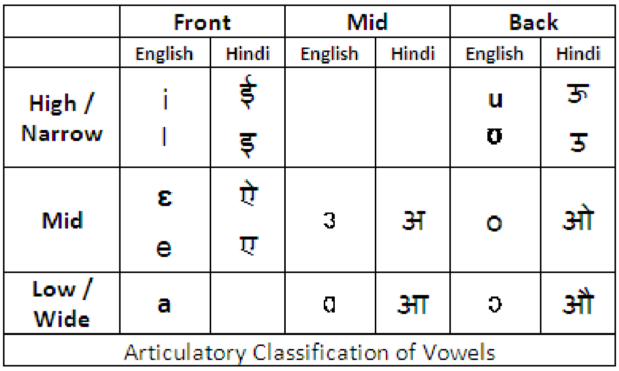


POA MOA		Bilabials		Dentals				Alveolar		Retroflex		Palatal/ Paleto Alveolar		Velar		Glottal	
				Labio Dentals		Lingua Dental											
Stop & Affricates	UvUa	p	प				त	t			ट	č	च	k क़	क		
	VoUa	b	ब				द	d			ड	ǰ	ज	g	ग		
	UvAs		फ				थ				ठ		छ		ख		
	VoAs		भ				ध				ढ		झ		घ		
Fricative	Uv	फ़		f		θ		s	स		ष	š	श	ख़		h	ह
	Vo			v		ð		z				ज़ ž		ग़			
Vowel like(Approximant)		w	व					r l	र ल		ड़ ढ़	Y	य				
Nasal		m	म					n	न		ण		ञ	ŋ	ङ		

MOA=Manner of Articulation
VoUa=Voiced Unaspiration
POA=Place of Articulation

UvUa=Unvoiced Unaspirated
VoAs=Voiced Aspirated
UvAs=Unvoiced Aspirated

Table 1: Phonetic Lexicon of Hindi and Indian English



	Front		Mid		Back	
	English	Hindi	English	Hindi	English	Hindi
High / Narrow	i ɪ	ई इ			u ʊ	ऊ उ
Mid	ε e	ऐ ए	ɜ	अ	o	ओ
Low / Wide	a		ɑ	आ	ɔ	औ
Articulatory Classification of Vowels						

Table 2: Phonetic Lexicon of Hindi and Indian English

The text database for Hindi and Indian English was designed keeping in mind the communication among people of different age groups. For this few domains of communication were identified which were observed to be most frequent in mobile communication. These domains helped to capture the versatility of communication. The 12 communication domains namely 'Vacation report', 'Change of Plans', 'Family Communication', 'Invitation', 'Congratulation', 'Travel Plans', 'Business', 'Feedback', 'Teenagers', 'School' and'Open Domain' were identified for creation of text corpus. The participants were explained about each category with the help of an example [1].

The domain 'School' was meant for the age group between 15-21 years , Domain , 'Business' was meant for persons in the age group of 22years and above. The other domains were open for persons of all age groups. The participants were briefed about the contents of the messages by giving examples.

The participants selected for this work belong to different age group, different professions and academic qualifications. The age criterion was categorized into 3 groups and based on percentage the participants were registered into the system. Table 3 shows the grouping of age and percentage of participants.

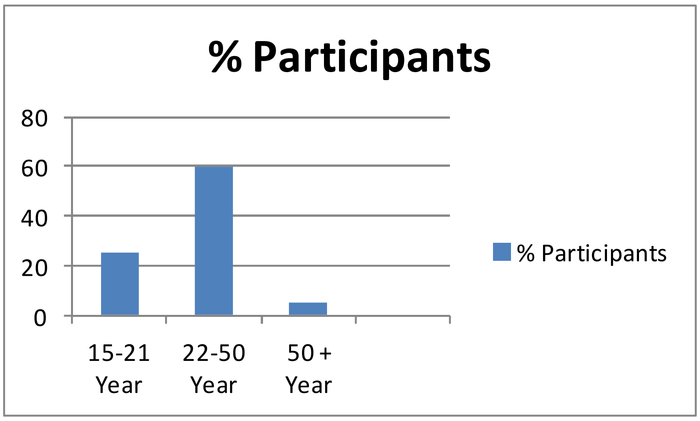


Figure 1: Percentage of participants based on age

All the participants registered in the system were necessarily having Hindi as their first language. As Hindi itself has many dialectal variations the Hindi speaking region of India was grouped into 5 groups [1]. This grouping helped us to capture the dialectal variation of different regions. By capturing these variations we look forward to develop more robust speech recognition system. For Indian English, participants all across India were selected for text and speech collection. These participants did not have English as their mother tongue. During selection process for speech collection efforts have been made so that dialectal variability of different parts of India could be captured.

The ratio of number of participants from each group represented the contribution of each group to the overall population of people with first language as Hindi. All the participants selected for Hindi and Indian English had to provide 10sms for each of the 11 domain based on their eligibility. In total, every participant had to give at least 110 messages.

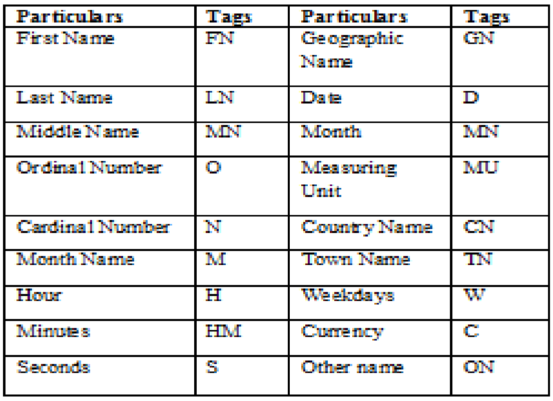
2.2 Message collection, expansion and tagging

The short messages were collected from the mobile sets of the participants with their permission, some messages were received through email and some other messages were received through an online system developed and implemented in a local area network. Since the QWERTY keyboard available with the computer does not provide friendly interface for Hindi language, a transliteration tool KTRANS based on the specifications of Itrans-3 was developed [3]. This tool receives input from the QWERTY keyboard in roman and converts the text into Hindi using Devnagri script. For Indian English text collection the conventional keyboard available with the computer was used for text collection. All these short messages were encoded in UTF-8 form. Initially these messages were in their raw form, using conventional grammar and slangs used in messaging through mobile.

There are a large number of words which are used as slang words in the actual messages. Some Examples of these words in Indian English are you = u, your = ur, please = plz, have = hv, and = n, are = r etc. In Hindi, however there are very few words which can be used as slang words. Moreover the Hindi messages are sent using roman script.

Once these raw messages were received they were next cleaned to check their correctness for grammar with reference to corresponding language grammar. Sentences which were meaningless/not clear were discarded during the process of cleaning. Sentences were manually corrected based on specifications/guidelines given by Nokia Research Center [4]. For example , it was instructed that a series of messages consisting of same or similar phrases while making the text message unique have to be altered a bit to create a more diverse database. Also it was instructed that if a message as a whole is good but it also contains a profanity, the message should be appropriately cleaned up rather than discarded as a whole. In case of Indian English initial word of sentences were capitalized and proper nouns used in the message text were also capitalized.

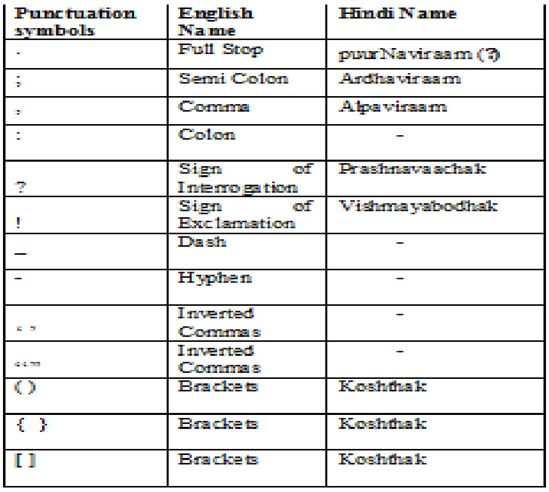
Once the messages had been checked for their correctness the next step was the expansion and tagging of messages. These expansions and tagging were done based on the specifications given by Nokia Research Center (NRC). The tagging is context specific and intent to simplify the meaning of the text. Table 4 shows the list of tags used to specify their meaning which varies depending upon context.



Particulars	Tags	Particulars	Tags
First Name	FN	Geographic Name	GN
Last Name	LN	Date	D
Middle Name	MN	Month	MN
Ordinal Number	O	Measuring Unit	MU
Cardinal Number	N	Country Name	CN
Month Name	M	Town Name	TN
Hour	H	Weekdays	W
Minutes	HM	Currency	C
Seconds	S	Other name	ON

Table 3: Tags used for specifying the context specific meaning

The punctuation marks and special symbols were expanded. Expansion of days, months, numbers and punctuation marks were done as their name in each language. For eg. In Indian English the weekdays Sun, Mon are expanded as Sunday, Monday etc whereas in Hindi सोम, मंगल are expanded as सोमवार, मंगलवार correspondingly. Similarly numerals zero to nine in Indian English is expanded as एक for १(one) and दो for २(two) and so on. Month name in Hindi and Indian English also differs. In Hindi name of month are as चैत्र, वैशाख etc. which corresponds to April and May months of English. During expansion all these issues were taken care of. Similar was the case with punctuation marks. Their expansion in both the databases has been done accordingly. Table 5 represents the expansion of punctuation symbols in both the languages.



Punctuation symbols	English Name	Hindi Name
.	Full Stop	puurNaviraam (?)
;	Semi Colon	Ardhaviraam
,	Comma	Alpaviraam
:	Colon	-
?	Sign of Interrogation	Prashnavaachak
!	Sign of Exclamation	Vishmayabodhak
—	Dash	-
-	Hyphen	-
‘ ’	Inverted Commas	-
“ ”	Inverted Commas	-
( )	Brackets	Koshthak
{ }	Brackets	Koshthak
[ ]	Brackets	Koshthak

Table 4: Expansions of Punctuations in English and Hindi

Since Indian mobile users widely uses words from other languages, during communication with others, special symbol was used for marking the foreign word i.e. word borrowed from other languages such as Urdu, Punjabi, English (in Hindi communication) and Hindi (in English communication).Hindi words like ‘thik hai’, ‘achha’, ‘namaste’ are widely used in communication in English.

Words such as ‘Hi’, ‘Great’, ‘good’, ‘train’ etc are the most commonly used words of English during communication in Hindi. The analysis of collected data shows that out of all collected unique words 15% are foreign words belonging to above mentioned languages.

Hindi	
Mode	**Message**
Raw	जयकुमार 1000 रु लेकर 10मार्च12 को दूसरी बस से लगभग 3:30 PM में दिल्ली के Hospital पहुँचा
Clean	जय कुमार 1000 रुपया लेकर 10 मार्च 2012 को दूसरी बस से लगभग 3:30 PM पर दिल्ली के Hospital पहुँचा।
Expend	FN/जय/FN LN/कुमार/LN C/1000/C रुपये लेकर D/10/D MN/मार्च/MN Y/2012/Y को O/दूसरी/O बस से लगभग <H/3/H HM/30/HM P_M> पर TN/दिल्ली/TN के &Hospital पहुँचा <पूर्ण_विराम>
English	
Mode	**Message**
Raw	hi D have a trip with 5000 discount from 3rd may12 to goa in summer ram’s family plz buy ticket 12:30 pm.
Clean	Hi Dear, Have a trip with discount ` 5000 from 3rd may 2012 to Goa in summer with Ram’s family, Please buy ticket by 12:30 PM.
Expend	&Hi Dear <COMMA> Have a trip with discount C/ `/C CN/5000/CN from D/3rd/D MN/ may/MN to TN/Goa/TN in summer with FN/Ram’s/FN family<COMMA> Please buy ticket by <H/12/H HM/30/HM P_M> <PERIOD>

Table 5: Modes of Hindi & Indian English Sentences

Table 5 shows an example of the manner in which the raw message of Hindi and Indian English have been cleaned, tagged and Expanded using the procedures explain above.

3. Speech data recording and annotation

The speech corpus for Hindi and Indian English was designed from the message information (text collected) in the following manner. Out of all the SMS texts collected, 13 prompt sheets of 630 phonetically rich sentences were created. These phonetically rich sentences were created to capture the versatility of mobile messages. To create phonetically rich sentences words were selected from the text corpus based on their frequency of occurrence in the whole text corpus [6] [7]. Meaningful sentences were framed using these words which were further manually corrected based on language grammar rules. Figure 1 shows the flow graph for the creation of phonetically rich sentences from the text corpus. Apart from these they were individually checked to ensure that there is nothing potentially offensive. Each prompt sheet was recorded as collection of continuous sentences, sentences with small gaps and in form of spelling sentences. It was ensured that every phoneme of the language is used in any of the 630 sentences of every prompt sheet. At least 7 speakers were recorded for each prompt sheet. The prompt sheets were recorded simultaneously through 3 channels.

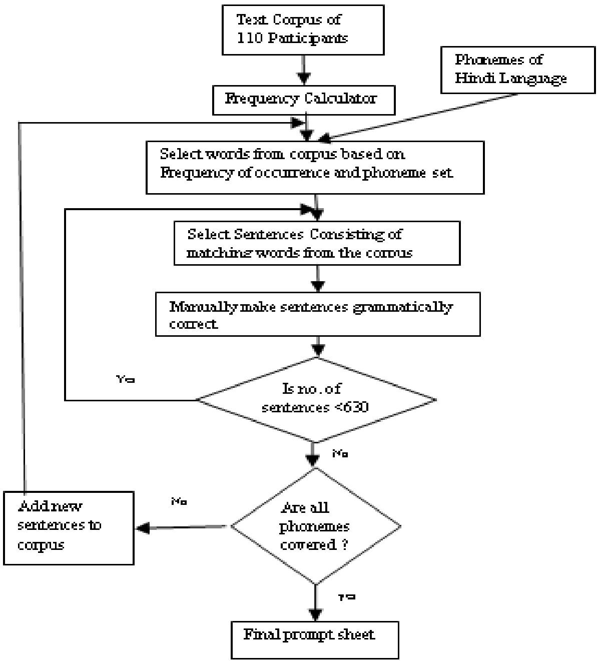

Text Corpus of 110 Participants
Phonemes of Hindi Language
Frequency Calculator
Select words from corpus based on Frequency of occurrence and phoneme set
Select Sentences Consisting of matching words from the corpus
Manually make sentences grammatically correct
Yes
Is no. of sentences <630
No
Add new sentences to corpus
No
Are all phonemes covered ?
Yes
Final prompt sheet

Figure 2: Flow graph for creation of phonetically rich sentence

The database obtained from speech recording consists of:

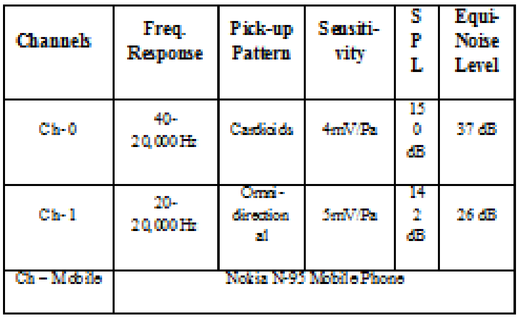
- Coverage of various dialectal variations.
- Coverage of phonetically rich words and sentences in communication domain
- Coverage of speaking styles(carefully pronounced , spontaneous speech)

- Coverage of environmental influence on recording through 3 channels.

The recorded speech was annotated based on the specifications of Nokia Research Center and unwanted sounds in the recording were marked. It was taken care that whenever the noise level or the quantity of unwanted sound in the recording exceeds the permissible percentage then that sentence has to be re-recorded.

3.1 Recording Environment

The speech signals were recorded through 3 channels simultaneously. The specification of each channel is given in table 6. The third channel used was the Nokia N-95 mobile phone.



Channels	Freq. Response	Pick-up Pattern	Sensiti-vity	SPL	Equi-Noise Level
Ch-0	40-20,000Hz	Cardioids	4mV/Pa	150 dB	37 dB
Ch-1	20-20,000Hz	Omni-directional	5mV/Pa	142 dB	26 dB
Ch – Mobile	Nokia N-95 Mobile Phone				

Table 6: Specifications of recording channels used for speech recording

The signals were recorded directly in the digital format. They were recorded with a sampling rate of 16 KHz, 16 bit quantization. In case of mistake or error in recording the same text was re-recorded. All the recordings were done in office environment at an SNR 40 db. The mobile phone was kept at a distance of approximately 15 inches apart from the speaker's mouth (see figure 2). Recording through mobile was done in speaker on mode. Auditech Audio recorder software was used for recording purpose. After the session was complete the data was transferred to the system through a Bluetooth data connection.

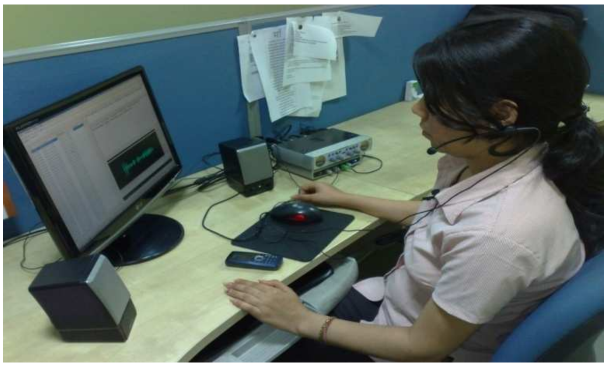


Figure 2: Recording through 3 channels simultaneously

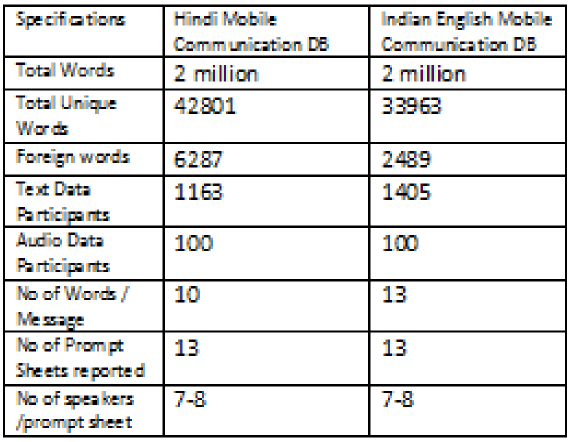
3.2 Speaker Specifications

For the purpose of Hindi database speakers having their first language as Hindi were only registered for recording.100 speakers from Hindi speaking regions representing different dialects were divided into different demographic criterion (age, gender and dialect) as set out in specifications. Also a minimum was imposed on different age group given in table 3. For Indian English 100 speakers from different part of India were selected for recording. During selection of speakers, issues like their acquaintance with English language and representation of different dialects and pronunciation variations were taken into consideration. The database comprised of 40% male data and 60% female recordings. After the recording of one speaker was over the recorded speech was annotated. This annotation was done to mark the unwanted sound, some particular noise and unvoiced portion of the speech.

4. Statistical analysis and observation

The database of Hindi and English included all the commonly used vowels and consonants, special letters, punctuation marks and symbols, numerals, name of months and currency etc. In Hindi Data Base foreign language words of English, Punjabi and Urdu were obtained. In Indian English very few foreign words were used but their occurrence was very frequent in the collected messages. The analysis shows that 15% of the unique words of Hindi and 8% of the unique words in Indian English are taken from foreign language during mobile communication in Hindi.

The comparative analysis has been outlined in table 7



Specifications	Hindi Mobile Communication DB	Indian English Mobile Communication DB
Total Words	2 million	2 million
Total Unique Words	42801	33963
Foreign words	6287	2489
Text Data Participants	1163	1405
Audio Data Participants	100	100
No of Words / Message	10	13
No of Prompt Sheets reported	13	13
No of speakers /prompt sheet	7-8	7-8

Table 7 : Comparative analysis of Hindi and Indian English Database in mobile communication.

From the recording of 100 speakers of different age group gender and dialect based upon the specifications in table 3 and table 4 it has been observed that the impact of dialectal variations were very much prominent also the foreign word pronunciation varied from one dialect to another.

During the analysis of speech data of English few observations were obtained which are outlined as:

- / θ / and / ð / are replaced by dental stops /t^h / and /d/ Substitution of unaspirated [p],[t],[k] for aspirated [p^h],[t^h],[k^h] at the beginning of accented syllables.
- Instead of /v/ and /w/ there is only one phoneme /ʋ/
- /□ / and /z/ are confused,/z/ or sometimes /□ / or /d_z/ are used for both.
- /e/ is replaced by either /æ/ or /ei/,/ei/ is replaced by /e/, /əu/ is replaced by / □ /, /□ / is replaced by /iə/
- English long vowels are made too short in final position by Hindi, Gujarati and Marathi speakers, especially from mofussil background.
- Final consonants are often followed by /ə / when they should not be, causing confusion between e.g bit and bitter.
- In Indian English rhythm is more like that of French than English. There is much less variation of length and stress and no grouping of syllables in to rhythm units as in English.

While observing the characteristics of the recording channel it was observed that the lapple microphone was the most sensitive, due to which ambience noise was also being captured. The S/N ratios of the headset microphones were good as they excluded the ambient noise more efficiently.

5. Validation

The specifications which the databases should meet are evaluated once the databases have been completed. The validation proceeds in four steps:

- Validation of text data to meet the specifications
- Validation of prompt sheets in order to check the corpus before recording begins and to make sure that it corresponds to the specifications and covers the versatility of text data.
- Pre validation of a small database of 10 speakers. The objective of this stage is to detect serious design error in the early stage.
- Validation of completed database. The database is checked against the specification.

After all the validation was over the speech and text database were finally stored on the permanent disk.

6. Experiences and current status

Different types of experiences done in the whole process are listed here. At the initial stage motivating around 1000 of participants for providing text data was the most difficult task. For Open Domain data some messages from Facebook and Orkut were collected. Their participation in research work was the sole motivation factor. Also convincing the participants to provide their personal mobile messages was very difficult. Most of the participants who provided messages from their mobile phones were faculty member, staff or students of KIIT college of Engineering belonging to different age and gender.

The text collection process for Hindi was tiresome and more complex and the sole reason for that was the lack of user friendly input unit for Hindi. The software KTRANS was used to fill the gap but then also manual task of checking was required. For English the data collection part was simple but expansion of Indian English messages was tiresome as they were mainly based on slangs and abbreviations. The speech collection process for Hindi was easier than Indian English even though the participants were required to be selected based on dialects of different regions. Selection of participants for recording of Indian English was more difficult as in some cases the regional dialectal influence was so strong that it was adversely affecting the clarity of pronunciation.

Since the number of sentences in the prompt sheets was large so recording was carried out in multiple sessions and hence recording parameters for speaker was required to be set in every session. Recording of one prompt sheet by single user required approximately 7 hours so 4 to 5 sitting were required to complete recording of one participant. The recording supervisors have to remain attentive during the whole process of recording to ensure that the speaker do not take a very casual approach and do the recording completely and in desired manner. There were problem of ambience noise getting recorded along with the sentences. Many times sentences were required to be discarded. The position of mobile phone from the speaker's mouth required extra attention during recording. The database of 2 million words and 100 speaker's sound recording for Hindi language and Indian English was created. This complete database for mobile communication is totally based upon the specifications provided by Nokia Research Centre.

7. Conclusion

The final database consists of a text corpus and a speech corpus for Hindi and Indian English recorded through three simultaneous channels of 100 speakers in mobile communication environment. The speakers male and female are of three different age groups and five dialectal regions for Hindi and from whole country for English. This database will be used for mobile based speech recognition services which would even help illiterate people of different age groups to use mobile channels for communication, making them easier to remain connected with their loved ones. The text database of 42801 unique words of Hindi and 33963 words of English would be used for developing language models, language features, language translation etc. This effort on message collection and analysis in mobile communication has been done for the first time. Thereby a new model for language and speech is being developed for a real life practical application.

8. Acknowledgement

We would like to thank Mr. A.K. Dhamija, for help in Recording PCOM Text and Speech Data and regularly maintaining it. Thanks are extended to our supporting staff Mr. Gautam Kumar, Mr. Yogesh chand, Mr. Heera Prasad and Mr. Jagmahinder for their continuous assistance in development of text and speech corpora from various sources. Our special thanks are due to KIIT management particulars Dr. Harsh Vardhan & Ms Neelima Kamrah for their constant support and providing conducive environment for this work.

9. References

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

[2] Gibbon, D., Moore, R. & Winski, R. (1997), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter. Berlin, New York. 1997

[3] ITRANS- Technology Development for printing in Indian Languages using English encoded input http://www.aczoom.com/itrans

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

[5] Karunesh Arora, Sunita Arora, S S Agrawal, Niklas Paulsson, Khalid Choukri, "Experiences in Development of Hindi Speech Corpora based on ELDA standards" Proceedings of O-COCOSDA 2006

[6] Sunita Arora, Karunesh Kr Arora, S.S.Agrawal "Vishleshika: Statistical Text Analyzer for Hindi and other Indian Languages" Proceedings of International Workshop on Spoken Language Processing, TIFR, Jan 9-11, 2003 pp 191-198

[7] Karunesh Arora, Sunita Arora, Kapil Verma, S S Agrawal, " Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages" INTERSPEECH2004 -ICSLP Jeju,Korea

[8] Jesper Olsen, Yang Cao, Guohong Ding, Xinxing Yang. " A decoder for large vocabulary continuous short message dictation on embedded devices". In Proceedings of ICASSP'2008. pp.4337-4340

[9] Zheng-Hua Tan (etac) "Speech Recognition on Mobile Devices: Distributed and Embedded Solutions" Inter Speech 2008, Brisbane, Australia -22-26 Sept.2008

[10] Z.H.Tan and B. Lindberg (Eds.) "Automatic Speech Recognition on Mobile Devices and over Communication Networks. Springer-Verlag, London-2008.

[11] Qiru Zhov and Imed Zitouni, in speech Recognition, Technology and applications, "Arabic Dialectical Speech Recognition in mobile communication". (ed.- Fance mihelic and Janez Zibert)2008, I-Tech, Vienna, Austria.