Indian English speech recognition system for commonly used English Words using CMU tools

Ms. Jasleen Kaur
Department of Computer Science and Engineering,
Research Scholar of Computer Science and Engineering,
Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib

Abstract: This paper is based on work done for the development of a speaker-independent Indian English speech recognition system. The work of recognition is performed using Sphinx tools. The database is a collection of Most Commonly used English words in routine. In this research initially, the system accuracy is evaluated using Carnegie Mellon university tools for the different age groups individually and later on these individual systems are combined together to compute the overall accuracy of the system. The system obtained best performance of 85.80 % for young (Eng2) model when trained using 16 GMMs (Gaussian Mixture Models). Also, the performance of the combined model is computed for different GMMs.

Keywords: Speech recognition; Indian English language; GMMs; CMU Sphinx; Acoustic model; Language model.

I. INTRODUCTION

A. *Automatic Speech Recognition*

Automatic speech recognition is a process in which an acoustic speech signal recorded from a speaker is converted into text by the computer [21]. Human beings have been inspired so long to generate a system that can understand and talk like a human. Since 1960s, scientists have been researching various methods, so as to make computer record, understand and interpret human speech [9] ASR is a process in which a computer takes human speech as an input and tries to convert it to a corresponding set of words using a specific algorithm.

Automatic Speech Recognition (ASR) is a system where a computer takes a speech from recorded audio signal and then converts it into the corresponding text [8]. Getting a computer to understand and react appropriately to spoken language [9]. It is the process by which a computer recognizes what a person said. The Speech recognition process involves an acoustic signal captured by a microphone and then it is converted to a set of words accordingly [29]. A computer system is enabled to identify and respond to the sound produced during human speech in known as speech recognition.

B. *Classification of Speech Recognition System*

ASR systems are essential part of various research field and there are different ASR systems found in literature. ASR systems [22] can be classified into following types:

1) Based on utterances

a) Isolated Words

This recognition system recognizes only a single word at one time. User needs to give only one word response or command. The main advantage of this system is: It is simple and easy to implement because word boundaries are obvious that can be easily detected and the words are pronounced very clearly [3].

b) Connected Words

This system is same as an isolated word system, but it permits separate words to run-together with a minimum stop between them.

c) Continuous Speech

This recognition system permits an individual to speak almost in a natural manner, during which the system computes its content. Generally, it is a computer dictation where closest words run together without any pause or division between them. Such systems are more complex.

d) Spontaneous Speech

In spontaneous speech recognition, system recognizes the natural speech. Spontaneous speech is a natural speech that comes suddenly through mouth. A spontaneous speech ASR system is capable of handling a variety of natural speech features i.e. words being run together along with mispronunciation, stutters and false starts etc.

2) Based on Speaker Model

Speech recognition system can be divided into three main categories as follows:

a) Speaker Dependent Models

Speaker-dependent system works only for a particular type of speaker. They are more accurate for a particular speaker, but are less accurate for other type of speakers. They are cheaper and are easier to develop. But they are not as flexible as speaker-independent systems. They can be used for security purpose.

b) Speaker Independent Models

Speaker-independent system can recognize a variety of speakers without any prior training. It can be used in Interactive Voice Response System (IVRS) that must accept input from a large number of different users. But it limits the number of words in a vocabulary and implementation is also very difficult. It is expensive and it is less accurate than speaker-dependent systems.

c) Speaker Adaptive Models

In such systems, the speaker-dependent data is used and is matched with the best-suited speaker to recognize the speech and to decrease an error rate after adaption [12]. They adapt operation according to characteristics of the speakers.

3) Based on Vocabulary

The size of vocabulary can affect the complexity, processing rate and the rate of recognition of ASR system. ASR systems are classified as follow:

- Small Vocabulary: contains 1 to 100 words or sentences.
- Medium Vocabulary: contains 101 to 1000 words or sentences.
- Large Vocabulary: contains 1001 to 10,000 words or sentences.
- Very-large vocabulary: contains more than 10,000 words or sentences.

C. Block Diagram of ASR System

The main components of a typical ASR system found in most of the applications are shown in "Fig. 1".

1) Input Speech

It is basically the recorded acoustic signal from different speakers. The acoustic signal is an analog signal. The analog signal cannot be directly transferred to the ASR system. So these signals are transformed into digital signal. This digital signal can now be processed.

2) Feature extraction

It helps to find the set of parameters of utterances that have acoustic relation with speech signals. These parameters are called features. The main goal of feature extractor is to discard the irrelevant information and keep only relevant one. There are several methods for feature extraction such as Mel-Frequency Cepstral Coefficient (MFCC) [6], Perceptual Linear Prediction (PLP), Linear Predictive Cepstral Coefficient (LPCC) and RASTA-PLP (Relative Spectral Transform) [17]etc.

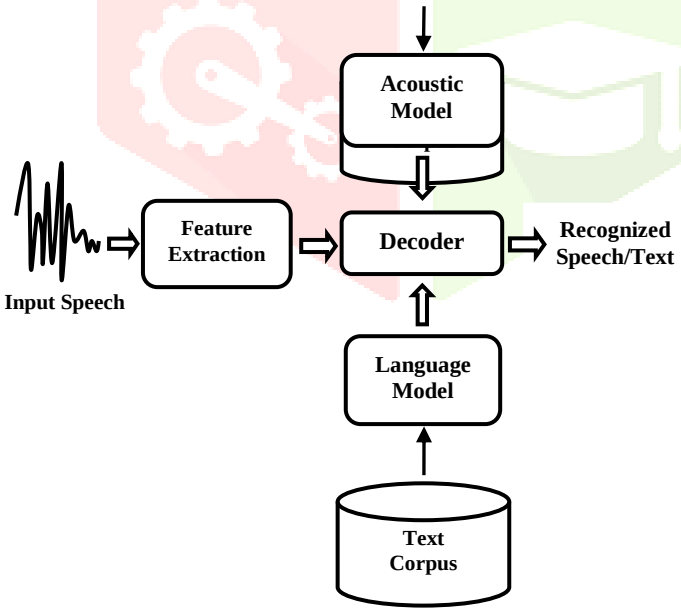

Figure 1 Block diagram of ASR System

3) Acoustic model

Acoustic modeling is the fundamental part of ASR system [1]. It is the main part of Training. The acoustic model provides a connection between the acoustic Information and phonetics. Acoustic model plays an important role in the performance of the ASR system and is responsible for computational load [12]. Acoustic model uses speech signals from training database. Hidden Markov Model (HMM) is widely used and accepted model [7] as it is efficient algorithm for training and recognition.

4) Language model

It is also the part of training. A language model contains the structural constraints available in the language to generate the probabilities of occurrence of a word followed by the sequence of n-1 words [21]. Various models are used to find the exact word

sequence like bi-gram, tri-gram, n-gram language models. This is done by predicting the likelihood of the nth word, using the n-1 earlier words. The language model finds and differentiates between word and phrase that has similar sound.

5) Decoder

Decoding (or recognition) is the process of comparing the unknown test pattern with each sound class reference pattern and computing the similarity between them to find the best match. After the completion of training phase, system testing is performed. In testing phase patterns are classified to recognize the speech. Finally, the output of this sub-system is text.

The paper is organized as follows: Sect. 2 presents a brief description of the Indian English language. In Sect. 3, presents the problem being formulated and major contribution of the proposed system. Sect. 4, presents a description on the Indian English speech recognition system and investigations to adapt the system to Indian English language. Section 5 investigates the experimental results. Finally, Sect. 6 presents conclusions and future scope.

II. INDIAN ENGLISH LANGUAGE

English is a bridge language of India. Even English is used as one of the two official languages in India, only a few hundred/thousand Indians use English language as first language. Idiomatic forms and vernaculars are absorbed into Indian English that are derived from Indian languages. But, the Indian English dialect remains homogeneous in vocabulary, phonetics and phraseology between variants.

Phonology: Indian accent for English language varies greatly from state to state. Some Indians speak English in an accent closer to British English. While the other Indians use vernacular, native-tinted accent for English language.

Vowels: It has been noticed that there are lot of differences in the way a vowel is pronounced in Indian English than that of British English and American English. Some of the common differences are discussed below:

There is only one phoneme /ə/ in Indian English that corresponds to British English phonemes/ʌ/, /ɜː/ and / ə /; i.e.

<sir> /sɜː(r)/ /sə(r)/

<pup> /pʌp/ /pəp/

<fear> /fɪə(r)/ /fɪə(r)/

Indian English has /ɒ/ corresponding to BE /ɒ/ and /ɔː/; i.e.

<raw> /rɔː/ /rɒ/

<pot> /pɒt/ /pɒt/

RP has two distinctive back vowels /ɒ/ and /ɔː/ whereas

Indian English has /ʊ/. Thus, in Indian English there is no difference in the pronunciation of:

<cot> /kɒt/ /kʊt/

<caugh> / kɔːt/ /kʊt/

Consonants: The distinctive features of consonants in Indian English are:

Pronunciation of consonants varies between rhotic and non-rhotic. Pronunciation with native phonology are rhotic and others being non-rhotic (imitative of British pronunciation). The consonant system of Indian English consists of 23 consonants including: Bilabial – /m/, /p/, /b/, Labio dental - /f/ , Dental - /t/, /d/, Alveolar - /n/, /s/, /z/, /l/, /r/, Palatal-alveolar - /ʃ/, /ʒ/, Palatal - /j/, /tʃ /, /dʒ /, Velar - / ŋ /, / k /, / g /, Retroflex -/T/, /D/, /N/, Glottal - /h/.

All native Indian languages lack the voiced palatal or post alveolar sibilant /dʒ/, as in 'treasure'. In Indian English, /w/ is generally replaced by /v/ because it doesn't point out any difference between /v/ voiced labio-dental fricative and /w/ velar semi-vowel. Many Indians pronounce a frictionless labio-dental approximant near to /v/ for both /v/ and /w/ graphemes (i.e. wine is spoken as vine and what as vat). Indian English lack the phonemes voiceless dental fricative / θ / and voiced dental fricative /ð/. So, the aspirated voiceless dental plosive [t̪ʰ] is substituted as /θ/ and the un-aspirated voiced dental plosive /d̪ʱ/ is substituted for /ð/. This creates confusion to native speakers. Phonemes /p/, /t/ & /k/ are slightly aspirated in British English when used in a word or syllable at initial position. While, in most of the Indian languages, the difference between non-aspirated and aspirated plosives is phonemic.

So, Indian English uses the corresponding non-aspirated voiceless plosives /p/, /t/, /k/ instead of /ph/, /th/ and /kh/. In place of corresponding alveolar plosives /t/ and /d/, Indian English uses retroflex plosives /T/ and /D/. Some of the Indian languages lack affricates. So, Indian pronunciation of English affricates /tʃ/ and /dʒ/ are corresponding to palatal plosives without the following friction. In the speech of some of the English speakers, syllables /l/, /m/, /n/ are usually replaced by voiced consonant clusters. In BE /r/ occurs only before a vowel whereas Indian English implies a very sharp as well as clear alveolar trill / ɾ / in all word positions.

III. RELATED WORK

Anukriti Tiwari and Bhattacharya [2] designed a method for a speaker-independent automatic speech recognition system that can be interpreted in any of the Indian languages (Evaluated for Hindi language and Bengali) and finally implemented on Windows 7 system. A speaker-dependent digit recognition system for isolated English digits was designed by Ganesh Pawar and Sunil Morade [19] using a database of 50 speakers. HMM was used as classifier and MFCC for features extraction. Training and testing purposes is done using HTK tool kit. The system provides an accuracy of 95 percent.

Recently, Satori and ElHaoussi [24] investigated the speaker-independent continuous speech recognition system in Amazigh language. The system is based on the CMU Sphinx tools. In the training and testing phase, Amazigh_Alphadigits corpus was

used. It consists of 60 Berber Moroccan speaker's speech and their transcription (30 males; 30 females) who are the native of Tarifit Berber. This system has ben working with a performance of 92.89 percent when it was trained on 16 GMMs. Joshi and Rao [11] worked on pronunciation assessment of vowels of Indian English uttered by speakers with Gujarati using confidence measures obtained by automatic speech recognition. It has been noticed that Indian English is represented more accurately by Hindi speech-models instead of American English speech-models. An isolated speech recognition system for English digit was developed by Limkara et al. [15] using MFCC and Dynamic time wrapping algorithm in which the system works with an accuracy rate of 90.50 percent. They provide a comparative study on the speaker-dependent and speaker-independent speech recognition by the designed digit recognition system. Instead of using MATLAB they used HTK for speech recognition. HTK is well-known open-source software but MATLAB is a commercial product. MFCC is used for extracting features because it is one of the effective algorithms. Hidden Markov Model is used as classifier instead of DTW algorithm because it is an easy method and it provides more accuracy in the recognition process. Finally for both the speaker-dependent and speaker-independent systems are compared on the basis of accuracy. Ma and Paulraj [16] worked on three accents of English language recorded from three main ethnicities in Malaysia namely Malay, Chinese and Indian. They used Mel-bands spectral energy as the statistical descriptors and neural network as a speech recognizer. They performed these experiments on three different independent datasets of 20%, 30%, and 40% of total samples. They obtained an average recognition rate of 95.59%. Mishra et al. [18] proposed a system on comparative study of isolated digits in Hindi language by using HMM & MFCC algorithm for extracting features. They designed the system for both HTK and MATLAB individually. Using HTK system provides an accuracy of between 99-100 percent which is 5 to 6 percent better as compared to MATLAB. In noisy environment, HTK give an accuracy rate from 89 to 94 percent. Phull and Kumar [20] worked on Large Vocabulary Continuous Speech Recognition (LVCSR) system for Indian English (IE) video lectures using CMU Sphinx tools. Speech data was video lectures on different Engineering subjects given by experts from all over the India as a part of NPTEL project of 23 hours. They obtained an WER of 38% and 31%, before and after the adaption of IE acoustic model respectively. The Results were comparable to American English (AE) and were 34% less than average WER for HUB-4 acoustic model. Cole et al. [5] developed a speaker-independent spoken English alphabet recognition system. The system was trained on one token for each alphabet from 120 speakers. Then it is tested on a new set of 30 speakers. They obtained a performance of 95%. But the performance is increased to 96% when tested on a second token of each letter from the same 120 speakers. Sarada et al. [23] worked on group delay based algorithm so as to automatically segment and label the continuous speech signal into syllable-like units for Indian languages. They used a new feature extraction technique. This technique uses features that are extracted from multiple frame sizes and frame rates. They obtained recognition rates of 48.7% and 45.36% for Tamil and Telugu languages respectively. Toma et al. [25] developed the system which describes the effect of Bengali accent on English vowel recognition. They noticed that Bengali-accented speech has a large influence on the spectral characteristics of different English vowel sounds. Walha et al. [26] proposed an approach on developing an HMM based ASR system for Standard Arabic Language to select the most appropriate acoustic parameters describing each audio frame, acoustic models and speech recognition unit. They analyze the effect of varying frame windowing (size and period), acoustic parameter number obtained from features extraction methods, number of embedded re-estimations of the Baum-Welch Algorithm, speech recognition unit and Gaussian number per HMM state. The corpus used is multi-speaker SA connected-digits. The corpus is transcribed and used in all experiments. Also the system is evaluated for speaker-independent continuous SA speech corpus. They obtained the phonemes recognition rate of 94.02% which is relatively high when compared to other ASR system using the same corpus.

IV. MOTIVATION FOR WORK

A. Problem formulation

ASR systems that have been developed so far are working online. Online systems permit you to work from any vendor, at anytime, anywhere in the world. But they require continuous and reliable internet connection. On the other hand, Offline systems have the ability to work even they are disconnected from the internet. Also they are fast, responsive and productive. Also in offline systems, there is no such system developed that can recognize the commonly used English words in north-west Indian English accent.

The systems that have been developed so far only uses other north Indian languages like - Hindi, Punjabi, Bengali, Bhojpuri, etc. but not English language. So it is proposed to develop a system that uses Indian English language for recognizing commonly used English words but in an accent used by native of Punjab (north-west region). So it is proposed to develop an Indian English (IE) acoustic model for training the ASR system. Also the phonetic dictionary used in the proposed system is focused on north-west Indian English accent. Speech recognition changes with age. So it is proposed to develop a system based on different age groups like – child, young, adult and old. Also a combine model is proposed.

B. Major contribution

The proposed system provides an Indian English (IE) acoustic model for training the speech recognition system for Most Commonly used English words in an accent used by native of Punjab (north-west region). Also the proposed system can be used by people of all age groups. The proposed system will work offline. So no internet connection is required. The proposed system can be used by physically disable people. This system can further be used in various applications.

V. INDIAN ENGLISH SPEECH RECOGNITION SYSTEM

A. *System overview*

This speech recognition system works on four different age groups namely – Children, Young, Adult and old. In this speech recognition system, initially data preparation is done in which speech recordings are collected from 76 speakers. The corpus consists of speech recordings of 500 most commonly used English words from each of the 76 speakers. Then the acoustic model and language model are built. Then the phonetic dictionary was made using 500 most commonly used English words [27] and their transcriptions. Both training and recognition are based on CMU Sphinx system. It is HMM-based, speaker-independent, isolated continuous speech recognition system capable of handling large vocabularies (CMU Sphinx Open Source Speech Recognition Engines) [10]. This approach of modeling Indian English sounds in CMU Sphinx system consists of generated and trained acoustic model along with language model and dictionary of Most Commonly used English words with their speech transcriptions.

B. *Speech database preparation*

The corpus, "Most Commonly used English words", is used in this work and it contains speech and their transcription of 76 Punjabi speakers. The corpus consists of spoken 500 words collected from each of the 76 speakers. The audio files were generated by speakers pronouncing the words in alphabetical order. So as to make the task of labeling speech signals easy. The sampling rate of the recording is 16 kHz, with 16 bits resolution. "Table 1" shows more speech corpus technical details.

Table 1 System parameters

Parameter	Value
Speaking mode	Isolated words
Sampling rate	16 kHz
Enrolment (or Training)	Speaker-independent
Vocabulary size	Medium (500 words)
Equipment	Good quality microphones and a Smart Voice Recorder application in mobile.
Speaking style	Read (dictation)
Number of channels	1, Mono
Audio data file format	.wav
Corpus	500 words
Number of speakers	76
Speakers' age	Age groups – children, young, adult, old
Accent	North-West Indian English
Rule set	20% of the total speech corpus
Size of training set	80% of the total speech corpus
Number of tokens	Total 38,000 tokens (500 tokens per speaker)

During the recording sessions, speakers were asked to utter the English words sequentially. Audio recording for a single word is saved into one ".wav" file. So, total 500 ".wav" files are stored for a single speaker and the same process is performed by all of the 76 speakers. It is time consuming to save every single recording once uttered. Hence, depending on this, the corpus consists of 38,000 tokens. Wrongly pronounced utterances were ignored and only correct utterances are kept in the database. "Table 2" shows the age groups used, the models named on these age groups and number of speakers in each group of this system.

Table 2 Age categories

Age group	Age (in years)	Model	No. of Speakers
Children	Below 15	Eng1	20
Young	15 to 25	Eng2	20
Adult	25 to 60	Eng3	20
Old	60 above	Eng4	16
ALL	Any	ENG	76

C. Pronunciation dictionary

This dictionary is also known as lexicon. It contains 500 most commonly used English words and their pronunciation (phonetic-transcription) based on the accent used in north-west region for Indian English (IE). This dictionary is created after a deep study on IE phonetics and then different rules are used to pronounce each word. "Table 3" shows the phonetic dictionary list for few words used to train the system. The pronunciation dictionary acts as an intermediary between the Acoustic Model and Language Model.

Table 3 The phonetic dictionary list used in the training

Word	Phones	Word	Phones
A	AY	DIRECT	D R EH K T
ABLE	AY B L	DO	D UW
ABOUT	A B AH U T	DOES	D UH Z
ABOVE	A B UH V		
ACT	EH K T	•	
BACK	B EH K	•	
BASE	B AY S	•	
BE	B E		
BEAUTY	B E Y U T E	YOU	Y UW
CLEAR	K L E ER	YOUNG	Y UH N G
CLOSE	K L O Z	YOUR	Y UW ER
COLD	K O L D		
COLOR	K L ER	So on, Y as last alphabet	
DIFFER	D IH FF ER		

It is assumed that Indian English speech model is better represented by Hindi speech models rather than American English models (US English model) [11]. In this work, a phonetic dictionary has been created on the basis of this concept. Various rules are used to create this phonetic dictionary. "Table 4" shows the basic rules to define some phonemes in an Indian English accent. Later on this dictionary has been validated by an expert.

D. Feature extraction

Feature extraction includes the extraction of speech features recorded from speakers. This sub-system plays a crucial role in the performance of speech recognition system. The parameters used in our system, were 16 KHz sampling as shown in "Table 1".

Table 4 Some Phonetic Rules for Pronunciation Dictionary

Punjabi alphabets	Phonetic transcrip -tion	Phone used	Words
□	ə	A	**A**GO, **A**MONG, **A**PPEAR
□	ɑ:	AH	**A**SK, CL**A**SS, F**A**ST
□	ɪ	IH	B**I**G, D**I**D, F**I**LL
□	i:	E	F**EE**L, F**EE**T, GR**EE**N
□	ʊ	U	B**OO**K, F**OO**T, G**OO**D
□	u:	UW	D**O**, F**OO**D, M**OO**N
□	eɪ	AY	**A**BLE, **A**GE, AG**AI**N
□	Æ	EH	**A**CT, **A**DD, **A**T, B**A**CK
□	əʊ	O	AG**O,** C**O**LD, H**O**LE
□	ɔ:	AW	B**O**X, C**A**LL, TH**OUGH**T
□□	ɜ:r	ER	F**IR**ST, G**IR**L, TOW**ARD**
□□	aɪ	AH E	CR**Y**, FL**Y**, HIGH
□□□	ɔɪ	OI	B**OY**, P**OI**NT, V**OI**CE
□□	aʊ	OW	FOLL**OW,** H**OW,** N**OW**
□	ʃ	SH	FI**SH, SH**APE, **SH**OW
□	ʌ	UH	AB**OV**E, B**U**T, C**U**T

E. Training

The Training of acoustic model is performed using CMU Sphinx tools that uses embedded training method based on the Baum-Welch algorithm [28]. Training is the process of building the knowledge base by learning the Acoustic Model and Language Model used by the speech recognition system.

1) Acoustic model

The acoustic model helps to map the observed features of phonemes (basic speech units) provided by the front-end of the system to the HMMs. The basic HMM model used is 3-states HMMs architecture for each English phoneme includes three states: begin, middle and end state, which join models of HMM units together in the ASR engine, as shown in "Fig. 2". Each emitting state consists of Gaussian mixtures.

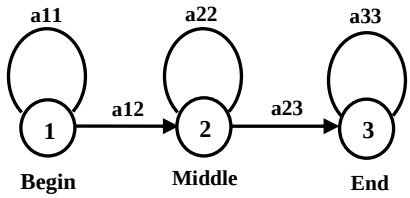

Figure 2 The 3-states HMM model.

The acoustic modeling is done by using speech signal from the training database. In this technique, words in the vocabulary are modeled as a sequence of phonemes, where each phoneme is modeled as a sequence of HMM states. Every recording is converted into a sequence of feature vectors. A set of feature files are generated for each recording using the front-end provided by Sphinxtrain. At this stage, the speech engine uses the phonetic dictionary (see "Table 3") which maps every used English word to a sequence of phonemes. During the training, a phone list is used to take all the phonemes. The phonemes are further refined into Context-Dependent (CD) tri-phones and are added to the HMM set.

2) Language model

In this ASR system, the n-gram language model is used to guide the search for correct word sequence. Search is done by predicating the likelihood of the nth word, using the n − 1 preceding words. The commonly used n-gram models are – uni-gram, bi-gram and tri-gram. The language model is created by computing the word's uni-gram counts, which are converted into a task vocabulary with word frequencies. The bi-grams and tri-grams are generated from the training text based on this vocabulary. In this work, the Cambridge statistical language modeling toolkit (CMU-CLMTK) is used to generate language model of our system [4].

F. Testing

Testing is also called as decoding. It is performed after the completion of training phase. It is necessary to test the quality of the trained database. So as to select the best parameters, to understand how the system performs and to optimize the performance of system. The decoding is a last stage of the training process. The output of the decoding phase is percentage of word error rate (WER) and sentence error rate (SER).

Table 5 System Recognition rate for experiment 1-4

	Recognition rate for 16 Gaussian mixtures (GMMs)		
	Model	WER (%)	Accuracy (%)
Exp1.	Eng1	19.00	81.00
Exp2.	Eng2	14.20	85.80
Exp3.	Eng3	18.40	81.60
Exp4.	Eng4	19.90	80.10

G. Recognition

After the completion of training phase, the acoustic model is generated by the system. The acoustic model can now be used for recognition. Recognition can be done by a recognizer. Sphinx4 and pocketsphinx are the basic recognizers provided by CMU Sphinx tools. Depending on the type of model trained, any of the above recognizer can be used for recognition. In this work, pocketsphinx is used as a recognizer. Speech is given to the system and it is converted into text.

VI. EXPERIMENTAL RESULTS

A. Evaluation Criteria

The performance of the proposed system can be evaluated by the recognition percentage [26] defined by the following formula:

$$\%\text{Recognition} = \frac{N - D - S - I}{N} * 100 \quad (1)$$

$$\%\text{WER} = \frac{I + D + S}{N} * 100 \quad (2)$$

where D, S, I and N are deletions, substitutions, insertions and the total number of speech units of the reference transcription respectively.

B. Result

After the completion of recognition process, the system evaluates results in the form of Word Error Rate (WER) and Sentence Error Rate (SER). The lower these rates are better will be the result. WER should be around 10% for a typical task of 10 hours. But for a large task, it could be around 30 percent. In this paper, the combined model ENG has 57-hours of task (approx. 45 minutes recording from single speaker).

In all the experiments, corpus subsets are disjointed and partitioned to training 80% and testing 20% in order to assure the speaker independent aspect. The system obtained the best performance of 85.80 % for Eng2 model when trained using 16 GMMs (Gaussian Mixture Models). The models for different age groups (i.e. Children, Young, Adult and Old) are named as Eng1, Eng2, Eng3 and Eng4 respectively as shown in "Table 5".

In the last, all models are combined into a single model named as ENG. "Table 6" shows the percentage accuracy of ENG for different GMMs. The combined model obtained best performance of 85.20 % for 128 GMMs.

Table 6 **System overall recognition rate for Combined model (ENG)**

	Overall system recognition rate for different Gaussian mixtures			
	16 GMM (%)	32 GMM (%)	64 GMM (%)	128 GMM (%)
WER	23.50	18.80	17.10	14.80
Recogni--tion	76.50	81.20	82.90	85.20

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we investigated the speaker-independent isolated word ASR system using a database of sounds corresponding to English words spoken in north-west Indian English language. This system is implemented by using CMU Sphinx tools based on HMMs. This system involves creating the speech database for English words, which consist of many subsets used in the training and testing phase of the system.

This work includes creating the speech database English words data of 500 words dictionary (i.e. Medium Isolated Vocabulary Speech Recognition), and also consists of recordings of 76 speakers, recorded using microphone which are used in the training and testing phase of the system.

The system obtained the best performance (accuracy) of 85.80% for Eng2 model (Young age group) which is 4.20%, 4.80%, 5.60% and 0.60% better than Eng3 (Adult), Eng1 (Children), Eng4 (Old) and ENG (combined) models respectively. The minimum %recognition is 80.10% for Eng4 model (Old).

In a future work, the proposed system can be improved by using a large vocabulary (1,000 of words) model. Key research challenges for the future are, use of multiple word pronunciations and the access of a very large lexicon. The obtained results can be improved by fine tuning the system by training with large vocabulary, by increasing the number of speakers for recordings and by the categorizing the speakers on the gender basis. So that the word error rate (WER) can be reduced to 10% and for improving the system accuracy.

VIII. ACKNOWLEDGMENTS

We would like to thank people involved in the development of the CMU Sphinx tools and making it available as open source. Also, big thanks to Prof. Harshvinder Singh (Head, PG department of English, Mata Gujri college, Fatehgarh Sahib) for validating the phonetic dictionary.

IX. REFERENCE

[1] R. K. Aggarwal, "Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement", PhD Thesis, 2012.

[2] Anukriti, S. Tiwari, T. Chatterjee & Bhattacharya, M., "Speaker Independent Speech Recognition Implementation with Adaptive Language Models", *International Symposium on Computational and Business Intelligence (ISCBI)*, New Delhi, India, pp. 7-10, 2013.

[3] G.K. Kharate & S.S. Bhabad, "An Overview on Technical progress in Speech Recognition", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, Issue 2, pp. 488- 497, 2013.

[4] CMU lmtool, Retrieved February 23, 2017, from http://www.speech.cs.cmu.edu/tools/lmtool-new.html., 2017.

[5] R. Cole, M. Fanty, Y. Muthusamy & M. Gopalakrishnan, "Speaker-independent recognition of spoken English letters", *International joint conference on neural networks (IJCNN)*, Vol. 2, pp. 45 – 51, 1990.

[6] B. P. Das & R. Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", *International Journal of Modern Engineering Research*, Vol. 2, Issue 3, pp. 854-858, 2012.

[7] Gaurav, D. S. Deiv, G. K. Sharma, & M. Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", *Journal of Signal and Information Processing*, Vol. 3, pp. 394-401, 2012.

[8] R. E. Gruhn, W. Minker & S. Nakamura, "Statistical Pronunciation Modeling for Non-Native Speech Recognition", In Signals and Communication technology, pp. 5-17, 2011. Berlin: Springer –Verlag.

[9] G. Hemakumar & P. Punitha, "Speech Recognition Technology: A Survey on Indian languages", *International Journal of Information Science and Intelligent System*, Vol. 2, Issue no. 4, pp. 1-38, 2013.

[10] X. D. Huang, X., "The SPHINX-II Speech Recognition System: An overview", *Computer Speech and Language*, Vol. 7, Issue 2, pp. 137–148, 1989.

[11] S. Joshi & P. Rao, "Acoustic models for pronunciation assessment of vowels of Indian English", *Conference on Asian Spoken Language Research and Evaluation*, pp. 1-6, 2013

[12] T. Choudhary, A. Kumar, & M. Dua, "Continuous Hindi Speech Recognition using Monophone based Acoustic Modeling", *International Journal of Computer Applications. In International Conference on Advances in Computer Engineering & Applications (ICACEA),* pp.163-167, 2014.

[13] P. K. Kurzekar, R. R. Deshmukh, V. B. Waghmare & P. P. Shrishrimal, "Continuous Speech Recognition System: A Review", *Asian Journal of Computer Science and Information Technology*, Vol. 4, Issue 6, pp 62-66, 2014.

[14] K. F. Lee, "Automatic Speech Recognition the development of the SPHINX system", Boston: *Kluwar*, 1989.

[15] M. Limkara, V. Sagvekar & R. Rao, "Isolated Digit Recognition Using MFCC AND DTW", International *Journal on Advanced Electrical and Electronics Engineering*, Vol. 1, Issue 1, pp. 59-64, 2012.

[16] Y. Ma, M. P. Paulraj, S. Yaacob, A. B. Shahriman & S. K. Nataraj, "Speaker accent recognition through statistical descriptors of mel-bands spectral energy and neural network model", *IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, Kuala Lumpur, Malaysia, pp. 262-267, 2012.

[17] S. B. Magre, R. R. Deshmukh & P. V. Janse, "A Review on Feature Extraction and Noise Reduction Technique", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, Issue 2, pp. 352-356, 2014.

[18] A. N. Mishra, A. Biswas, & M. Chandra, "Isolated Hindi Digits Recognition: A Comparative Study", *International Journal of Electronics and Communication Engineering*, Vol. 3, Issue 1, pp. 229-238, 2010.

[19] Pawar, G. S., & Morade, S. S., "Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit", *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 4, Issue 6, pp. 781-784, 2014.

[20] Phull, D. K., & Kumar, G. B., "Investigation of Indian English Speech Recognition using CMU Sphinx", *International Journal of Applied Engineering Research*, Vol. 11, Issue 6, pp. 4167-4174, 2016.

[21] Saini, P., & Kaur, P., "Automatic Speech Recognition: A Review", *International journal of Engineering Trends and Technology,* Vol. 4, Issue 2, pp. 132-136, 2013.

[22] Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R., "A Review on Different Approaches for Speech Recognition System", *International Journal of Computer Applications,* Vol. 115, Issue 22, pp. 23- 28, 2015.

[23] Sarada, G. L., Lakshmi, A., Murthy, H. A., & Nagarajan, T., "Automatic transcription of continuous speech into syllable-like units for Indian languages", Vol. 34, Issue 2, pp. 221-233, 2009.

[24] Satori, H., & ElHaoussi, F., "Investigation Amazigh speech recognition using CMU tools", *International Journal of Speech Technol*, Vol. 17, pp. 235–243, 2014.

[25] Toma, T. T., Md Rubaiyat, A. H., & Asadul Huq, A. H. M., "Recognition of English vowels in isolated speech using characteristics of Bengali accent", *International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 405- 410, 2013.

[26] Walha, R., Drira, F., EI-Abed, H., & Alimi, A. M., "On Developing an Automatic Speech Recognition System for Standard Arabic Language", *International Journal of Electrical, Computer, Energetic, Electronics and Communication Engineering*, Vol. 6, Issue 10, pp. 1138-1143, 2012.

[27] World-English, Retrieved: February 10, 2017, from http://www.world-english.org/english500.html, 2017

[28] Young, S., Evermann, G., & Odell, J., "The HTK Book", (for HTK Version 3.4), USA: Cambridge University, 2009.

[29] Zue, V., Cole, R., & Ward, W., "Survey of the state of the art in human language Technology", USA: Cambridge University Press and Giardini, 1997.