# On Developing an Automatic Speech Recognition System for Commonly used English Words in Indian English

Ms. Jasleen Kaur (Research Scholar)
Department of Computer Science and Engineering
Maharaja Ranjit Singh Punjab Technical University
Bathinda, India
*jasleenkaurbhogal@gmail.com*

Prof. Puneet Mittal (Assitant Professor)
Department of Computer Science and Engineering
Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib, India
*puneet.mittal@bbsbec.ac.in*

**Abstract—**Speech is one of the easiest and the fastest way to communicate. Recognition of speech by computer for various languages is a challenging task. The accuracy of Automatic speech recognition system (ASR) remains one of the key challenges, even after years of research. Accuracy varies due to speaker and language variability, vocabulary size and noise. Also, due to the design of speech recognition that is based on issues like- speech database, feature extraction techniques and performance evaluation. This paper aims to describe the development of a speaker-independent isolated automatic speech recognition system for Indian English language. The acoustic model is build using Carnegie Mellon University (CMU) Sphinx tools. The corpus used is based on Most Commonly used English words in everyday life. Speech database includes the recordings of 76 Punjabi Speakers (north-west Indian English accent). After testing, the system obtained an accuracy of 85.20 %, when trained using 128 GMMs (Gaussian Mixture Models).

***Keywords***:- *Automatic Speech recognition, Indian English, CMU Sphinx, Acoustic model*

## I. INTRODUCTION

### A. Automatic Speech Recognition

Automatic speech recognition is a process in which an acoustic speech signal is transformed into the text by the computer [1]. Automatic Speech Recognition means a system that takes human speech as an input and tries to generate a corresponding set of words using a specific algorithm [2]. In Automatic Speech Recognition, a computer initially processes a speech from recorded audio signal and then changes it to the corresponding text [3]. It is the process by which a computer respond to what a person said rather than who said. In Speech recognition, an acoustic signal recorded by a microphone or a telephone is converted to a set of words by computer [4]. Speech Recognition is a process of making a computer able to identify and respond to the sound produced during human speech. Recognition means getting a computer to react appropriately to spoken language [5]. A speech recognition system consists of a microphone, speech recognition software for a computer to take and interpret the speech, a good quality soundcard for input/output. [1].

### B. Types of Speech Recognition System

ASR system can be categorized into different classes based on various parameters. The various categories [6] are described below:

#### 1) Based on utterances

*a)* Isolated words

In Isolated word recognition, system recognizes single word. Speaker needs to give only one word command at a time. It is simple and easy to implement because word boundaries are easily detected and the words are pronounced clearly [7].

*b)* Continuous speech

This system allows the speaker to speak almost naturally, while the computer determines its content. Basically, it is computer dictation. Closest words run together without any pause or other division between words. They are more complex.

*c)* Spontaneous speech

This system recognizes the natural speech that comes suddenly through mouth. This type of ASR system is able to handle a variety of natural speech features i.e. words being run together may include mispronunciation, silence and false starts.

#### 2) Based on Speaker Model

Each speaker has a unique voice. Speech recognition system is categorized as follow:

*a)* Speaker-dependent Models:

Speaker-dependent systems are developed for a particular type of speaker. They are more accurate for the particular speaker, but are less accurate for other speakers. These systems are cheaper, easier to develop. But they are not as flexible as speaker independent systems.

*b)* Speaker-independent Model

In speaker-independent models, a system can recognize a variety of speakers without any prior training. The drawbacks of this type of models are: limits on the number of words in a vocabulary; hard implementation; expensive; and accuracy is lower than speaker-dependent models.

**3) Based on Vocabulary**

The size of vocabulary affects the complexity, processing and rate of recognition (accuracy) of ASR system. ASR systems are classified as follow:

- Small Vocabulary: 1–100 words or sentences
- Medium Vocabulary: 101–1000 words or sentences
- Large Vocabulary: 1001–10,000 words or sentences
- Very large vocabulary: Above 10,000 words or sentences

### C. Block Diagram of ASR System

The common components of the ASR system found in most of the applications or area are shown in Figure 1.

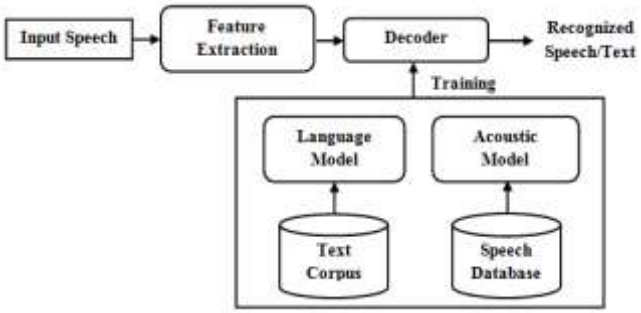


**Figure 1:** Block diagram of ASR System

**1) Input Speech:** In this part, the speech is recorded from the different speakers through a microphone. This speech signal is in analog form, so it cannot be directly transferred to the ASR system. For further processing, the speech signal is first transformed into the digital signal.

**2) Feature extraction:** It is a technique that helps us to find the set of parameters for the utterances that have acoustic relation with speech signals. The feature extractor keeps only relevant information and reject irrelevant one. Various methods used for feature extraction are like– Mel-Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), Linear Predictive Cepstral Coefficient (LPCC) and RASTA-PLP (Relative Spectral Transform) etc.

**3) Acoustic model:** It is the main part of Training. The acoustic model provides a mapping of the acoustic information and phonetics. It uses speech signals from training database. Various models are available for acoustic modeling. Hidden Markov Model (HMM) is widely used and accepted model [8] for training and recognition.

**4) Language model:** It is also the part of training. The 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 [1]. Speech recognition system uses bi-gram, tri-gram, n-gram language models. The language model differentiates between word and phrase that has similar sound.

**5) Decoder:** A decoder simply compares the unknown test pattern with each sound class reference pattern and finds similarity between them. It is also called the testing phase of the system. In testing phase patterns classification is done to recognize the speech. Finally, the output of this sub-system is text.

The paper is organized in the sections as: Sect. 2 provides a description of the Indian English language. In Sect. 3 problem being formulated and objectives of the proposed system are shown. In Sect. 4, we describe the Indian English speech recognition system and our investigations to adapt the system to Indian English language. Sect. 5 presents the experimental results. Finally in Sect. 6, conclusions and future scope is provided.

## II. INDIAN ENGLISH LANGUAGE

Indian English is any of the forms of English characteristic of the Indian_subcontinent. Idiomatic forms that are derived from Indian languages and vernaculars are absorbed into Indian English. But, the Indian English dialect remains homogeneous in vocabulary, phonetics and phraseology between variants.

**Phonology:** Indian accents for English language vary from state to state (from the southern part of India to the northern part). Some Indians speak English in an accent closer to British English while other Indians uses vernacular, native-tinted accent for English speech.

**Vowels:** In north-eastern part of India, many speakers do not make a clear distinction between /ɒ/ and /ɔː/ (eg. cot and caught). Most of north-eastern Indian English speaker pronounce /æ/ as /a/ (eg. man) and /oo/ is pronounced as /u/ (eg. pool) etc. Similarly, in South India /ɒ/ is pronounced as /ɑː/ (i.e. coffee is pronounced as kaafi, copy will be kaapi, lock will be laak etc.).

**Consonants:** Pronunciation of consonants varies between rhotic and non-rhotic. Speakers from north-eastern part of India differentiate between /v/ and /w/ where a group of speakers pronounce it as /ʋ/ for both the consonants. The voiceless plosive /p/, /t/, /k/ are non-aspirated in Indian English language. But they are aspirated in word initial or stressed syllables in other English accents (American, British). So 'Pin' is pronounced as [Pɪn] in Indian English but [Pʰɪn] in other dialects. The north Indian English accent lacks the dental fricatives (/θ/ and /ð/; spelled with th). The voiceless dental plosive [[t̪ʰ] is substituted for /θ/ and possibly aspirated version [d̪ʱ] is substituted for /ð/ (eg. "thin" would be realized as [t̪ʰɪn] instead of /θɪn/). Indian languages lack the voice alveolar fricative /z/. The unvoiced /s/, often use the postalveolar /dʒ/. English alveolar /t/ would be articulated as the Indian retroflex /ʈ/ or as the dental /t̪/ in different phonological environments. Indian English have a reduced vowel system; /r/ tends to become a flap or retroflex flap; the consonants /p/, /t/, /k/ tends to be unaspirated; in some regions, /v/ and /w/ are not distinguished, also in others regions, /p/ and /f/, /t/ and /θ/, /d/ and /ð/, and /s/ and /š/ are not. A Tamil native who uses Tamil language to communicate would speak English with a shadow of Tamil accent over it. Similarly, for all south Indian languages like Malayalam, Kannada, Telugu, Konkani and all north Indian languages like Hindi, Punjabi, Bengali, Bhojpuri, etc.

## III. RELATED WORK

Ganesh S. Pawar and Sunil S. Morade [9] developed a speaker-dependent isolated English digits recognition system

with a database of 50 speakers. HMM is used as a classifier and MFCC as a features extraction algorithm. HTK is used for training and testing. The system achieved 95% of accuracy. Hassan Satori and Fatima ElHaoussi [10] worked on the development of a speaker-independent continuous Amazigh speech recognition system. The system is based on the CMU Sphinx tools. In the training and testing phase, Amazigh_Alphadigits corpus was used. This corpus consists of speech and their transcription of 60 Berber Moroccan speakers (30 males; 30 females). The system achieved an accuracy of 92.89% when trained using 16 GMMs. Joshi et al. [11] worked on pronunciation assessment of vowels of Indian English uttered by speakers with Gujarati using confidence measures obtained by automatic speech recognition. It is observed that Indian English speech is better represented by Hindi speech models rather than by American English models for vowels common to the two languages. Maruti Limkar [12] proposed a speech recognition system for isolated English digit using MFCC and DTW (Dynamic time wrapping) algorithm which gave an accuracy rate of 90.5%. They provide a comparative study on the speaker-dependent and speaker-independent speech recognition by the designed digit recognition system. HTK is used for speech recognition compared to MATLAB. HTK is open-source while MATLAB is a commercial product. MFCC is used for features extraction as it is one of the effective features extraction algorithms. HMM is used as classifier rather DTW algorithm because it is easier and brings more accuracy in recognition. The accuracy is compared for both the speaker-dependent and speaker-independent. Mishra et al. [13] worked on speaker independent connected digits with R-PLP, BFCC and MFCC to describe the robust features for digit recognizing both in noisy and clean environment based upon Hidden Markov Model and using Hidden Markov Tool Kit for MFCC. All other features are extracted using MATLAB and saved in HTK format. The result shows features extraction with MFCC having an accuracy rate of 98% for clean data which is 1% lesser than MF-PLP. Disha Kaur Phull and G. Bharadwaja Kumar [14] 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 average 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. Lakshmi Sarada et al. [15] 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. [16] 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.

## IV. PROBLEM FORMULATION

ASR systems that have been created so far are working on the web. Online systems enable you to work from any place, at anytime. But, they require continuous and reliable web connection. Offline systems can work even they are separated from the internet. They are quick and responsive. Also in disconnected systems, there is no such system built up that can recognize the commonly used English words in north-west Indian English accent.

The systems that have been developed so far just uses other north Indian dialects like - Hindi, Punjabi, Bengali, Bhojpuri, and so forth yet not English dialect. So it is proposed to build up a system that uses Indian English dialect for recognizing commonly used English words [17] in an accent used by native of Punjab (north-west region). So it is proposed to build up an Indian English (IE) acoustic model for preparing the ASR system. Additionally the phonetic dictionary used as a part of the proposed system is centered on north-west Indian English accent.

**Objectives:**

The important objectives of the proposed work are:

- To study various Automatic Speech Recognition (ASR) systems.
- To develop an Acoustic model for commonly used English words in north-west Indian English accent.
- To develop the language model for Commonly used English words.
- To compute the accuracy of an Acoustic model being developed in north-west Indian English accent.
- To develop a system that can recognize the English words in English accent used by Punjabi people.

## V. INDIAN ENGLISH SPEECH RECOGNITION SYSTEM

### A. System overview

In this speech recognition system, initially data preparation is done in which speech recordings of 500 commonly used English words are collected from each of the 76 Punjabi speakers. The text corpus consists of grammar for 500 English words. Then the phonetic dictionary is prepared using phonetic transcriptions or commonly used English words. Then, an acoustic model and language model are developed. Both training and recognition are based on CMU Sphinx system. It is HMM-based, speaker-independent automatic speech recognition system enable of handling large vocabularies (CMU Sphinx Open Source Speech Recognition Engines) [18][19].

### B. Speech Database Preparation

The database for “Most Commonly used English words” is used in this work and it contains a corpus of speech and their transcription. The corpus contains 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

| | |
|---|---|
| **Speaking mode** | Isolated words |
| **Sampling rate** | 16 kHz |
| **Enrolment (or Training)** | Speaker-independent |
| **Vocabulary size** | Medium |
| **Equipment** | Microphones and a Smart Voice Recorder application. |
| **Number of channels** | 1, Mono |
| **Audio data file format** | .wav |
| **Corpus** | 500 words |
| **Number of speakers** | 76 |
| **Accent** | North-West Indian English |
| **Size of training set** | 80% of the total speech corpus |
| **Rule set** | 20% 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. 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.

### C. Pronunciation dictionary

The pronunciation dictionary is also called as lexicon. It contain all the 500 most commonly used English words followed by their pronunciation (phonetic transcription) based on north-west Indian English (IE) accent. This dictionary has been created after a deep study of Indian English phonetics and different rules are used to pronounce the words. Table 2, shows the phonetic dictionary list for some words used to train the system. The pronunciation dictionary acts as an intermediary between the Acoustic Model and Language Model.

### D. Feature extraction

This sub-system performs the extraction of speech features form the recorded audio files. Feature extraction plays an important role in the performance of speech recognition system. As seen in Table 1, the parameters used in this system, were 16 KHz sampling rate with a 16 Kbit sample.

### E. Training

Training is the process of learning the Acoustic Model and Language Model along with the pronunciation dictionary so as to build the knowledge base used by the recognition system. The Training of acoustic model is performed using CMU Sphinx tools that uses embedded training method based on the Baum-Welch algorithm [20].

**Table 2:** The phonetic dictionary list used in the training

| | | | |
|---|---|---|---|
| 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 | So on, Y as last alphabet | |
| COLOR | K L ER | | |
| DIFFER | D IH FF ER | | |

#### 1) Acoustic model

In acoustic model, the observed features of phonemes (basic speech units) are mapped to HMMs. The words in the vocabulary are modeled as a sequence of phonemes, and each phoneme is modeled as a sequence of HMM states. The basic HMM model used is 3-states HMMs architecture for each English phoneme. It includes three states: begin state, middle and end state. States join models of HMM units together in the ASR engine, as shown in Figure 2.

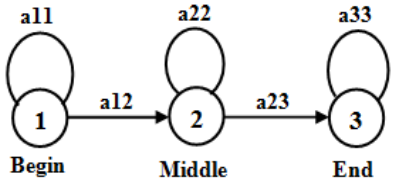


**Figure 2** The 3-states HMM model.

The acoustic modeling is done by using speech signal from the training database. 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 [21]. At this stage, the speech engine uses the phonetic dictionary (see Table 2) 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.

#### 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, [21]

the Cambridge statistical language modeling toolkit (CMU-CLMTK) is used to generate Language model of this system.

### F. Testing

Testing is also called as decoding. It is performed after the completion of training phase. It is very important to test the quality of the trained database, so as to select only best parameters, to know how the system performs and to optimize the performance of the system. For this, a testing (decoding) step is needed. 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).

### 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. Finally a recognized text is generated by the ASR system.

## VI. EXPERIMENTAL RESULTS

### A. Evaluation Criteria

The performance of the proposed work can be evaluated by the Recognition percentage 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

When the decoding job is complete, the script evaluates Word Error Rate (WER) and Sentence Error Rate (SER). The lower these rates the better the recognition will be. For typical 10-hours task WER should be around 10%. For a large task, it could be like 30%. In this paper, the 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 85.20 %, when trained using 128 GMMs (Gaussian Mixture Models). Table 3 shows the results of the experiments.

**Table 3** System overall recognition rate for model (ENG)

| Model | GMMs | SER (%) | WER (%) | Recognition (or accuracy) |
|---|---|---|---|---|
| ENG | 128 | 13.80 | 14.80 | 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 Commonly used English words spoken in north-west Indian English accent. This work includes 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 the speech database for Commonly used 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 85.20%.

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 categorizing the speakers on the gender or age basis. So that the word error rate (WER) can be reduced to 10% for improving the system accuracy.

## VII. ACKNOWLEDGMENTS

We would like to thank Head of department for providing lab facilities and guidance. Also thank to the people involved in the development of the CMU Sphinx tools and making it available as an open source. Also, thank to all the people involved in this work for its completion.

## REFERENCES

[1] Preeti Saini and Parneet kaur, “Automatic Speech Recognition: A Review”, International journal of Engineering Trends and Technology, Vol. 4, Issue 2, 2013.

[2] Al-Zabibi, M., “An acoustic-phonetic approach in automatic Arabic Speech Recognition”, The British Library in Association with UMI, 1990.

[3] Gruhn, R.E.; Minker, W.; Nakamura, S., “Statistical Pronunciation Modeling for Non-Native Speech Recognition”, Vol. 10, 114 p., 2011.

[4] Zue, V.; Cole, R.; Ward, W., “Survey of the state of the art in human language Technology”, Survey on Speech recognition, USA, 1996.

[5] Hemakumar and Punitha, “Speech Recognition Technology: A Survey on Indian languages”, International Journal of Information Science and Intelligent System, Vol. 2, No. 4, 2013.

[6] 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, No. 22, April 2015.

[7] Bhabad S.S.; Kharate G.K., "An Overview of Technical Progress in Speech Recognition" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 3, March 2013.

[8] Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, Mahua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing, 2012, 3, 394-401.

[9] Ganesh, S. Pawar ; Sunil, S. Morade, "Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit", International Journal of Advanced Research in Computer Science and Software Engineering, Jaunpur, Uttar Pradesh, India, Vol. 4, Issue 6, 2014.

[10] Hassan Satori and Fatima El Haoussi, "Investigation Amazigh speech recognition using CMU tools", International Journal of Speech Technol, vol. 17, pp. 235–243, 2014.

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

[12] Maruti Limkara, "Isolated Digit Recognition Using MFCC AND DTW", International Journal on Advanced Electrical and Electronics Engineering, India, Vol. 1, Issue 1, 2012.

[13] Mishra A. N., "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 4, Issue 2, 2011.

[14] Disha Kaur Phull and G. Bharadwaja Kumar, "Investigation of Indian English Speech Recognition using CMU Sphinx", International Journal of Applied Engineering Research ISSN 0973-4562, Vol. 11, Issue 6, pp. 4167-4174, 2016.

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

[16] 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.

[17] World-English, Retrieved Feb 2, 2017., online available at: http://www.world-english.org/english500.html

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

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

[20] S. Young et al., "The HTK Book (for HTK Version 3.4)", Cambridge University, March 2009.

[21] Lmtool-new, Retrieved date: Feb 23, 2017., online available at from http://www.speech.cs.cmu.edu/tools/lmtool-new.html.