P Jeethendra et al, International Journal of Computer Technology & Applications,Vol 8(3),349-360

Accuracy of Telugu Language Speech Recognition corresponding to Size of the Speech Corpus Database

P. Jeethendra
Research Scholar (PP ECE 0125),
Rayalaseema University, Kurnool (AP) India, email: jeethendrapathak@gmail.com

M. Chandrashekar
Centre for Advanced Systems, SFD-BDL,
Kanchanbagh, Hyderabad, Talangana, India, email: cmatham@gmail.com

ABSTRACT

In this paper, efforts were put in finding development of Indian language speech databases in general and Telugu in particular for building large vocabulary speech recognition systems. The speech to text database for the Speech and Hearing Disabled needs a higher order of accuracy and wide database and a high word recognition rate with lowest error. The design and methodology of collection of various speech databases is discussed.

One of the major challenges in the field of Automatic Speech Recognition (ASR) is Modeling of rich transcription, modeling. The speech database should be rich in many dimensions. such as in text. environments. transducer type, number of recording sessions, recording system, the transmission channel, the country of origin, and the mother language. Research in any Language processing and very useful in many speech processing tasks. such as speaker recognition, speech recognition, and accent identification.

It was found that databases for European Language are widely and abundantly available (2) but a little database is available for Indian Languages. The accuracy of the speech recognition depends upon the size and amount of database in that language.

Keywords : Database, Large Vocabulary, Automatic Speech recognition, Accuracy, Indian Languages.

I. INTRODUCTION

The development of the database requires a set of phonetically rich sentences (1). To select a phonetically rich sentence, one of the important decisions to be made is the choice of a huge text corpus source from which the optimal sub-set has to be extracted. The reliability and coverage of the optimal text and of the language model largely depends on the quality of the text corpus chosen. In this paper speech corpora and some methodological concerns about widely used database, design techniques, experiments on Word error rates in speech recognition system for the English and Indian Languages are surveyed and discussed. It was observed that the corpus should be unbiased and large enough to convey the entire syntactic behavior of the language.

Speech database is a core part to evaluate the performance of the system in speech processing field. The developed system can be deployed successfully in real life only if it is evaluated by a versatile and relevant database.(7) There are many databases in major languages like English, Spanish, German, Japanese, Chinese, etc. These databases are rich in number of speakers, amount of speech, variability of speakers and texts, environments, and transmission channels. However, Telugu databases are few in numbers and most of them are conservative. Therefore there is a need for publicly available comprehensive Telugu speech database which shall cater the need of even the Speech and Hearing Disabled. The computers System which can understand the spoken language can be very useful in domains like health care, Improvement of life style and Education of Speech & Hearing Disabled, Visually Disabled and Government services. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. Speech recognition needs developing system for analysis and classification of the speech signals. Since, 1960s computer scientists have been researching ways and means to make computer record, interpret and understand human speech by taking the various aspects of the Speech such as its Formants, Phones, Pitch etc.

In order to get adequate amounts of speech to train and test the speaker recognition system, speech databases are needed. There are several applications of speaker recognition, leading to a diversity of the structure and content of speaker recognition databases. The most important benefit of using standard and readily available databases is that system performances using different techniques on the same database become comparable and standardized, hence, enabling quantitative evaluation of methods and speaker recognition protocols. In a search we have found 36 existing databases including both public and proprietary bases that have been used in speaker recognition studies, a comprehensive review of databases has earlier been given in the project report (23) We here provide an overview and aspects of a taxonomy of speech databases, in order to facilitate future case studies and new database design. We also describe a new database ELSDSR which was created to meet the needs of our own recent effort in speaker recognition, and which will be freely available for research

II. AN OVERVIEW OF THE DATABASES OF THE INDIAN LANGUAGES

A Speech Database of Hindi language for Automatic Speech Recognition system for Travel domain has been developed at C-DAC Noida. The database consists of training data collected from 30 female speakers in a noise free environment consisting of approximately 26 hours of speech recordings. Total 8,567 sentences consisting 74,807 words were recorded by the speakers uniformly distributed over all age group from 17 to 60 years.(2)

A Punjabi language Speech Database has been developed for Text to Speech synthesis system at Department of Computer Science, Punjabi University, Patiala. The syllables were considered for developing said speech database for Text to Speech Synthesis system because the researchers have selected syllables as the basic unit of concatenation. This Punjabi language speech database consists of 3,312 syllables which account for more than 99% of commutative percentage frequency in the selected corpus.(5).

A Garhwali speech database is being developed for development of Automatic Speech Recognition system for Garhwali language at Government P.G. College, Rishikesh. A total number of 100 speakers consisting of 50 male and 50 female would be selected to speak the selected words or sentences. All speakers are from different district of Uttarakhand i.e. out of 13districts of Uttarakhand. They have considered Tehri Garhwal, Pauri Garhwal, Chamoli, Rudraprayag and Uttarakashi districts of Uttarakhand for recording the speech. In these districts of Uttarakhand Garhwali is spoken quite frequently. For developing the speech database a text corpus consisting 11,188 isolated Garhwali tokens/words has been prepared. For recoding the speech data PRAAT would be used. The speech recording would be done in the lab in noisy environment which would be helpful for the development of the robust speech recognition system (17)

A General purpose, multi speaker, Continuous Speech Database has been developed for Hindi Language by the researchers of TIFR Mumbai and CDAC Noida. The Hindi Speech database is comprehensive enough to capture phonetic, acoustic, intra-speaker and inter speaker variabilitys in Hindi Speech. This database consists of sets of 10 phonetically rich Hindi sentences spoken by 100 Native speakers of Hindi language. The speech data was digitally recorded using two microphones in a Noise free environment. Each speaker was asked to read the 10 sentences consisting 2 parts. The first part consists of two „Dialecf sentences which preferably covers the maximum phonemes of Hindi language. Every speaker was asked to speak these two sentences. The second part consisted of 8 sentences which covered maximum possible phonetic context. Though this continuous speech database was developed for training speech recognition system for Hindi language, it has been designed and developed in such a manner that is can also be used in tasks such as speaker recognition, study of acoustic-phonetic correlation of the language (3).

A Speech database has been developed for developing a Text to Speech Synthesis system in Kannada Language at Mysore. The basic entity selected for the speech synthesis in this project was phonemes. This speech database consists of total 1,605 phonemes. The phonemes were recorded using the utility tool PRAAT on Windows Operating System platform. The sampling frequency used for recording the speech was 16,000 Hz. The recording was done using the standard microphone in lab. The recorded phonemes include vowels, semi vowels, stops, fricatives, nasals etc. (18).

A MIS (i.e. Mandi Information System) for retrieval of commodity price of market using mobile/telephone system has been developed at IIIT Hyderabad. The proposed MIS was in Telugu

language. Speech corpus consisting of 17 hours of speech data recorded from 96 speakers in noisy environment using mobile phones. A total of 500 words were recorded from each speaker. Approximately 15 hours of recorded speech data has collected to build the acoustic model of ASR (21)

A speech to speech synthesis system for travel and emergency services in Indian languages is developed at IIIT Hyderabad. The speech databases developed include English, Telugu and Hindi speech corpus from 15 different speakers. All the recordings were done using a laptop and a standard microphone in a room in noise free environment (22).

III. CORPORA OF SPEECH DATABASE

Current ASR systems are usually based on Hidden Markov Models (HMM). HMM based recognizer consists typically of 3 principal function modules: feature extraction (parameterization), acoustic modeling, and decoding.
The performance of the speech recognition systems is given in terms of a word error rate (%) as measured for a specified technology, for a given task, with specified task syntax, in a specified mode, and for a specified word vocabulary. The first step followed in creating a speech database for building an Automatic Speech Recognizer (ASR) is the generation of an optimal set of textual sentences to be recorded from the native speakers of the language (6). The selected sentences should be minimal in number to save on manual recording effort and at the same time have enough occurrences of each type of sounds to capture all types of co- articulation effects in the chosen language. In this section, the various stages involved in the generation of the optimal text are described.

A principles of speaker recognition databases may be based on features such as the recording protocol, the population of participating subjects, the recording device, language(s), type of verbal statement, and the intended use, etc. The intra-speaker and inter-speaker variability are important parameters for a speech database. Intra-speaker variability can be very important for speaker recognition performance and can be estimated if the same sentence is read several times by the same subjects. The intra-speaker variation can originate from a variable speaking rate, changing emotions or other mental variables, and in environment noise. The variance brought by different speakers is denoted inter-speaker variance and is caused by the individual variability in vocal systems involving source excitation, vocal tract articulation, lips and/or nostril radiation (9). If the inter-speaker variability dominates the intra-speaker variability speaker recognition is feasible. Speech databases are most commonly classified into single-session and multi-session. Multi-session databases allow estimation of temporal intra-speaker variability. Combination sets are also possible including single-session recording with a larger set of speakers and multi-session recordings with a smaller set of speakers, for instance, SpeechDat, Switchboard-1, SIVA and Gandalf,(23). For sampling of low interspeaker variability subjects, which is relevant, e.g., for admission control systems, some databases even include close relatives among speakers , or human mimicry and technical mimicry (24). With respect to input devices the most common means of recording are microphones or telephone handsets, the latter can be modified by being over local or long distance telephone lines. According to the acoustic environment, databases are recorded either in noise free environment, such as in the sound booth, or with office/home noise. Moreover, according to the purpose of the databases, some corpora are designed for developing and evaluating speech recognition, for instance TIMIT (25), and some are specially designed for speaker recognition, such as SIVA, Polycost and YOHO . Many databases were recorded in one native language of recording subjects; however there are also multi-language databases with non-native language of speakers, in which case the language and speech recognition become the additional use of those databases.

IV. SPEECH CORPORA FOR ENGLISH LANGUAGE

There are many Speech structure development tools referred as Corpora, for speech recognition. The most popular bases are TIMIT and its derivatives, Polycost, and YOHO.

IV.a. TIMIT Corpus

The TIMIT corpus (TIMIT - Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)),of read speech has been designed to provide speech data for the acquisition of acoustic phonetic knowledge and for the development and evaluation of automatic speech recognition systems (25). Although it was primarily designed for speech recognition, it is also widely used in speaker recognition studies, since it is one of the few databases with a relatively large number of speakers. It contains 630 speakers' voice

messages (438 M/192 F), and each speaker reads 10 different sentences. It is a single-session database recorded in a sound booth with fixed wideband headset. The derivatives of TIMIT are: CTIMIT, FFMTIMIT, HTIMIT, NTIMIT, VidTIMIT.

IV.b.. Recording of the TIMIT

They were recorded by playing different recording input devices, such as telephone handset lines and cellular telephone handset, etc. TIMIT and most of the derivatives are single-session, and are thus not optimal for evaluating speaker recognition systems because of lack of intra-speaker variability. VidTIMIT is an exception, being comprised of video and corresponding audio recordings of 43 subjects. It was recorded into 3 sessions with around one week delay between each session. It can be useful for research involving automatic visual or audio-visual speech recognition or speaker verification (10).

IV.c. Polycost

The Polycost corpus was an activity of the so called COST 250 European project. It includes both native and non-native English from 134 speakers (74 M/60 F) from 13 European countries. Therefore it can not only be used in speaker recognition, but language and accent recognition as well. It has more than 5 sessions recorded over weeks in home/office environment by variable telephone handsets through digital ISDN.

IV.d. YOHO

The YOHO corpus was designed for evaluating speaker verification in text-dependent situation for secure access applications. It consists 138 speakers' speech messages (106 M/32 F). It was recorded in multi sessions over a three months period by fixed high-quality handset in the office environment. The text read was prompted digit phrases.

V. WORD ERROR RATE (WER)

The accuracy of ASR system is generally evaluated using corpora of grammatically sound, speech or natural spontaneous speech. This prohibits an accurate estimation of the performance of the acoustic modeling part of ASR because the language modeling performance is inherently integrated in the overall performance metric (13)

Word Error Rate (WER)is a measure (metric) of the performance of an automatic speech recognition, machine translation etc. The function is intended for calculation of WER between word sequence H (Hypothesis) and word sequence R (Reference). Such a performance is computed by comparing a reference transcription with the transcription output by the speech recognizer. From this comparison it is possible to compute the number of errors, which typically belong to 3 categories as mentioned above i.e.

Insertions I (when in the output of the ASR it is present a word not present in the reference) Deletions D (a word is missed in the ASR output) Substitutions S (a word is confused with another one), Thus Word Error Rate (WER) is given as

WER= (S+D+I)/N

Where N is the number of words in the reference transcription

The main issue in computing this score is the needed alignment between the 2 word sequences. This can be obtained through dynamic programming, using the so called Levenstein distance. For calculation we use Levenshtein distance on word level. Levenshtein distance is a minimal quantity of insertions, deletions and substitutions of words for conversion of a hypothesis to a reference. WER=D(H,R)/N, where D(H,R) is a Levenshtein distance between H and R and N is the number of words in the reference R. H and R are cell arrays of words (for example after using TEXTSCAN) or cells with word sequences or strings. Types H and R may be different.

Minimum Edit Distance it is the distance in words between the ASR hypothesis and the reference transcription

• Edit Distance = (Substitutions + Insertions + Deletions)/N

.• For ASR, usually all weighted equally but different weights can be used to minimize difference types of errors – WER = Edit Distance * 100

$$\text{Word Error Rate} = 100\,\frac{(\text{Insertions} + \text{Substitutions} + \text{Deletions})}{\text{Total Word in Correct Transcript Alignment}}$$

For the analysis of knowing the need of the database for improving the accuracy the following three examples were taken :

Example (1):

An experiment carried out by S.Nareshkumar, N.Mariappan, K.Thirumoorthy , of Dept. CSE, Shenk college, Sivakasi, and found the details as follows :

REF: portable **** PHONE UPSTAIRS last night so
HYP: portable FORM OF STORES last night so
Eval I S S

WER = 100 (1+2+0)/6 = 50%

A small database with 335 words and 1212 sentences finally trained the database with Word Error Rate=7.8% and Sentence Error Rate = 37.5%. The previous word and sentence error rate were 26.9% and 76% respectively , Word Error Rate is a common metric of the performance of a speech recognition. Word Error Rate is the sum of number of substitutions, number of deletions and the number of insertions divided by the number of words in the reference. Sentence Error Rate(12) is the sum of number of substitutions, number of deletions and the number of insertions divided by the number of sentences in the reference.

TABLE 1: Processing a developed system. Pattern Conversion method

No. of speech sources	56 persons
Total no. of samples taken	1335
Total Samples Accepted	1212
Total no. of words used	335

Example 2.

An experiment conducted by N. UshaRani and G. Srinivasulu of Sri Venkateshwara University (14), consisting of Twenty speakers (10 female + 10 male) and recorded the interrogative sentences. Sensitive microphone is used for recording in normal environmental conditions. Each speaker asked to record the same 50 sentences. All speakers recorded their voices in normal health conditions. All speakers are under the age group 20-24 years old. Totally 1000 sentences are recorded for the present work. SPHINX-3 speech recognition system is used for training and decoding.

The experiment they conducted as follows:

10 male speakers speech is trained and 10 female speakers speech is tested on the trained data to get the accuracy of the speech recognition system. Later the process is continued for the 10 female speakers speech is trained for testing 10 male speakers speech consists of 500 sentences. The substitution errors, Insertion errors and deletion errors are responsible for the degradation of the performance of the speech recognition system. Errors that will substitute in the place of correct words are called substitution errors. Some extra words are inserted along with the correct words called as Insertion errors. Some words are deleted while recognition leads to deletion errors. These errors are responsible for the low performance of the speech recognition system [6]. The following is one of the results of the sphinx-3 decoder. Reference (REF) and Hypothesis (HYP) is denoted as:

REF: EYPI EXPRESS TAIMINGS EMITI (ఏ పీ ఎక్స్ప్రెస్ టైమింగ్స్ ఏమిటి)

HYP: SREE ETU EXPRESS TAIMINGS EMITI (శ్రీ ఎటు ఎక్స్ప్రెస్ టైమింగ్స్ ఏమిటి)

In the above, SREE is inserted as additional word and the ETU is substituted in the place of EYPI. This additional word is also present in the pronunciation dictionary. This type of errors leads to low accuracy rate. Let the another result of the sphinx-3 decoder,

REF: PLAATFAAM TICKET DHARA ENTHA (ప్లాటుఫామ్ టికెట్ ధర ఎంత)

HYP: PLAATFAAM TICKET DHARA (ప్లాటుఫామ్ టికెట్ ధర)

In the above example, the ENTHA word is deleted and not recognized which causes deletion errors results low accuracy rate.

SENTENCE RECOGNITION

The following table 2 shows the number of sentences correctly recognized, substitution errors, Insertions errors, Deletion errors.

Table. 2. Sentence Recognition for the Male and Female Speech

	Male speech is tested on Female trained speech	Female speech is tested on Male trained speech
Sentences correctly recognized	157	69
Substitutions	315	396
Insertions	14	8
Deletions	244	167

The following table 3 shows the number of words correctly recognized, substitution errors, Insertion errors, Deletion errors.

WORD RECOGNITION

Table. 3. Word Recognition for the Male and Female Speech

	Male speech is tested on Female trained speech	Female speech is tested on Male trained speech
Words recognized correctly	1301	646
Substitutions	492	482
Deletions	577	1242
Insertions	14	8
Confusion Pairs	195	167

Some words are recognized (substituted) in the place of original words, but these words will present in the dictionary (lexicon) used in the speech recognition system. These pair of words is called confusion pair. This confusion occurs because of the similarity present in the phoneset. Word accuracy rate is the ratio of the number of words recognized correctly to the total number of words in the reference. Word error rate is the ratio of the sum of the substitutions, insertions and deletions to the total number of words in the reference.

The above result is represented in the pictorial representation as below:

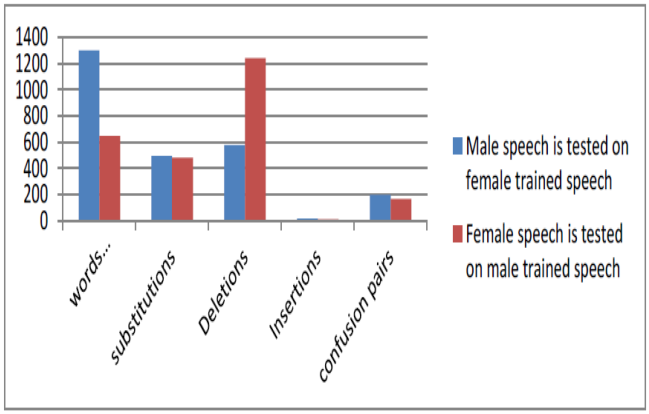

Fig. 4. Word recognition between male and female speech

From the above figure, the experimental results show that insertions are observed to be very less in number when compared with substitutions, deletions. The percentage of word recognition is the number of words recognized correctly to the total number of words in the hypothesis. The percentage of word error rate is 100 * [1- (N-D-S-I)/N] where N is the total number of words in the hypothesis D is the number of words deleted S is the number of words substituted I is the number of words inserted If the male speech is tested on the female trained speech, the percentage of word recognition accuracy is 54.895% and the percentage of word error rate is 45.696%. If the female speech is tested on male trained speech, the percentage of accuracy rate is 27.257% and the percentage of word error rate is 73.08%. Some words are confused between various nouns, verbal words are confused in the place of nouns and nouns are confused in the place of verbs. In the present work, it has been observed the more confusion occurred between the nouns and also more deletions taken place when female speech is tested on male trained speech.

VI. CORPUS OF THE TELUGU DATABASE

The intention of creating a Telugu language speech database for speaker recognition is to obtain rich voice messages with respect to measure inter and intra speaker variability. Subjects are recruited in a normal environment. Most of them are native Telugu speakers.

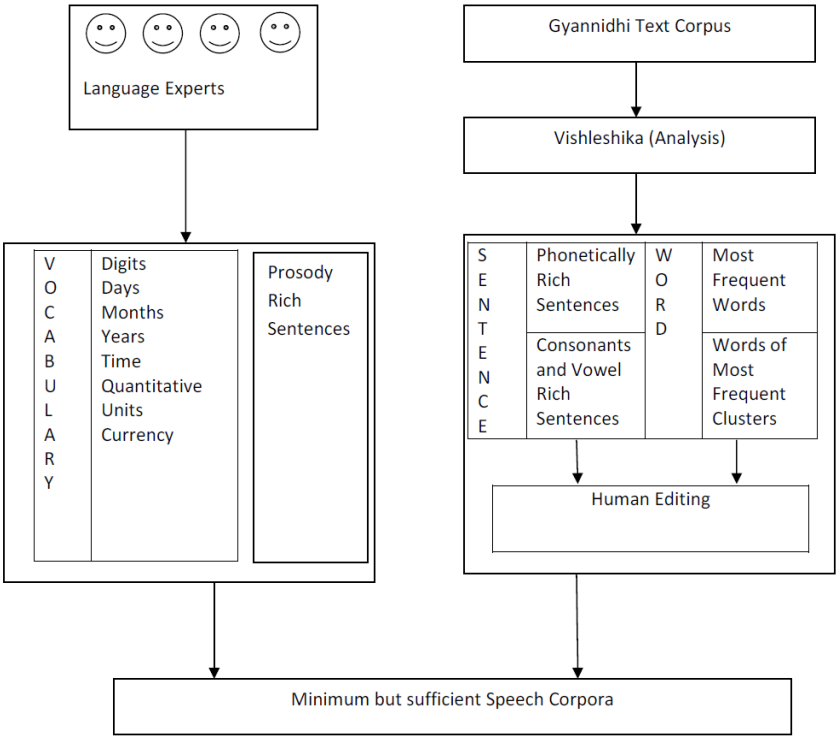

Fig. 5. Corpora Development Methodology (Reproduced)

Preparing Telugu Speech Corpora

The various steps involved in building Telugu Speech copora are , the recording media is chosen so as to capture the effects due to channel and microphone variations. For the databases that were built for the Indian language ASRs, the speech data can be recorded over calculated number of landline and cellular phones using a multi-channel computer telephony interface card and any other medium. The following procedure is followed:

1. Speaker Selection

Speech data is collected from the native speakers of the language who were comfortable in speaking and reading the telugu language. The speakers were chosen such that all the diversities attributing to the gender, age and dialect are sufficiently captured. The recording is clean and has minimal background disturbance. Any mistakes made while recording have been undone by re-recording or by making the corresponding changes in the transcription set. This database has to be evaluated for speaker recognition (11), and did provide a good speaker recognition rate which can be from desired and optimal number of Telugu speakers.

2. Data Statistics and Recording Set ups

Speakers from various parts of the respective states (regions) should be carefully recorded in order to cover all possible dialectic variations of the language. Each speaker has to be recorded optimal sentences of the optimal text. During recording sessions, speech data was recorded simultaneously through a microphone and a telephone line. During recording sessions, volunteer speakers were seated at a table with a telephone set and a microphone (connected to a laptop) and were required to speak into the microphone and telephone simultaneously. The microphone rested on the table near the speaker's

mouth and the telephone receiver had to be held up to the speakers ear by hand. These were depicting the situation in which the ASR system being developed will eventually be used. Table 3 gives the number of speakers recorded in each language and in each of the recording modes landline and cellphones. To capture different microphonic variations, four or more different cellphones can be used while recording the speakers. They can be tabulated in the tables as shown. Table 4 gives the age wise distribution of the speakers in the language. Table 5 shows the gender wise distribution of the speakers of the language. The recording procedure can be done as follows :

Recording hardware. A laptop or any storgae device can be used to record speech through a USB Desktop microphone. and the recordings can be captured through A Linksys SPA400 telephony gateway

Recording software. Praat software can also be used on the laptop to capture and manage the speech received over the microphone. Microphone speech was recorded at 16 kHz and stored in .wav format. Telephone speech can be recorded at 8 kHz and managed through Trixbox, an Asterisk-based PBX phone system

Recording locations. Office rooms and a student lab were used to conduct the recordings. External noise in the office environment was contributed by the opening and shutting of doors and drawers, people talking, printers, telephones etc. Recording sessions were conducted in the student lab almost always when it was completely empty. In addition to these environments, some recording sessions were also conducted in home environments.

3.Speech Data Processing and Segmentation

This section describes how the speech data was processed after it had been acquired through the recording session. Recorded speech from volunteer speakers was manually split into smaller portions, about 10 seconds long, using Praat (15), such that they were suitable for use as training data for CMU Sphinx speech recognition toolkit (16) The basic rule followed during this process was 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 10 seconds long) can be produced. Any portions that included disruptive noises, such as a telephone ring, a drawer opening or closing, or someone else speaking, in close proximity to the speaker, were marked as unusable for the training process.

4. Transcription Correction

Even with the care taken to record the speech with minimal background noise and mistakes in pronunciation, some errors have crept in while recording. These errors had to be identified manually by listening to the speech. If felt unsuitable, some of the utterances have been discarded. In the case of the data collected in the three Indian languages, the transcriptions were manually edited and ranked based on the goodness of the speech recorded. The utterances were classified as Good , With Channel distortion , With Background Noise and Useless whichever is appropriate. The pronunciation mistakes should be carefully identified and if possible the corresponding changes were to be made in the transcriptions so that the utterance and the transcription correspond to each other. The idea behind the classification was to make the utmost utilization of the data and to serve as a corpus for further related research work.

The Silence, Vocalization and Breath tags can be defined to represent non-speech areas in the segments. All silences or pauses during speech as audible or viewable in the waveform displayed on Praat (15) shall be 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 shall be marked by a vocalization tag within the transcription. Breath sounds identified within segments to be marked with a breath tag.

VII. SPEECH CORPORA FOR OTHER INDIAN LANGUAGES

The process of converting the acoustic signals in to text is defined as speech recognition. The main components of speech recognition system are acoustic model, language model and lexicon. It consists of two phases namely training and decoding (recognition). SPHINX-3 is one of the most important speech recognition systems which are used for the research work [2]. It is based on Hidden Markov Model (HMM). HMM is a method of directly estimating the conditional probability of an observation sequence for the given a hypothesized sequence of data.It consists of observed sequence and the state sequence. The state sequences are hidden,

hence the name Hidden Markov Model. HMM involve Evaluation problem, Training and Decoding. Evaluation problem is solved by using the forward algorithm to maximize the probability of observed sequence.

The Linguistic Data Consortium for Indian Languages (LDC-IL) is the Consortium established for developing a similar activity like Linguistic Data Consortium (LDC) at the University of Pennsylvania. The services of LDC-IL are been hosted and Managed by CIIL Mysore. It is also supported by the Central Government India. The LDC-IL will be responsible to create the database but will also provide forum for the researchers all over the world to develop speech application using the collected data in various domains. The LDC-IL has collected Speech databases in various Indian Languages. The table 3 shows the Speech corpus collected by LDC-IL in hours (20).

Table 6. . Showing the length of the Speech corpora available (Source LDC-IL)

SPEECH CORPORA (Raw Data) As on July 2014		
Sl No.	**Languages**	**Hours**
1	Assamese	105:51:38
2	Bengali	138:18:47
3	Bodo	201:10:48
4	Dogri	111:32:11
5	Gujarati	156:23:04
6	Hindi	269:09:50
7	Indian English Bengali	34:12:57
8	Indian English Guajarati (MP3 Format)	21:40:00
9	Indian English Kannada	37:01:33
10	Kannada	198:51:03
11	Kashmiri	44:59:07
12	Konkani	195:14:47
13	Maithili	95:59:54
14	Malayalam	265:24:18
15	Manipuri	187:35:13
16	Marathi	168:13:50
17	Nepali	145:04:46
18	Oriya	165:30:05
19	Punjabi	187:53:28
20	Tamil	213:37:27
21	Telugu	50:51:36
22	Urdu	124:19:58

Table: Database of Indian Languages Speech corpora (Source LDC -IL)

the speech database that are been developed for speech recognition system, text to speech synthesis system in some Indian languages re compared with that of the instruments used for recordings, number of speakers, language, type of speech, the recording environment, language in which database is created and the application of the database. The table 4 shows the basis on which we have compared these different speech databases. The developed speech databases are either for general purpose application or for task specific application.

Sr. No.	Database Developed by	Recording Environment	No. of Speakers	Recording Device Used	Application of Database	Language
1.	TIFR Mumbai and IIT Bombay	Noisy Environment	1500	Cell Phone & Voice Recorders	Speech Recognition System for Agriculture Purpose	Marathi
2.	C-DAC Noida	Noise Free Environment	30	Standard Mics	Speech Recognition System for Travel Domain	Hindi
3.	IIIT Hyderabad	Noisy Environment	96	Mobile Phones	Speech Recognition System for Agricultural commodity Price Enquiry	Telugu
4.	IIIT Hyderabad	Noise Free Environment	15	Standard Microphone and Laptop	Travel and Emergency Services	Telugu, Hindi & English
5.	Government P.G. College, Rishikesh	Noisy Environment	100	Standard Microphones	Speech Recognition System	Garhwali
6.	Punjabi University, Patiala	Studio Environment	1	Standard Microphone	Text to Speech Synthesis System	Punjabi
7.	Utkal University, Bhubaneswar	laboratory environment	Not Known	Noise Cancellation Microphone	Text to Speech Synthesis System	Hindi, Odiya, Bengali & Telugu
8.	TIFR Mumbai and C-DAC Noida	Noise Free Environment	100	Standard Microphone	General Purpose	Hindi
9.	IIT Kharagpur	Studio Environment	92	Standard Microphone	General Purpose	Hindi, Telugu, Tamil, & Kannada
10.	Islampur, Maharashtra	Laboratory Environment	Not Known	Standard Microphone	Text to Speech Synthesis System	Konkani (Goan)
11.	SJ College of Engineering, Mysore	Laboratory Environment	Not Known	Standard Microphone	Text to Speech Synthesis System	Kannada
12.	KIIT, Bhubaneswar	office environment	200	Cellphone, Omni directional & cardioid Microphone	Mobile based speech recognition	Hindi & Indian Spoken English
13.	IIIT Hyderabad and HP Labs Bangalore	Noisy Environment	559	Mobile Phone and Landline	General Purpose	Marathi, Tamil & Telugu

Table 7. . Comparison of Databases of Indian Languages (Source 19.)

VIII. DISCUSSIONS

From the above two table it was found that Indian Language speech corpora is still at the initial stages and a lot of inputs are needed. Although Telugu is the second largest spoken language, The speech corpora for the Telugu Language is just 50 Hours compared to Malyalam and Hindi with 265 and 269 Hrs respectively. From the Fig 2 above it mentions as maximum applications for General Purpose only, and no reference for any Disability applications

IX. CONCLUSION

1.Existing speech recognition systems are mostly command based recognition systems[2] .The proposed system gives a common framework for applications which involves database operation. It increases the interaction between user and computer. Users can be relieved from the complexities of the query. Sphinx4[5] toolkit have been studied for speech recognition. Converting the textual representation of queries into standard SQL format is done for some standard SQL queries.. The proposed system is being implemented to verify the given concept.

2. The present work observed that the testing female speech on the male trained speech gives low performance of the speech recognition system. The low accuracy rate is due to the differences in pitch, formant frequencies etc., between male speech and female speech. In speech recognition system, the training speech data should be the combination of the female and male speech, then the models will be developed for both the male and female speech in acoustic model, then the recognition accuracy will be improved. Still the accuracy rates can be improved by incorporating semantic rules during decoding phase.
3. Existing speech recognition systems are mostly command based recognition systems[2] .The proposed system gives a common framework for applications which involves database operation. It increases the interaction between user and computer. Users can be relieved from the complexities of the query. Sphinx4[5] toolkit have been studied for speech recognition. Converting the textual representation of queries into standard SQL format is done for some standard SQL queries.. The proposed system is being implemented to verify the given concept
4. In this paper demonstrate the results of recognition give user option to select the most accurate result using POS tagging. This work is planned to implement the model of speech recognition for matching user speech and the output results. This system is active on the Internet. If the connection is lost, this application will not be used. This work is especially for different kinds of places such as pagoda, hotel, market, etc. in our country. That is why the words are related to our Myanmar phonetics. This work presents a framework for extraction of linguistic details from standard word error rates WER and Leveshtein distance use for automatic error analysis. This paper demonstrated an approach for the decomposition of the standard word error rates, inflectional error analysis, missing words analysis over ten POS classes.
5. . Future work would include increasing the corpus by adding speech from new speakers and also improving the process in order to capture more speech per recording session. One suggestion is to use a set of objects or pictures during the recording session and to ask speakers to describe them. Volunteers may speak more freely with this method as opposed to the question-answer style adopted in this work.

X. REFERENCES

1. Gopalakrishna Anumanchipalli, et al , International Institute of Information Technology, Hyderabad, India & Hewlett Packard Labs India, Bangalore, India Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems, Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece, Oct 2005.

2. Pukhraj P. Shrishrimal ,Ratnadeep R. Deshmukh, Vishal B. Waghmare, Dept. of CS and IT Dr. B. A. M. University, Aurangabad-431004, India, Indian Language Speech Database: A Review, International Journal of Computer Applications (0975 – 888) Volume 47– No.5, June 2012

3. Samudravijaya K., P. V. S. Rao and S. S. Agarwal. 2000. Hindi Speech Database. In Proceedings of Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.

4. Sanghamitra Mohanty, "Syllable Based Indian Language Text To Speech System", International Journal of Advances in Engineering & Technology, 2011. Vol. 1,Issue 2.

5. Parminder Singh, Gurpreet Singh Lehal. 2006. Text-ToSpeech Synthesis System for Punjabi Language. In Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain.

6. Singh, S. P., et al Building Large Vocabulary Speech Recognition Systems for Indian Languages , International Conference on Natural Language Processing, 1:245-254, 2004.

7. Mansour M Alsulaiman et al, Kinf Saud University, Saudi Arabia, IEEE Conference paper 2011

8. Ling Feng and Lars Kai Hansen, Informatics and Mathematical Modelling, Technical University of Denmark, A NEW DATABASE FOR SPEAKER RECOGNITION

9. Deller, J.R., Hansen, J.H.L., Proakis, J. G., "DiscreteTime Processing of Speech Signals", IEEE Press, New York, NY, 2000.

10. "Home Page of the VidTIMIT Database", 2001. http://rsise.anu.edu.au/~conrad/vidtimit/

11. Feng, L., "Speaker Recognition", Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2004

12. Yousafzai, J. , Cvetkovic, Z. , Ager, M , Sollich, P —Redundancy in speech signals and robustness of automatic speech recognition XIII International Symposium on Problems of Redundancy in Information and Control Systems (RED), 2012.

13. Amit Juneja et al, A comparison of automatic and human speech recognition in null grammar- The Journal of the Acoustical Society of America - published February 2012

14. N. Usha Rani, G. Srinivasulu, Comparison of Telugu Speech Recognition Accuracy among the Male and Female Speech,Department of CSE, SVU College of Engineering, Tirupati, India, International Journal of Engineering Sciences Research-IJESR, Vol 04, Special Issue 01, 2013

15. Praat: doing phonetics by computer, www.fon.hum.uva.nl/praat, accessed June 2010.

16. CMU Sphinx Open Source Toolkit for Speech Recognition Project by Carnegie Mellon University,http://cmusphinx.sourceforge.net/, accessed June 2010.

17. R. K. Upadhyay and M. K. Riyal. 2010. Garhwali Speech Database. In Proceedings of O-COCOSDA 2010, Kathmandu, Nepal.

18. D. J. Ravi and Sudarshan Patilkulkarni, "A Novel Approach to Develop Speech Database for Kannada Text-to-Speech System", Int. J. on Recent Trends in Engineering & Technology, 2011, Vol. 05, No. 01, in ACEEE.

19. Pukhraj Shreemal, Deshmukh et al, - Indian Language Speech Database - A Review, International Journal for Computer Applications (0975-888), Vol.44, No.5 June 2012

20. Size of Speech Corpora (As on Dec 2011) , Available at: //www.ldcil.org/resourcesSpeechCorp.aspx

21. Gautam Varma Mantena et al 2011. A Speech- Based Conversation System for Accessing Agriculture Commodity Prices in Indian Languages.In Proceeding of Joint Workshop on Handsfree Speech Communication and Microphone Arrays (HSCMA), Edinburg, Scotland.

22. Anandaswarup V, Karthika M, Nagaswetha G, et al 2010. Rapid Development of Speech to Speech Systems for Tourism and Emergency Services in Indian Languages. In Proceeding of International Conference on Services in Emerging Markets, Hyderabad, India.

23. Melin, H. (2000), "Databases for Speaker Recognition,Activities in COST250 Working Group 2", In: COST250 - Speaker Recognition in Telephony, Final Report 1999 (CD-ROM), European Commission DG-XIII, Brussels, August 2000.

24. Cole, R., Noel, M., Noel, V. (1998). "The CSLU speaker recognition corpus", *ICSLP'98*, Sydney, Australia, November 30-Decenber 4, pp.3167-3170.

25. Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J.G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," *NIST*, 1993.