Open access Journal International Journal of Emerging Trends in Science and Technology
IC Value: 76.89 (Index Copernicus) Impact Factor: 4.219 **DOI:** https://dx.doi.org/10.18535/ijetst/v4i7.07

An experimental framework of speaker independent speech recognition system for Kashmiri language (K-ASR) system using Sphinx

Authors
Vivek Bhardwaj[1], Virender Kadyan[1], Amitoj Singh[2], Rohit Sachdeva[3]
[1] Department of Computer Science and Engineering,
Chitkara University Institute of Engineering and Technology
Chitkara University, India
[2] Maharaja Ranjit Singh Punjab Technical University
Bathinda
[3] Multani Mal Modi College, Patiala

Abstract.
Speech to text conversion in various languages have been performed so far but no process has defined for the Kashmiri language. There has been no research done on Kashmiri speech recognition. So in this work, we describe the development as well as implementation of first CMU Sphinx-3 based speech recognizer for the Kashmiri language. Recognition of the words have been done by using hidden markov models (HMMs). Dictionary consists of 100 words, representing Kashmiri digits from one (akh) to hundred (hat). Here, we developed a speaker independent, Kashmiri - Automatic Speech Recognition (K-ASR) system. The System is trained and tested for 1200 words spoken by 12 male and female speakers. Maximum Accuracy of 78.33% was achieved by the K-ASR system.
Keywords: *Speech recognition (SR), Kashmiri language, HMM, MFCC, Sphinx*

1 Introduction

In today life, ASR is the best way to recognize the spoken words recorded by the microphone. ASR is the process of converting input speech signal into text and one of the most developing field of Natural Language Processing (NLP).
Kashmiri is one of the important language of India and is a Dardic subgroup language belonging to the Indo-Aryan languages. It is mostly spoken in the Kashmir Valley and Chenab regions of Jammu and Kashmir. There are approximately 55 million speakers in India who speaks Kashmiri language, according to 2001 census [10]. Beside Kashmir, this language is also spoken in Azad Kashmir, Pakistan. Kashmiri has distinct 15 Vowels and 27 Consonants which it does not share with different Indo-Aryan languages [10]. But there is no ASR system developed for the Kashmiri language. The proposed research will focus on the speech recognition for Kashmiri language and purpose of this work is to design and train a speech recognition system that could be used by application developers to develop applications that will take Kashmiri language speakers aboard the current information and communication. In this paper, we developed a speaker-independent, continues speech recognizer for the Kashmiri language, who does not require any training and can be used by any speaker. The system is developed by using Sphinx toolkit and called as K-ASR system.

2 Related Work

In this section literature of speech recognition works in different languages discussed. An Arabic ASR system was developed using sphinx toolkit in (Hyassat and Zitar 2006) [1]. Three Arabic corpuses HQC-1 - Holly Qura'an Corpus (18.5 hours), CAC1 - command and control corpus (1.5 hours) and ADC - Arabic digits

corpus (less than 1 hour) were created for testing and training of the system. The accuracy of 70.8 %, 98.1 % and 99.2 % was achieved for three different corpuses.

The development of an Arabic ASR engine using htk and MFCC to extract feature vectors shown in (Al-qatab and Ainon 2010) [2]. Isolated words, as well as continuous speech were recognized by the engine. Maximum of 98.01 % accuracy was achieved by the system.

An Assamese language isolated word speech recognition system was developed by (Bharali and Kalita 2015) [3]. The System was trained for ten different Assamese words (0 to 9) spoken by fifteen speakers. For clean data using MFCC feature extraction technique, the system provides maximum accuracy of 80 % and 95 % for noisy data using LPCEPSTRA.

A connected word Hindi ASR system has been developed in (Kumar et al. 2012) [4] using htk. The vocabulary of 102 words created by 12 male and female speakers. The system utilizing MFCC for feature extraction and gives word accuracy of 87 %.

Dua et al. (2012) [5] provides a Punjabi speech recognition system for isolated words using htk. Speech corpus consists of 115 different Punjabi words. Maximum accuracy of 95.63 % was achieved by the system using MFCC and HMM.

Another ASR system for Tamil has been developed in (Krishna et al. 2014) [6]. Features are extracted by using Integrated Phoneme Subspace (IPS), MFCC and HMMs are used for acoustic modeling. MFCC gives 74.6 % word accuracy whereas IPS gives word accuracy of 84 %.

3 K-ASR system architecture

The overall architecture of the K-ASR system is shown in figure 1 and divided into five parts namely feature extraction, language model, building Kashmiri dictionary, acoustic model and decoder.

3.1 Feature Extraction

The first step of ASR system is to extract acoustic feature vectors corresponding to each input waveform. These acoustic feature vectors are computed with the help of executable program provided within sphinx-3 training package. Feature extraction process starts with Pre-emphasis where input speech signal is sent to high pass filter to boost the lower frequency energy to higher frequency energy. After that frame speech signal into shorter frames of 20-30 ms and then multiply each frame with hamming window [12]. After framing, if the signal is denoted by s(n) where n=0,…, N-1 then s(n)*w(n) is the signal after the windowing process, where w(n) is hamming window given by equation 1 [3,10].

$$w(n) = 0.54 - 0.46 \cos\left(2\pi n/N-1\right) \quad 0 \le n \le N-1$$

As windowing process is completed fast Fourier transform (FFT) is applied to extract spectral information for the windowed signal. In the next step mel frequency spectrum is computed by applying bandpass filter to the output of FFT. Then compute the logarithm of all filtered signals. In the last discrete cosine transform (DCT) is applied to the log filtered signals to obtain MFCC.

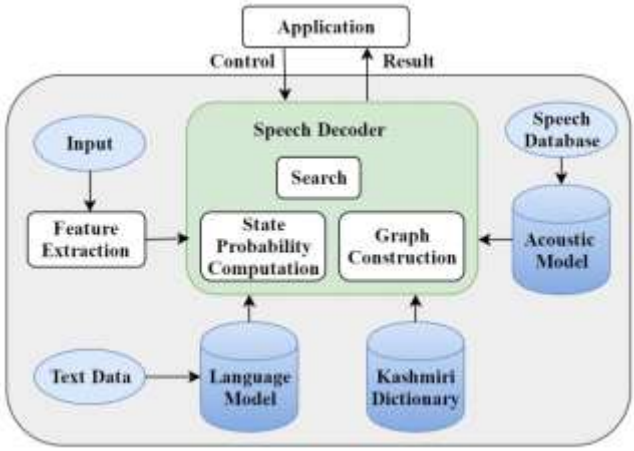

Figure 1: Architecture of K-ASR system

3.2 Transcription file

Transcription is the representation of any language in written form or mapping of spoken words onto written words [1]. In speech recognition contents of each waveform is represented in the text form. Figure 2 shows the

contents of transcription file for the Kashmiri language. Transcription of the waveform is generated by writing corresponding words and after that name of the waveform. The whole process is done manually.

<s> AKH ZU TRE TSOR PAANS SHE SATH </s> (a1)
<s> AETH NAV DAH KAH BAW TROVHA CUDHA </s> (a2)
<s> PANDHA SHURA SADHA ARDHA KANVO WOO AKWOO </s> (a3)
<s> ZUTOWOO TREWOO SHOEWOO PENCHEWOO SHATWOO SATOWOO EATAWOO </s> (a4)
<s> KUNTRO TREE AKTREE DUATREE TREETREE CHUTREE PANTREE </s> (a5)
<s> SHATREE SATTREE AKTREE KUNTEJI CHADGEE AKTEJI DWATEJI TEETEJI </s> (a6)
<s> CHURTEJI PANTEJI SHAITEJI SATTEJI ARTEJI KUNWANZAH PANCHAH </s> (a7)
<s> AKWANZAH DUANWANZA TREWANZA CHUWANZA PANWANZA SHEWANZA SATHWANZA </s> (a8)
<s> AARWANZA KANWANZA SHEETH AKHHEART DOOHEART TREHEART CHOHEART </s> (a9)
<s> PANCHHEART SHEHEART SATHEART ARHEART KUNHEART SATHHAT AAKSAATHE </s> (a10)
<s> DOSAATHE TREESAATHE CHOSAATHE PANSAATHE SHAHSAATHE SATSAATHE ARSAATHE </s> (a11)
<s> KUNSHEET SHEATH AAKSHEATH DUASHEATH TREESHEATH CHOSHEATH PANCHSHEATH </s> (a12)
<s> SHAHSHEATH SATSHEATH ARSHEATH KUNSHEATH NAMHAT AKNAMHAT </s> (a13)
<s> DONAMHAT TREENAMHAT CHONAMHAT PANCHNAMHAT SHENAMHAT </s> (a14)
<s> SATNAMHAT AKNAMHAT NANNAMHAT HAT </s> (a15)

Figure 2: Contents of transcription file

3.3 Language Model

A Language Model (LM) [9] is used to provide probabilities for a word sequence in order to predict the word which is spoken next. LM does not include silence or filler words. An n-gram LM is an adjacent sequence of n words from a given sequence of speech or text. In n-gram LM, for a given sequence of m words W= (W1, …, Wm) probability of LM is estimated by equation 2.

$$P(W1,\ldots,Wm) = \prod_{i=1}^{m} P(Wi \mid W1,\ldots,Wi-1) \approx \prod_{i=1}^{m} P(Wi \mid Wi-(n-1),\ldots,Wi-1)$$

An n-gram LM, with n=1 called as unigram, n=2 is a bigram and n=3 is a trigram LM. The LM used by our proposed K-ASR system is trigram LM in ARPA format. A trigram LM mainly consists of unigrams, bigrams and trigrams. In trigram, LM probability is computed by P (W3 | W1, W2) or we can say that W3 follows the sequence W1 W2 of words.

3.4 Building Kashmiri and filler dictionary

Dictionary file contains the words pronunciations and breaks words into the subwords present in the acoustic model. Any spoken word that is not present in the dictionary doesn't recognize and is called out of vocabulary (OOV). There is a possibility of more than one pronunciation for a single word. In that case, they are differentiated by a unique parenthesis. For example:

PANDHA P AE N D HH AH PANDHA (2) P AE N D AH CMU provide lexicon tool to generate your own dictionary and also provides the implementation of the dictionary interface to support the CMU dictionary [7]. Our dictionary consists of 100 Kashmiri digits, one (akh) to hundred (hat) generated by CMU lexicon tool.

Decoder present in sphinx also needs another dictionary called filler dictionary which consists of words that are not present in the LM. These words are called as non-speech sounds. Filler dictionary must always contain these three entries <sil>, <s>, and </s> [1]. It may also contain non speech-sounds such as um, uh and breath noise sounds made during the speech.

3.5 Phone list

Phone list is the collection of acoustic units used during the training of the models. Sphinx phone list does not allow other acoustic units than those used in the dictionary. Our phone list consists of 31 phones including silence (sil) shown in figure 3.

SIL	AA	AE	AH
AO	B	CH	D
DH	EH	EY	HH
IY	JH	K	M
N	NG	OW	P
R	S	SH	T
TH	UH	UW	V
W	Y	Z	

(2)

Figure 3: Phone list

3.6 Acoustic model

Acoustic model refers to the process of mapping between speech and statistical representations (HMMs) [11] that are generated against different features extracted using MFCC [9]. Acoustic model uses Baum-Welch or forward-backward training algorithm to create HMMs for each phone. For each word, K-ASR system generates a sequence of tri-phone HMMs using the Kashmiri dictionary. Then find the best state through the triphone HMM, for the equivalent feature vectors [8]. Figure 4 shows a HMM with

three states, using senones (0, 1, 2) and a word of fourteen frames of feature vectors. All the feature frames are assigned with a senone ID [5].

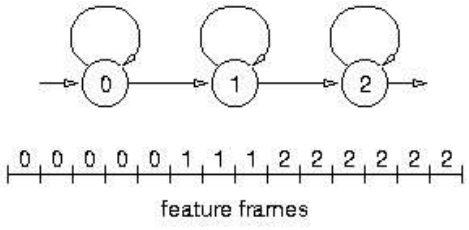

Figure 4: 3 state HMM [8]

3.7 Decoder

The sphinx-3 decoder uses dynamic programming algorithm called Viterbi algorithm. Viterbi algorithm is used to find the most probable order of hidden states through a probabilistically scored time or state lattice, called Viterbi path [8]. We need Audio files, Acoustic Models, Dictionary files, Language model as input for the decoding process and produces hypothesis file (best recognition result for each word), word file (word graph of all possible words recognized) and log file as output. Figure 5 shows the input files and output files generated by the decoder.

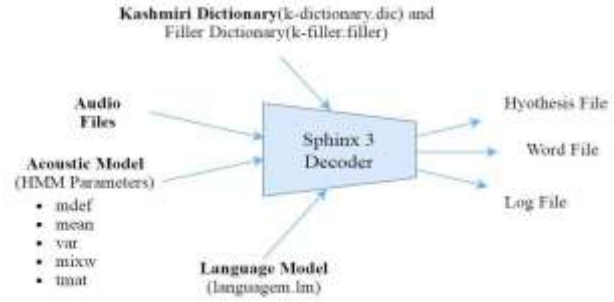

Figure 5: Sphinx-3 Decoder

4 Results

In this section, the experimental results of the K-ASR system are presented. For analyzing the results of the ASR system following metrics are used.

Percent correct (PC) =100× (WC/TW)

Word accuracy (WA) =100 × ((TW – S- D – I)/TW) (4)

Word error rate (WER) =(S+D+I)/TW (5) Where, WC = Words correct, TW = Total Words, S = Substitutions, I = insertions and D = deletions.

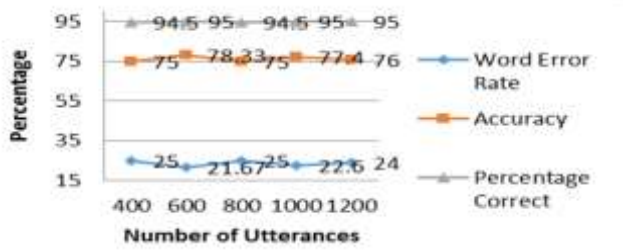

Figure 6: Recognition rate of the system

Results of the K-ASR system using Sphinx decoder has been evaluated using five experiments shown in table 1 and graphically represented in figure 6. Training and testing of the system is done using five Kashmiri digits corpuses during the experiments. Corpuses consist of 400, 600, 800, 1000 and 1200 words. Five Kashmiri speakers who did not involve in the training phase were also involved in the testing of the K-ASR system. The recognition rate of the system lies between 75 to 78.33 %. Maximum accuracy of 78.33 % obtained when the system was tested for the corpus of 600 words.

Table 1: Test Results using Sphinx decoder

Experiment Number	Data set	Result
Experiment 1	1. Untrained speakers 2. Total Speakers = 4 (2 M and 2 F) 3. Noise - free environment	Total Words = 400 Correct =378 Errors = 100 Percentage Correct = 94.5 % Accuracy = 75.00 % Error = 25.00 %
Experiment 2	1. Untrained speakers 2. Total Speakers = 5 (3 M and 2 F) 3. Noise - free environment	Total Words = 600 Correct = 570 Errors = 130 Percentage Correct = 95.0 % Accuracy = 78.33 % Error = 21.67 %
Experiment 3	1. Untrained speakers 2. Total Speakers = 5 (2 M and 3 F) 3. Noise - free environment	Total Words = 800 Correct = 756 Errors = 200 Percentage Correct = 94.5 %

		Accuracy = 75.00 % Error = 25.00 %
Experiment 4	1. Trained and untrained speakers 2. Total Speakers = 12 (7 M and 5 F) 3. Noise - free environment	Total Words = 1000 Correct = 950 Errors = 226 Percentage Correct = 95.0 % Accuracy = 77.40 % Error = 22.60 %
Experiment 5	1. Trained and untrained speakers 2. Total Speakers = 12 (8 M and 4 F) 3. Noise - free environment	Total Words = 1200 Correct = 1140 Errors = 288 Percentage Correct = 95.0 % Accuracy = 76.00 % Error = 24.00 %

4. Conclusion and Future Work

In this work, the first speaker independent ASR system for the Kashmiri language has been developed using Sphinx-3 toolkit. HMMs are used for the building of the acoustic model. The system has been implemented on one to hundred digits spoken by 12 male and female Kashmiri speakers. Experimental results show the maximum accuracy of 78.33 % by using MFCC as a feature extraction technique. The developed framework can be used by application developers to build up applications for the recognition of Kashmiri words. Moreover, the research work done in this paper could form the basis for further research in Kashmiri ASR systems.

The work can be extended further by increasing size of the vocabulary and recording the speech corpus in different noisy environments to improve the system efficiency.

References

1. Hyassat, H., & Zitar, R. A. (2006). Arabic speech recognition using SPHINX engine. International Journal of Speech Technology, 9(3-4), 133-150.
2. Al-Qatab, B. A., & Ainon, R. N. (2010). Arabic speech recognition using hidden Markov model toolkit (HTK). In Information Technology (ITSim), 2010 International Symposium in (Vol. 2, pp. 557-562). IEEE.
3. Bharali, S. S., & Kalita, S. K. (2015). A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language. International Journal of Speech Technology, 18(4), 673–684.
4. Kumar, K., Aggarwal, R. K., & Jain, A. (2012). A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering, 1(1), 25-32.
5. Dua, M., Aggarwal, R. K., Kadyan, V., & Dua, S. (2012). Punjabi automatic speech recognition using HTK. IJCSI International Journal of Computer Science Issues, 9(4), 16940814.
6. Krishna, K. M., Lakshmi, M. V., & Laksmi, S. S. (2014). Feature extraction and dimensionality reduction using IPS for isolated tamil words speech recognizer. International Journal of Advanced Research in Computer and Communication Engineering, 3(3).
7. The CMU pronouncing dictionary. Retrieved February 19, 2017, from http://www.speech.cs.cmu.edu/cgi-bin/cmudict
8. Sphinx-3 s3.X Decoder (X=6). Retrieved February 19, 2017, from http://www.cs.cmu.edu/~archan/s_info/Sphinx3/doc/s3_description
9. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.
10. Kashmiri language. Retrieved February 19, 2017, from https://en.wikipedia.org/wiki/Kashmiri_language
11. Ashraf, J., Iqbal, N., Khattak, N. S., & Zaidi, A. M. (2010, March). Speaker independent Urdu speech recognition using HMM. In Informatics and Systems

(INFOS), 2010 7th International Conference on (pp. 1-5). IEEE.

12. Satori, H., & ElHaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. International Journal of Speech Technology, 17(3), 235-243.