Grenze International Journal of Computer Theory and Engineering , Special issue

Speech Corpus Development for Speaker Independent Speech Recognition for Indian Languages

Amaresh P Kandagal[1]* and V Udayashankara[2]
[1]Research Scholar - Sri Siddhartha Academy of Higher Education – Tumkur – India
Email - amareshpk@gmail.com
[2]Professor - Sri Jayachamarajendra College of Engineering – Mysore – India
Email - v_udayashankara@sjce.ac.in

***Abstract*— In this paper, we discuss development of speech corpus for speaker independent speech recognition for Indian airports and it is extended for continuous speech recognition Indian languages. We have collected the speech corpus from 801 speakers to build large vocabulary ASR engine. Speech corpus recorded over telephone line and microphone. It is recorded from speakers ranging from age group between 20 to 60 years. 4.5 hours of microphone data recorded from 375 male and 244 female voices. The telephonic data is of 1.3 hours which includes male and female voices. Total 6.2 hours of speech corpus is collected. The recording was conducted at office, college and home environments. We also discuss preliminary isolated speech recognition results using the acoustic models created on these corpus using Hidden Markov Model toolkit (HTK).**

***Index Terms*— Speech recognition; isolated word; data analysis and visualisation; speaker independent; accent variation.**

I. Introduction

Population of world is getting lot of benefit from smart devices by exploring and accessing online information in their day to day activity. For Human Computer Interaction (HCI) keyboard or keypad are the key interfaces to communicate with smart devices such as smart phones, computers, access controllers etc. But this interface is not user friendly and not flexible to use for illiterate and semi-literate population. And moreover to type or write data for interacting with smart devices is expensive (time consuming).Speech is the intuitive way to interact and is the common interface for human to human interaction. Hence robust Automatic Speech Recognition is crucial for HCI. Definition of ASR engine is a machine maps an acoustic speech signal to some form of abstract meaning of the spoken information, and translates it in the form of text or command with which machine can react .Though there are several ASR engines are available in market, the performance of is far from the perfect. However it is ease and more helpful for end users if there are speech interfaces to provide information in their local languages. ASR systems are very useful for visually challenged people.

Country like India is a multilingual society which has about 1652 dialects/native languages. There are 22 languages that are been recognized by the constitution of India. The languages recognized by Indian Constitution are: 1) Assamese 2) Bengali 3) Bodo 4) Dogri 5) Gujarati 6) Hindi 7) Kannada 8) Kashmiri 9) Konkani 10) Maithili 11) Malayalam 12) Meitei (Manipuri) 13) Marathi 1 4) Nepali 15) Odia 16) Punjabi 17)

Grenze ID: 01.GIJCTE.3.4.14

Sanskrit 18) Santhali 19) Sindhi 20) Tamil 21) Telugu and 22) Urdu [1], [2]. Hindi is the official language of India which written in Devanagari script. For Indian languages the amount of work has not yet reached to a critical level, speech domain to be used as real communication tool, as that in other languages of developed countries. HP Labs India and IBM research lab attempts to develop speech recognition system [1], [3]. Indian languages which are of different variations however, there is lot of scope to develop language technology. To achieve such ambitious goals, the collection of standard speech databases is prerequisite.

This paper describes the design and development of speech corpora/database for Indian airport places and Indian accent English words. Section 2 describes scripts and sounds of Indian languages. Section 3 describes literature survey on speech corpus for Indian languages by the Linguistic Data Consortium for Indian Languages (LDC-IL). Section 4 gives the transcription and labelling tools for editing acoustic signal Section 5 briefly explains about major components of an ASR system Section 6 describes performance evaluation method and the conclusion and discussion is in section 7.

II. Scripts and Sounds of Indian Languages

The Brahmi script is the earliest writing system developed in India after the Indus script. The scripts for Indian languages are derived from the ancient Brahmi script. Aksharas are the basic units of the writing system. The attributes of Aksharas are as follows: (1) An Akshara is an orthographic representation for a acoustic signal in an Indian language;(2) Aksharas are syllabic in nature; (3) The typical forms of an Akshara are CV, CVC, CCV,V, CCCV, VC, and thus have a generalized form of C*V. Here C denotes a consonant and V denotes a vowel sounds.

A. Convergence and Divergence

Except English and Urdu language most of the languages in India share a common phonetic base, i.e., they have a common set of speech sounds. A single unit of sound is known as phoneme. The phonemes are broadly classified as vowels and consonants. In most of the Indian languages there are closely fifty phonemes in which fifteen are vowels and remaining are consonants. Vowels are subdivided into semi-vowel and Diphthongs and consonants are classified as nasals, fricatives, stops, affricates and whisper. Devanagari is the common script for Hindi, Marathi and Nepali languages. South region of Indian languages such as Kannada, Telugu, Tamil and Malayalam have their own scripts.

Languages are differentiated by phonotactics. Phonotactics are acceptable combinations of phones that can coincide in a language. This predicates that the distribution of syllables encountered is different in each of these languages. The Indian languages are significantly distinguished by the property called Prosody (intonation, duration and prominence) associated with a syllable [4].

III. Literature Review

IIIT Hyderabad and HP Labs Bangalore is developed a large vocabulary speech database. The collected database is for Marathi, Tamil and Telugu languages. The recorded acoustic data is by using landline and Smartphone. Different age groups of 559 speakers acoustic data is recorded all three different languages. It was collected from the native speakers of the language. The speech corpus composed of background noise and disturbance caused due to use of phone line [1].

Rajarambapu Institute of Technology Sakharale, Islampur, Maharashtra developed a Text to Speech Synthesis for Konkani Language. A limited vocabulary speech data base is developed to build TTS system. The speech corpus is recorded for 1000(thousand) Konkani language regularly used words. The speech data is collected over a microphone on computer. Complete recording were conducted in computer lab. Corpus size is around three thousand wave files composed of Vowels, Characters, Barakhadi, and half characters [5].

A Kannada language Speech database has been developed to build TTS system at Mysore. Phonemes are the building blocks to develop TTS engine. The speech corpus consists of total 1605 phonemes. The phonemes were recorded by using PRAAT utility tool on Windows Operating System platform. Speech corpus sampling frequency is 16000 Hz. All the recording is conducted in indoor environment. Speech data is captured from microphone sensor. Speech corpus is consists of vowels, fricatives, stops, nasals, diphthongs, affricatives etc [6].

Hindi and Indian Spoken English TTS project is developed at KIIT, Bhubaneswar. It was sponsored by Nokia Research Centre, China. The speech corpus was performed using 13 prompt sheets containing 630 phonetically rich sentences in each language prepared after collecting text messages in Hindi and Indian

Spoken English. Hundred speaker’s speech data was collected. Unique words 42,801 for Hindi and 33,963 for Indian spoken English text corpus was used to collect the speech samples. Recordings were performed in three different channels (i.e. mobile phone, cardioid microphone and Omi directional microphone). The speech corpus is composed of 60% female voices and 40% male voices of 16 KHz sampling rate [7].

The Linguistic Data Consortium for Indian Languages (LDCIL) is the Consortium established after a long persuasion for developing a similar activity like Linguistic Data Consortium (LDC) at the University of Pennsylvania. LDC-IL is supported by the Central Government India. The LDC-IL is accountable to create the database and also will provide opportunity for the research and developers to build speech applications using the collected data in various domains. The LDC-IL has collected Speech databases in various Indian Languages. Table-1 describes the size of Speech Corpora by LDC-IL (as on July 2014) [8].

IV. Speech Corpus

The development of speech corpus is carried out either application specific or general purpose. This paper, describes development of speech corpus for speaker independent speech recognition for Indian airports and it is extended for continuous speech recognition Indian languages. We have collected the speech corpus from 801 speakers to build large vocabulary ASR engine. Speech corpus recorded over telephone line and microphone. It is recorded from speakers ranging from age group between 20 to 60 years. 4.5 hours of microphone data recorded from 375 male and 244 female voices. The telephonic data is of 1.3 hours which includes male and female voices. Total 6.2 hours of speech corpus is collected. The recording was conducted at office, college and home environments.

The scope of speech corpus is to develop immersive language learning tools and to build interactive voice response (IVRS) system applications such as flight and movie ticket booking system. The corpus is consists of Indian airport names, pronunciations of months, week days, am/pm and pronunciation of digits in Indian local languages. The text corpus is designed according to phonemes histogram. Text corpus is of isolated and connected words. The development of speech corpus is broadly described in 2 phases 1) Recording setup and 2) Speech data processing.

A. Recording Setup

While recording speakers were seated at a table microphone connected to a laptop. Speakers were advised that to speak in natural way, not to add any artificial expressions or emotions during voice sample collection. It’s been taken care that speaker should be comfortable zone to readout the text corpus. These were depicting the situation in which the ASR system being developed will eventually be used. Telephonic speech database is collected with a setup of CTI Dialogic analog card.

Recording Software

Gold wave version 5.4 and Wave surfer tools were used for voice recording and editing purpose. Microphone speech was recorded at16 kHz sampling rate and these were stored in .wav format. Telephone speech was recorded at 8 kHz sampling rate and managed through CTI dialogic card for an Asterisk-based PBX phone system.

Recording Locations

Recording are conducted at office rooms and a student lab, class rooms and home environments. External noise in the office, college, and home environment was contributed by the opening and shutting of doors and drawers, people talking, printers, telephones etc.

B. Speech Data Processing

This section describes speech data processing after it had been acquired through the recording session.

Speech cleaning and segmentation

Voice sample recordings was manually split into smaller chunks, about 10 to 30 seconds long, using wave surfer speech editing tool. Labelling or marking speech data as silence, noise and pronunciation are three main aspects considered in cleaning of the speech corpus. The fundamental rule applied during acoustic data cleaning and segmentation is to only mark a boundary during silence (though desired, it was not always aligned with a phrase or a sentence boundary). Thus smaller .wav files (not more than 30 seconds long) were produced.

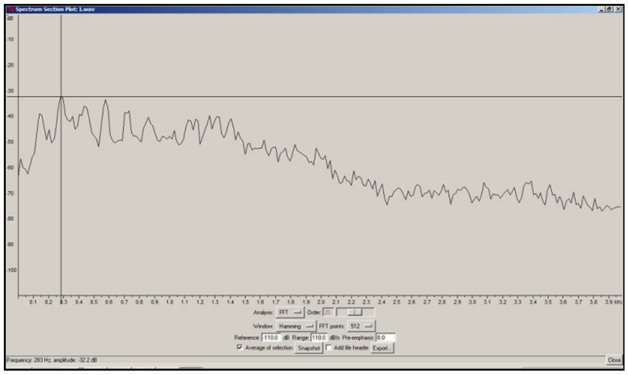
TABLE I.LDC-IL SPEECH CORPUS FOR INDIAN LANGUAGES

Sl No.	Languages	Raw Data	**Segmented Data**			**Annotated Data**	**Pronunciation Dictionaries (studio recording)**
		HH:MM:SS	Dialog	**HH:MM:SS**	Speakers	**HH:MM:SS**	**HH:MM:SS**
1	Assamese	105:51:38	Upper Assam, Lower Assam	80:8:4	306	28:18:56	36:31:28
2	Bengali	138:18:47	SCB (Kolkata) & Barendri (North Bengal)	125:19:53	697	39:12:31	21:55:46
3	Bodo	201:10:48	Standard and Non Standard	198:10:48	416	30:45:56	50:38:55
4	Dogri	111:32:11	Standard	17:10:58	61		
5	Gujarati	156:23:04	Avadhi, Bhojpuri, Magahi & Standard	20:4:18	52	2:39:39	49:00:00
6	Hindi	269:09:50	Standard And South Gujarati	34:12:57	53	80:01:48	45:51:32
7	Indian English Bengali	34:12:57	Indian	77:38:53	302		
8	Indian English Guajarati (MP3 Format)	21:40:00	Indian	16:52:24	62		
9	Indian English Kannada	37:01:33	Standard, Bhojpuri & Magahi	174:31:50	586		
10	Kannada	198:51:03	North-East(Hyderabad Karnataka), North-West(Mumbai Karnataka) and Canara	157:50:25	642	69:05:56	58:20:43
11	Kashmiri	44:59:07	Standard	29:26:13	149	6:28:25	
12	Konkani	195:14:47	Standard	193:14:47	454	37:00:00	32:29:53
13	Maithili	95:59:54	Standard	88:0:43	301	30:10:40	
14	Malayalam	265:24:18	Standard	103:16:12	307	92:40:43	33:03:05
15	Manipuri	187:35:13	Standard and Kakching	175:26:9	668	109:48:27	49:41:18
16	Marathi	168:13:50	Standard	89:18:43	306		
17	Nepali	145:04:46	Darjeeling and Assamese	114:39:29	351	39:33:34	23:23:35
18	Oriya	165:30:05	Standard	165:30:5	474	62:33:15	40:00:33
19	Punjabi	187:53:28	Standard	187:53:28	468	47:07:13	33:30:06
20	Tamil	213:37:27	Standard	211:37:27	446	72:09:09	48:12:10
21	Telugu	50:51:36	Standard	13:3:43	56		
22	Urdu	124:19:58	Standard	110:46:15	342	36:33:55	34:02:39

The speech data that included disruptive noises, such as street noise, babble and traffic noise. Noisiness can affect the performance of ASR system. 10dB threshold is set to check the noise level. The speech data SNR is less than 10 dB then the file is discarded from the training process. "Fig. 1," describes the spectrum section of acoustic segment for SNR analysis. A telephone ring, a drawer opening or closing, or someone else speaking, in close proximity to the speaker, were marked as unfeasible data for the training process.

Speech Transcription

A team of linguists, the segmented speech files were transcribed manually. Each speech segment file had a corresponding phoneme transcription. The orthographic transcription was converted into phonemic transcription using HMM force alignment by HTK tool kit. In addition to the transcription of speech in segments, the Silence, Vocalization and Breath tags were defined to represent non-speech areas in the segments. All silences or pauses during speech as audible or viewable in the waveform displayed on wave surfer tool were marked with a silence tag, in particular at the start and end of segmented portions. Sounds produced by the speaker that could not be classified as speech were marked by a vocalization tag within the orthographic transcription. Breath sounds identified within segments were marked with a breath tag. “Fig. 2,” describes the labelling of speech signal.


Spectrum Section Plot: 1.wav
dB
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 kHz
Snapshot
Export...
Close

Analysis: FFT Order: 20

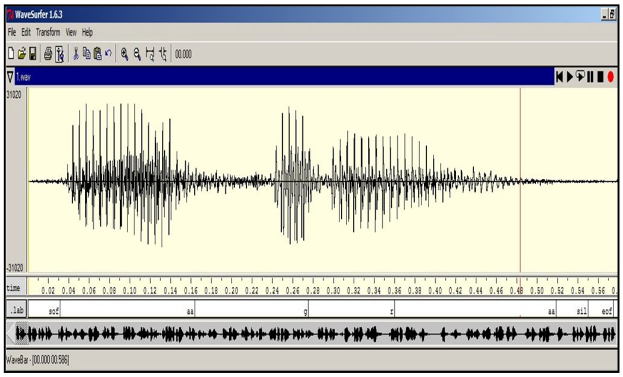
Window: Hamming FFT points: 512

Reference: -110.0 dB Range: 110.0 dBfs Pre-emphasis: 0.0

☑ Average of selection ☐ Add file header

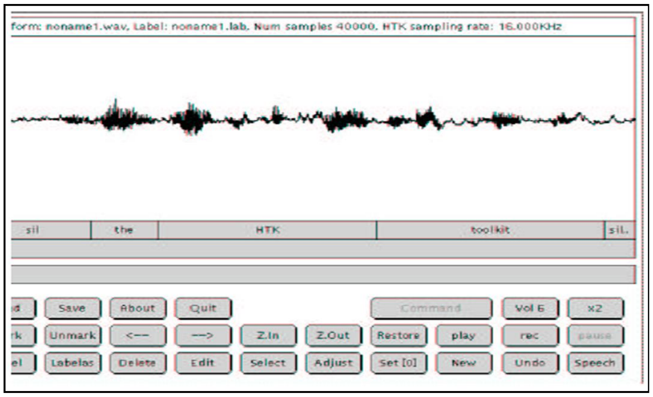
Frequency: 283 Hz, amplitude: -32.2 dB

Figure 1. Acoustic Signal Spectrum Section



WaveBar - [00.000 00.586]

Figure 2. Labeling of Speech Signal by Wave Surfer Tool



form: noname1.wav, Label: noname1.lab, Num samples 40000, HTK sampling rate: 16.000KHz

Figure 3. Hslab Display Window

Acoustic transcription and labelling is also carried out by using HTK tool Kit. HSLab is an HTK library function and it is an interactive label editor for manipulating speech label files. An example of using HSLab would be to determine the boundaries of the speech units of interest and assign labels to the desired acoustic signal. HSLab is the only tool in the HTK package which makes use of the graphics library H_{GRAF}. “Fig. 3,” illustrate HSLab window which is split into 2 parts: a display section and a control section. The display section contains the plotted speech waveform with the associated labels [9].
HSLab is invoked by typing the command line: *HSLab [options] dataFile*

V. BUILDING ASR SYSTEMS

In general, ASR systems consist of three major components- the acoustic models, the phonetic lexicon and the language models . In the subsections that follow, each of these components is briefly discussed.

A. Acoustic Model

An acoustic model is a file link to the observed features of acoustic signal represented in statistically for each of the distinct sounds that makes up a word. Acoustic model is the first key constituent of the ASR systems.

B. Pronunciation / Lexical Model

A lexicon or dictionary is referred to provide the mapping between words and phones or sub-word units. Which is developed to provide the pronunciation of each word for a specific language i,e it contains information about how words are pronounced[10].

C. Language Model

The language model tries to convey the behaviour of the language. By the means of probabilistic information, the language model is developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary [11]. It is the single largest component trained with several words, consisting of billions of parameters.The probability of occurrence of a word sequence W is calculated as

$$P\ (W) = P\ (w_1,\ w_2,\ ..,\ w_{n-1},\ w_n) = P\ (w_1).P\ (w_2|w_1).P\ (w_3|w_1w_2)\ ...\ P\ (w_n|w_1w_2\ ...w_{n-1})$$

Language models are cyclic and nondeterministic.

VI. BASELINE ASR SYSTEM PERFORMANCE

This section describes the performance analysis of ASR system for the acoustic models created on this corpus using Hidden Markov Model toolkit. The evaluation of the experiment was made according to the recognition accuracy and computed using the word error rate (WER) metric which aligns a recognized word string against the correct word string and computes the number of substitutions (S), deletions (D) and Insertions (I) and the number of words in the correct sentence (N) [12]. WER = 100 * (S+D+I) / N

The preliminary performance of an ASR system was computed for vocabulary size of sixty five words. Acoustic models were created by training collected speech corpus (microphone data). The system performance is computed for nine speakers (Untrained voice samples). The accuracy is of the system is achieved 93.46% for isolated speaker independent ASR system. The whole experiment is performed on microphonic speech corpus by using HTK tool kit.

VII. CONCLUSION AND FUTURE WORK

The current work describes the optimal development and the use of speech corpus for Indian languages. In this paper we have discussed some of the speech corpus developed in different Indian languages for various applications. We presented the simple methodology of database creation, acoustic data cleaning, and labelling processes. We conducted a preliminary experiment on collected data to build isolated speaker independent ASR engine with HTK tool kit. We have achieved 93.46% accuracy. The WER can be minimised by pre-processing and tuning of acoustic models. We expect the ASRs created will serve as baseline systems for further research on improving the accuracies. Our future work is focused in tuning these models and test them using language models built using a larger corpus.

ACKNOWLEDGEMENTS

This work has been conducted through the projects, automatic airline ticket booking system and immersive languages learning tools for Indian languages at Sprec India Pvt Ltd and Focus Softek Pvt Ltd respectively. Would like to acknowledge both the organisations for providing support and help to carry this research work.

REFERENCES

[1] Gopalakrishna Anumanchipalli, 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”. In Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece. 2005.

[2] “ https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India”

[3] Chalapathy Neti, Nitendra Rajput, Ashish Verma. “A Large Vocabulary Continuous Speech Recognition system for Hindi” In Proceedings of the National conference on Communications, Mumbai, 2002, pp. 366-370.

[4] Kishore Prahallad , E.Naresh Kumar , Venkatesh Keri , S.Rajendran , Alan W Black The IIIT-H Indic Speech Databases

[5] Sangam P. Borkar and Prof. S. P. Patil. “Text To Speech System For Konkani (Goan) Language” , In Proceedings of W3C Workshop on Internationalizing the Speech Synthesis Markup Language III Agenda. 2007.

[6] D. J. Ravi and Sudarshan Patilkulkarni, “A Novel Approach to Develop Speech Database for Kannada Text-to-Speech System”, International Journal on Recent Trends in Engineering & Technology, in ACEEE Vol. 05, No. 01, 2011.

[7] Shyam Agrawal, Shweta Sinha, Pooja Singh, Jesper Olsen. “Development of Text and Speech Database for Hindi and Indian Enlish specific to Mobile Communication Environment” In Proceeding of International Conference on The Language Resources and Evaluation Conference, LREC, Istanbul, Turkey. 2012.

[8] “ http://www.ldcil.org/resourcesSpeechCorp.aspx ”

[9] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., And Woodland, P. The HTK Book (for HTK Version 3.3). Cambridge University Engineering Department, April 2005. http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf.

[10] Prashanth Kannadaguli,Ananthakrishna Thalengala,"Phoneme modeling for speech recognition in Kannada using Hidden Markov Model",International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES),IEEE 2015.

[11] X. Huang, A. Acero, H. Hon, Spoken Language Processing: A Guide to Theory, System and Algorithm Development , New Jersey, Prentice Hall, 2001

[12] Geetha.K,Chandra.E,"Automatic Speech Recognition - An Overview ",International Journal Of Engineering And Computer Science,Page No. 633-639,2013