International Journal of Computational Intelligence and Informatics, Vol. 1 : No. 4, January - March 2012

Text-independent Mono and Cross-lingual Speaker Identification with the Constraint of Limited Data

Nagaraja B G and H S Jayanna
Department of Information Science and Engineering
Siddaganga Institute of Technology
Tumkur-572103, India
E-mail: {nagarajbg, jayannahs} @gmail.com

Abstract

Speaker recognition is a biometric process of automatically recognizing speaker who is speaking on the basis of speaker dependent features of the speech signal. Nowadays, speaker identification system plays a very important role in the field of fast growing internet based communication/transactions. In this paper, closed-set text-independent speaker identification in the context of Mono and Cross-lingual are demonstrated for Indian languages with the constraint of limited data. The languages considered for the study are English, Hindi and Kannada. Since the standard Multi-lingual database is not available, experiments are carried out on an our own created database of 30 speakers, who can speak the three different languages. Speaker identification system based on Mel-frequency cepstral coefficients–Vector Quantization (MFCC-VQ) framework is considered. It was found out in the experimental study that the Mono-lingual speaker identification gives better performance with English as a training and testing language though it is not a native language of speakers considered for the study. Further, it was observed in cross-lingual study that the use of English language either in training or testing gives better identification performance.

Keywords: Speaker identification, Mono-lingual, Cross-lingual, MFCC and VQ.

1. Introduction

Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV) systems have always been demanding in terms of robustness and accuracy for the modern state-of-the-art security applications [1]. The speaker verification involves accepting or rejecting the identity claim of a speaker. In speaker identification since there is no identity claim, the system identifies the most likely speaker of the test speech signal [2]. Speaker identification can be classified into Closed-set and Open-set identification [2]. The task of identifying a speaker who is known a priori to be a member of the set of N enrolled speakers is known as Closed-set Speaker identification system. On the other hand, Speaker identification system which is able to identify the speaker who may be from outside the set of N enrolled speakers is known as open-set Speaker identification [2]. Depending on the mode of operation, Speaker recognition can be classified as text-dependent recognition and text-independent recognition [3]. The text-dependent recognition requires the speaker to produce speech for the same text, both during training and testing whereas the text-independent recognition does not rely on a specific text being spoken [4]. Countries like India, more than fifty languages are officially recognized and citizens in India can speaks more than one language fluently. Therefore, development of Multi-lingual system is a challenging task. Multi-lingual speaker recognition and language identification are key to the development of spoken dialogue systems that can function in Multi-lingual environments [5]. In order to identify a speaker, speaker recognition system needs sufficient data. The availability of sufficient data to speaker recognition system provides sufficient information which can discriminate speaker well. As a result, the system yields good recognition performance [6]. Speaker recognition in limited data condition aims at recognizing speaker with the constraint that both training and testing data are limited. In the present work sufficient data is used to symbolize the case of having speech data of few minutes (> one minute). Alternatively, limited data symbolizes the case of having speech data of few seconds (≤ 15seconds). Since the amount of data available is small in the limited data conditions, the number of feature vectors we obtain is less which are insufficient to model and discriminate speaker well. Therefore, it is a challenging task to improve the speaker recognition in such situation. As we mentioned earlier in India people have been trained themselves to speak in many languages. This advantage can be utilized in machine learning to build a robust speaker recognition system. However, nowadays we cannot ask people to give data for a long period of time as the sufficient speaker recognition system expects. Further, due to increase in the use of communication and internet services for speech mode applications, it is desirable to

work with limited data and as well as in Multi-lingual environment. Speaker recognition under limited data conditions could be used in the following applications:

1) To locate the segment of given speaker in an audio stream such as teleconference or meetings, such data segments usually contain short utterances whose speaker needs to be identified.

2) In forensic application also the data available may be limited which may be recorded during casual conversation or by tapping the telephone channel.

3) Remote biometric person authentication for electronic transactions where speech is the most preferred biometric feature.

4) Criminals often switch over to another language, especially after committing a crime. So, training a person's voice in one language and identifying him in some other language or in a multilingual environment is a challenging task especially in the Indian context [15].

An attempt was made to recognize Multi-lingual speaker in [7]. In this work, training data of 60 seconds and for different testing data of 1, 3, 7, 10 and 15 seconds are considered for Mono and Cross-lingual experiments. Also, a Polynomial classifier of 2nd order approximation is built for Speaker Modeling. Recently, some attempts have been made to identify the speakers under limited data condition using the concept of Universal Background Model (UBM) to mitigate the sparseness, which requires additional speech data to train the Gaussian mixture model-Universal Background Model (GMM-UBM) [2]. A novel Multi-lingual text-independent based speaker identification algorithm was proposed by Geoffrey Duron in [8] and investigated 2 facets of speaker recognition: cross-language speaker identification and the same language non-native text independent Speaker identification. The results indicated that how Speaker identification performance will be affected when speakers do not use the same language during the training and testing or when the population is composed of native speakers.

In an another attempt the authors have proposed that by selecting only the feature vectors which are discriminating the speakers it is possible to identify speaker under limited data [13]. In our previous work, we made an attempt to use the concept of Multiple Frame size and Rate (MFSR) analysis technique to mitigate the sparseness of limited speaker-specific feature vectors during training and testing to improve the speaker recognition performance under limited data conditions [13]. Since the literature reveals that there are no enough studies on Multi-lingual speaker recognition system with the constraint of limited data, in this work we have made an attempt to identify speaker using Mel-Frequency Cepstral Coefficients (MFCC) as feature vectors and Vector Quantization (VQ) as modeling technique. Fig. 1 shows the overall Block diagram of Speaker identification System. The following steps show the complete speaker identification process:

a) Choose the training data.

b) Extract the features using MFCC.

c) Generate the speaker model using VQ.

d) Choose the testing data.

e) Extract the features using MFCC separately.

f) Compare test features with speaker model.

g) Use the Decision logic to find out the winner.

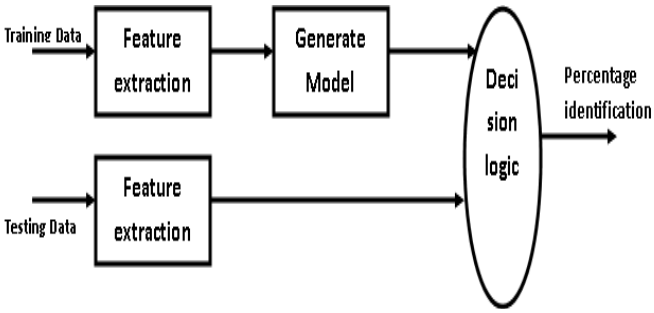

Fig. 1 Block Diagram of Speaker identification system.

The remainder of the paper is organized as follows: Section 2 describes the database used for the experiments. Feature extraction using MFCC and speaker modeling using VQ techniques are presented in Section 3. In Section 4, Mono-lingual speaker identification is presented. The Cross-lingual speaker identification is presented in Section 5. Section 6 gives Summary of the present work and scope for the future work.

2. Speech Database for the study

The speech database for the experiments was collected from 30 speakers. The database includes 17-males and 13-females speakers. All the 30 speakers were trilingual and their voice was recorded in English, Hindi and Kannada. The voice recording was done in an engineering college laboratory. The speakers were undergraduate students and faculties in an engineering college. The age of the speakers varied from 18-35 years. The speakers were asked to read the small stories in three different languages. The training and testing data were recorded in different sessions with a minimum gap of two days. The approximate training and testing data length is two minutes. Recording was done using free downloadable Wave surfer 1.8.8p3 software and beetel Head phone-250 with a frequency range 20-20 kHz. The speech files are stored in .wav format. The experiments are conducted using different sizes of training and testing data to study the effectiveness of the speaker recognition system. The detail specifications used for collecting the database are shown in Table 1.

Table. 1 Description of Database

Item	Description
Number of Speakers	**30**
Sessions	**Training and Testing**
Sampling Rate	**8kHz**
Sampling Format	**1-channel, Lin16 sample encoding**
Languages covered	**English, Hindi and Kannada**
Microphone	**beetel Head phone-250**
Recording Software	**WaveSurfer 1.8.8p3**
Maximum Duration	**120 seconds/story/language**
Minimum Duration	**Depends on Speaker**

3. Feature extraction and Modeling

The purpose of feature extraction stage is to extract the speaker-specific information in the form of feature vectors at reduced data rate [2]. In this work, features are extracted using MFCC technique. The state-of-the-art speaker identification system uses MFCC as a feature for recognizing speakers [6]. Fig.2 shows the block diagram representation of the MFCC method. Speech recordings were sampled at the rate of 8 kHz. Frame duration of 20 msec and a 10 msec for overlapping durations are considered. After framing, windowing (Hamming) method is carried out to minimize the spectral distortion. The mathematical expression for the Hamming window is as follows:

$$h(n) = 0.54 - 0.46 \cos(2\pi n / N-1), \quad (1)$$

Fourier transform is then applied on the windowed frame signal to obtain the magnitude frequency response. A magnitude spectrum (in human perception, it is more important to model the magnitude spectra of speech than their phase [14] is computed. The resulting spectrum is passed through a set of triangular band pass filters. We have considered 35 filters. These filters are equally spaced along the Mel-frequency scale. The Mel scale is a mapping between the real frequency scale (Hz) and the perceived frequency scale (Mels). The mapping from linear scale to Mel scale is given in equation 2

$$f_{mel} = 2595 \log_{10}(1+f/700), \quad (2)$$

In order to get the cepstral coefficients, Discrete cosine transform (DCT) is applied. Using DCT rather than Discrete Fourier transform (DFT) magnitude is that

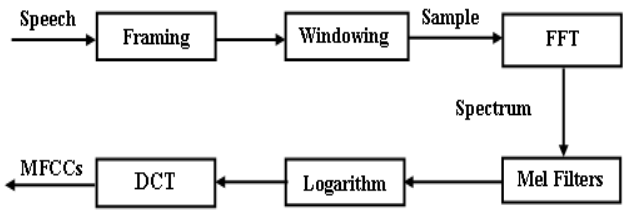

Fig. 2 Block Diagram of MFCC technique

it retains the relative phases of the feature coefficient trajectories, and hence, it can preserve both phonetic and speaker-specific information [8]. In this work, first 13 coefficients are considered as feature vectors. Since the 0th coefficient can be regarded as a collection of average energies of each frequency bands, it is unreliable [10].

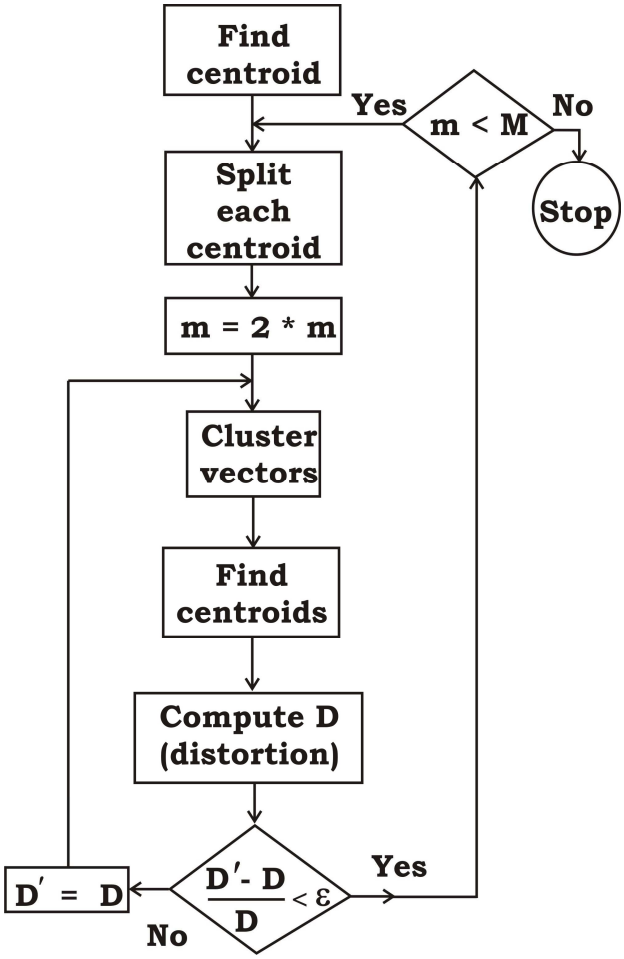

Fig.3 LBG Algorithm

The feature vectors of each speaker are further processed by a suitable modeling technique called Vector Quantization (VQ) [2]. VQ is a process of mapping vectors from large vector space to finite number of regions in the space. The vector quantization method is explained in the form of a flow chart shown in the Fig. 3. Most of the computation time in VQ-based speaker identification consists of distance computations between the unknown speaker's feature vectors and the models of the speakers enrolled in the system database [11]. In this work, the Linde-Buzo-Gray (LBG)-VQ technique is used with a splitting parameter (€) of 0.05. The initial codebook is obtained by the splitting method. In this method, an initial code vector is set as the mean of the entire training data. This code vector is then split into two and the algorithm runs with these two codebooks. Later these two codebooks are split into four codebooks and the iterative algorithm is repeated until the desired codebook size is achieved. We have generated different codebooks of sizes 16, 32, 64 and 128.

4. Mono-lingual Speaker Identification

In Mono-lingual speaker identification, training and testing languages are same for a speaker [15]. Since the data is collected in three languages to study the robustness of the system, the experiments are conducted in three cases with a speech data of 10 and 15 seconds.

1. Training and testing with English language.
2. Training and testing with Hindi language.
3. Training and testing with Kannada language.

The Mono-lingual experimental results for 10 seconds of training and testing data are shown in Fig. 4. Note: A/B indicates training with language A and testing with language B. eg. E/K indicates training with English language and testing with Kannada language. The results show that the speaker identification system yields good performance of 73.33% for codebook size of 128 when trained and tested with English language. The performance of the speaker identification system trained and tested with Hindi language is 73.33% for codebook sizes of 64 and 128. The performance of speaker identification system trained and tested with Kannada language is 70% for codebook sizes of 32, 64 and 128.

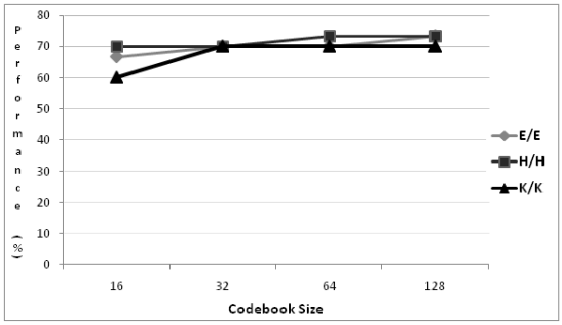

Fig. 4 Performance of Mono-lingual speaker identification system for 10 seconds of data.

The Mono-lingual experimental results for 15 seconds of training and testing data are shown in Fig. 5. The results show that the speaker identification system yields good performance of 90% for codebook size of 128 when trained and tested with English language. The highest performance with English language may be due to the speakers considered for the study. The speakers considered for the study (undergraduate students and faculties of Engineering College) are more comfortable with English language as they are studying/teaching in English medium and used to it.

The speaker identification system trained and tested with Hindi language gives the highest performance of 86.66% for codebook size of 128. This performance is better than the Kannada language. This is because almost all the speakers had taken additional time to practice the Hindi story which was given to read out in the different sessions and thus their fluency was significantly improved. The performance of the speaker identification system trained and tested with Kannada language is 73.33% for codebook size of 128. The poor performance may be due to the speakers difficulty in reading Kannada language since they had just studied this language as one of the languages subject in school days.

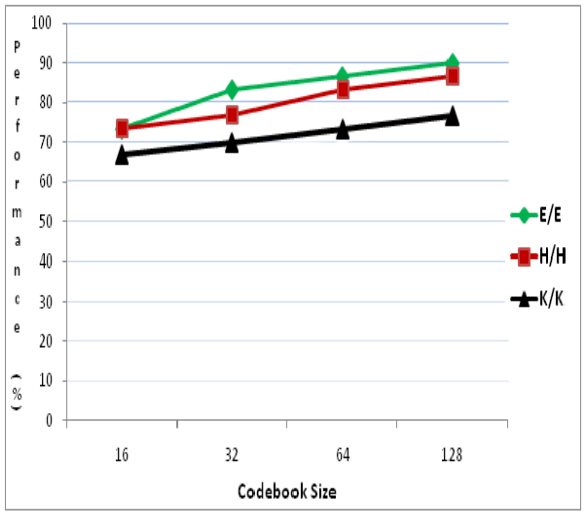

Fig. 5 Performance of Mono-lingual speaker identification system for 15 seconds of data.

5. Cross-lingual Speaker Identification

In Cross-lingual speaker identification, training is done in one language (say A) and testing is done in another language (say B) [15]. In this section, the impact of language on speaker identification system is presented. In order to demonstrate this, we have conducted six different Cross-lingual experiments with the speech data of 10 and 15 seconds.

The speaker identification system trained with Hindi and Kannada, and tested with English language for 10 seconds of training and testing data is shown in Fig. 6. The speaker identification system yields 63.33% and 66.66% for codebook size of 128 for H/E and K/E, respectively. The speaker identification system trained with Hindi and Kannada, and tested with English language for 15 seconds of training and testing data is shown in Fig. 7. The speaker identification system yields 76.66% for codebook sizes of 128 and 64 for H/E and K/E, respectively. With English as a testing language, no much difference in identification performance was observed in comparison with Hindi and Kannada as training languages. Even though regional language is Kannada, the speakers are used to this language colloquially but not in other terms. English language is used in each and every sector of everyday life so the speakers are having better reading and pronunciation of the text material for English language.

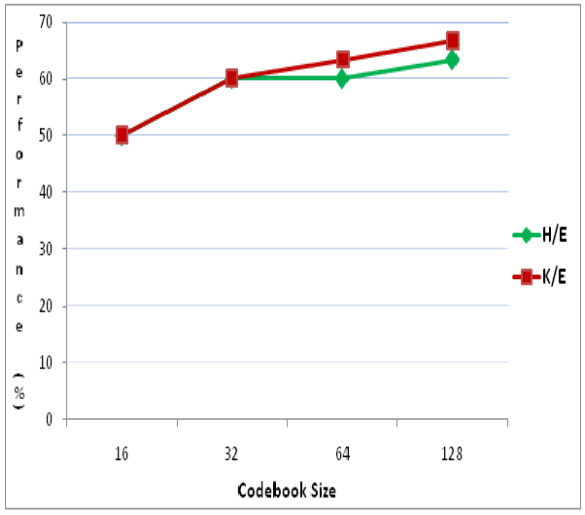

Fig.6 Performance of Cross-lingual speaker identification system for 10 seconds of data: Hindi and Kannada are the training languages and English is testing language.

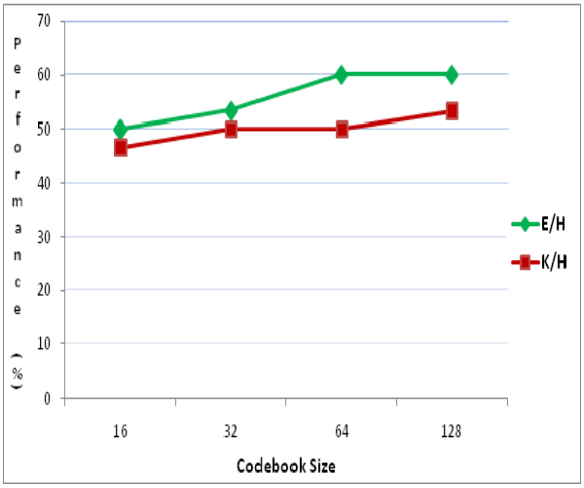

Fig. 8 Performance of Cross-lingual speaker identification system for 10 seconds of data: English and Kannada are the training languages and Hindi is testing language.

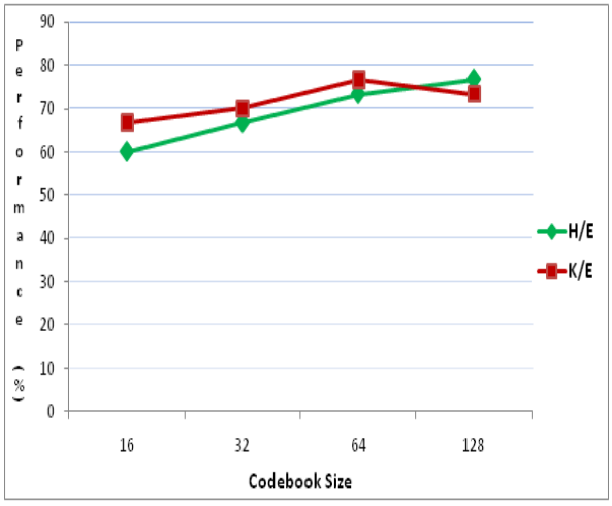

Fig.7 Performance of Cross-lingual speaker identification system for 15 seconds of data: Hindi and Kannada are the training languages and English is testing language.

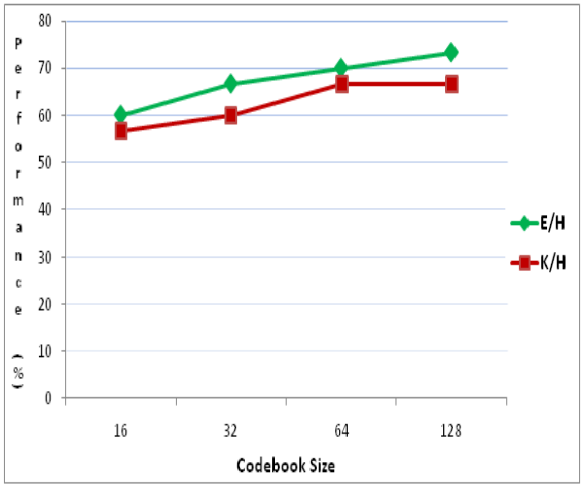

Fig. 9 Performance of Cross-lingual speaker identification system for 15 seconds of data: English and Kannada are the training languages and Hindi is testing language.

The speaker identification system trained with English and Kannada, and tested with Hindi language for 10 seconds of training and testing data is shown in Fig. 8. The speaker identification system yields 60% for codebook sizes of 64 and 128 and 53.33% for codebook size of 128 for E/H and K/H, respectively. The speaker identification system trained with English and Kannada, and tested with Hindi language for 15 seconds of training and testing data is shown in Fig. 9. The speaker identification system yields 73.33% for codebook size of 128 and 66.66% for codebook sizes of 64 and 128 for E/H and K/H, respectively. The speaker identification system trained with English and Hindi, and tested with Kannada language for 10 seconds of training and testing data is shown in Fig. 10. The speaker identification system yields 66.66% for codebook sizes of 64 and 128 and 60% for codebook size of 128 for E/K and H/K, respectively.

The speaker identification system trained with English and Hindi, and tested with Kannada language for 15 seconds of training and testing data is shown in Fig. 11. The speaker identification system yields 76.66% for codebook size of 128 and 63.33% for codebook sizes of 64 and 128 for E/K and H/K, respectively. It was observed in Figs. 8, 9, 10 and 11 that the performance with training in English and testing with Hindi or Kannada languages are decreased because duration characteristics, and stress patterns are different from one language to another in addition to the reasons quoted in the above.

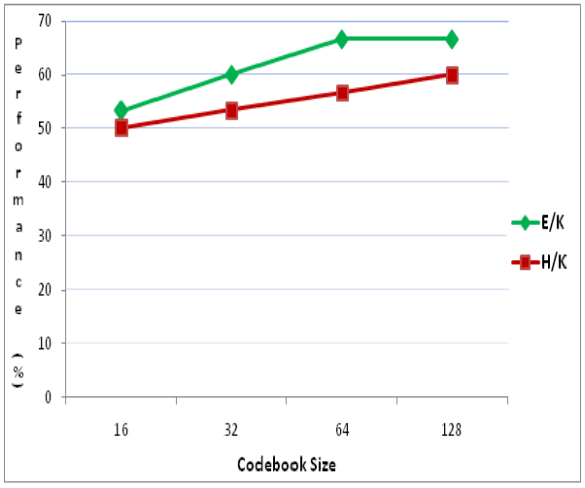

Fig. 10 Performance of Cross-lingual speaker identification system for 10 seconds of data: English and Hindi are the training languages and Kannada is testing language.

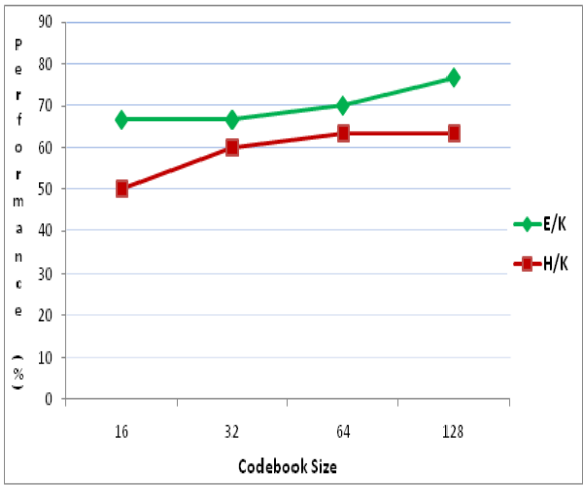

Fig. 11 Performance of Cross-lingual speaker identification system for 15 seconds of data: English and Hindi are the training languages and Kannada is testing language.

Some of the observations can be made from the experimental results are as follows:

i) Mono-lingual results are better with English language.

ii) The Mono-lingual results are better than the Cross-lingual experiments.

iii) As the amount of speech data increases the performance (% identification) also increases in all the experiments.

iv) Use of English language either in training or testing in cross-lingual study gives better identification performance.

6. Conclusion

In this paper, Mono-lingual and Cross-lingual speaker Identification systems are demonstrated using English, Hindi and Kannada languages. We observed that speaker identification system with English language provides good performance in Mono-lingual study. Further, we observed that speaker identification with English language for testing also provides good performance in Cross-lingual study. The experimental studies reveal that better feature extraction and modeling techniques are required in order to improve the performance in both Mono-lingual and Cross-lingual speaker Identification system. Therefore, the high level features like pitch, intonation etc. and modeling techniques like GMM, GMM-UBM and Neural networks can be used to improve the performance. In order to study the robustness of the system needs to be verified with different languages, different data sizes and large amount of speaker set.

Acknowledgment

This work is supported by Visvesvraya Technological University, Belgaum-590018, Karnataka, India.

References

[1] Ahmad Salman, Ejaz Muhammad and Khawar Khurshid, “Speaker Verification Using Boosted Cepstral Features with Gaussian Distributions”, *Proc.IEEE*, 2007.

[2] H. S. Jayanna and S. R. Mahadeva Prasanna, “Analysis,Feature extraction,modeling and Testing techniques for Speaker Recognition”, *IETE Technical Review*, vol. 26, pp. 181–190, May-june 2009.

[3] B. S. Atal, “Automatic recognition of speakers from their voices”, *Proc.IEEE*, vol. 64(4), pp. 460–475, April 1976.

[4] Campbell JP Jr., “Speaker recognition : A Tutorial ”, *Proc.IEEE*, 2007, vol. 85, No. 9, pp. 1437-62, Sep 1997.

[5] Li Deng, Jasha Droppo, Dong Yu, and Alex Acero., “Learning Methods in Multilingual Speech Recognition”, Speech Research Group Microsoft Research Redmond, WA 98052.

[6] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, “Extraction of speaker-specific excitation information from linear prediction”, *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[7] Hemant A. Patil, Sunayana Sitaram, and Esha Sharma, “DA-IICT Cross-lingual and Multilingual Corpora for Speaker Recognition”, *Proc.IEEE*, pp. 187–190, 2009.

[8] Geoffrey Durou., ``Multilingual text independent speaker identification,'' pp. 115-118.

[9] Douglas A. Reynolds, “Automatic Speaker Recognition: Current Approaches and Future Trends”, *ICASSP*, 2001.

[10] Tomi Kinnunen and Haizhou Li, “An Overview of text-Independent speaker Recognition: From Feature to Super Vectors”, *ELSEVIER., Speech Communication*, vol. 52, pp. 12-40, Jan. 2010.

[11] T. Kinnunen, E. Karpov, and P. Franti, “Real-Time Speaker Identification and Verification”, *Proc IEEE. Audio, Speech, Language Processing*, vol. 14(1), pp. 277-288, Jan. 2006.

[12] S. Kwon and S. Narayanan, “Robust speaker identification based on selective use of feature vectors”, *Pattern Recognit. Lett.*, Press, vol. 28, pp. 85-89, Jan. 2007.

[13] H. S. Jayanna and S. R. M. Prasanna, “Variable segmental analysis based speaker recognition in limited data

conditions", *IEEE-Int. Conf. Signal Image Processing.*, Hubli, Karnataka, Dec. 2006.

[14] D.O'Shaughnessy, "Linear Predictive Coding", *IEEE Potentials*, vol. 7, no. 1, pp. 29-32, Feb. 1998.

[15] Patil Hemant Arjun, "Speaker Recognition in Indian Languages: A Feature Based Approach", Indian Institute of Technology, Kharagpur, INDIA , July 2005.

[16] Lawrence R. Rabiner and Ronald W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, First edition, 1978.

[17] Lawrence Rabiner and Biing-Hwang Juang, "Fundamental of Speech Recognition", Pearson Education, Second Impression, 2007.