I.J. Image, Graphics and Signal Processing, **2013, 9, 14-20**
Published Online July 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2013.09.03

Kannada Language Parameters for Speaker Identification with The Constraint of Limited Data

Nagaraja B.G.
Department of Information Science and Engineering
Siddaganga Institute of Technology, Tumkur-572103, Karnataka, India
nagarajbg@gmail.com

H.S. Jayanna
Department of Information Science and Engineering
Siddaganga Institute of Technology, Tumkur-572103, Karnataka, India
jayannahs@gmail.com

Abstract — In this paper we demonstrate the impact of language parameter variability on mono, cross and multi-lingual speaker identification under limited data condition. The languages considered for the study are English, Hindi and Kannada. The speaker specific features are extracted using multi-taper mel-frequency cepstral coefficients (MFCC) and speaker models are built using Gaussian mixture model (GMM)-universal background model (UBM). The sine-weighted cepstrum estimators (SWCE) with 6 tapers are considered for multi-taper MFCC feature extraction. The mono and cross-lingual experimental results show that the performance of speaker identification trained and/or tested with Kannada language is decreased as compared to other languages. It was observed that a database free from ottakshara, arka and anukaranavyayagalu results a good performance and is almost equal to other languages.

***Index Terms*—**Speaker identification, monolingual, crosslingual, multilingual, language parameters, Kannada

I. INTRODUCTION

Speaker identification is a one-to-many comparison, ie., system identifies a speaker from a database of *N* known speakers. Depending on the mode of operation, speaker identification can be either text-dependent or text-independent [1]. In the former case, the speaker must speak a given phrase known to the system, which can be fixed or prompted. In the latter case, the system does not know the phrase spoken by the speaker. Speaker identification can be performed in monolingual, crosslingual and multilingual mode.

In monolingual speaker identification, training and testing languages for a speaker are the same whereas in crosslingual speaker identification, training is done in one language (say x) and testing is done in a different language (say y) [2]. In multilingual speaker identification, speaker specific models are trained in one language and tested with multiple languages.

The state-of-the-art speaker identification systems work only in a single language environment (monolingual) using sufficient data. There are many countries including India are multilingual and hence the effect of multiple languages on a speaker identification system needs to be investigated. Data sparseness is becoming a crucial research concern in automatic speaker recognition system. In non-cooperative scenario such as forensic investigation, the speech data may be a few seconds and task is to identify the speaker. In such application, it is required to validate the speaker using limited amount of speech data. Speaker identification in limited data condition refers to the task of recognizing speakers were both the training and test speech present only for a few seconds [3]. In this work, limited data refer to the case of having 10 seconds of speech data. 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.

Speaker recognition in Indian languages like, Marathi, Hindi, Urdu, Bengali, Oriya, Telugu and Tamil was carried out in [2]. The linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), MFCC and teager energy based MFCC (T-MFCC) features and polynomial classifiers (2^{nd} and 3^{rd} order) were used to build the system. The results showed that the testing language significantly affects the recognition performance in the crosslingual scenario as compared to the training language. The performance degradation due to the mismatch between Mandarin and Sichuan dialect in training and testing was studied in [4]. In that study, a combined GMM, trained by Mandarin and Sichuan dialect was proposed to alleviate the mismatch problem. The result showed that the combined GMM is more robust for the language mismatch condition than the GMM trained solely using Mandarin or Sichuan dialect speech.

The impact of mismatch in training and testing languages on a speaker verification system using English, Hindi and Arunachali languages was carried out in [5]. Speaker verification system was developed using 38-dimensional features (19-MFCC and its first derivatives) and GMM-UBM modeling approach. A training data of 120 seconds and different testing speech data were considered. It was observed that the recognition performance greatly dependent on the training and testing languages. Further, it was observed that if the system is trained with more than one language, the relative recognition performance of the system degrades compared to that of the single language scenarios (monolingual).

In our previous work [6], an attempt was made to identify the speaker in the context of mono and cross-lingual with the constraint of limited data using MFCC as feature vectors and vector quantization (VQ) as modeling technique. It was observed in crosslingual study that the use of English language either in training or testing gives better identification performance.

The state-of-the-art speaker recognition system uses MFCC as a feature for identifying speakers. Tomi kinnunen et al. demonstrated the use of multi-taper MFCC features for speaker verification task in [7]. The basic idea in multi-tapering is to pass the analysis frame through the multiple window functions and then estimate the weighted average of individual sub-spectra to obtain the final spectrum [8]. The experimental results on NIST-2002 and NIST-2008 databases indicated that multi-tapers outperform conventional single-window technique. This work concentrates on text-independent mono, cross and multi-lingual speaker identification under limited data condition using multi-taper MFCC features and GMM-UBM classifier.

The remainder of the paper is organized as follows: Section II describes the database used for the study. Feature extraction using multi-taper MFCC and speaker modeling using GMM-UBM technique is presented in Section III. Section IV outlines the mono, cross and multi-lingual experimental results. The language parameters for Kannada are presented in Section V. Finally, Summary and conclusions of this study are mentioned in Section VI.

II. DATABASE FOR THE STUDY

Since the standard multilingual database is not available, experiments are carried out on an our own created database of 30 speakers who can speak the three different languages (E-English, H-Hindi and K-Kannada). The database includes 17-male and 13-female speakers. The voice recording was done in the 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 smaller 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 detail specifications used for collecting the database are shown in Table 1.

TABLE I. 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

III. FEATURE EXTRACTION AND MODELING

Speech recordings were sampled at the rate of 8 kHz and pre-emphasized (factor 0.97). Frame duration of 20 msec and a 10 msec of overlapping durations are considered. Let $F = [f(0)\ f(1)\ ...\ f(N-1)]^T$ denotes one frame of speech of N samples. Windowed discrete Fourier transform spectrum estimate is given by [7] [9] [10]

$$\hat{S}(f) = \left|\sum_{t=0}^{N-1} w(t)\, f(t)\, e^{-i2\pi ft/N}\right|^2 \tag{1}$$

where $W = [w(0)\ w(1)\ ...\ w(N-1)]^T$ is the time-domain window (Hamming) function. Fig. 1 shows the block diagram representation of the multi-taper MFCC method. Multi-taper spectrum estimator is given by [7]

$$\hat{S}(f) = \sum_{j=1}^{K} \lambda(j) \left|\sum_{t=0}^{N-1} w_j(t)\, f(t)\, e^{-i2\pi ft/N}\right|^2 \tag{2}$$

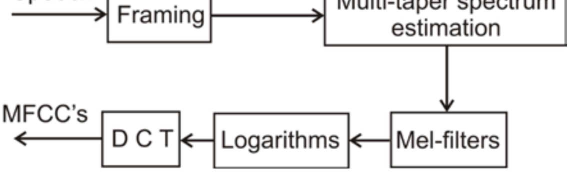

Figure 1. Block diagram of multitaper MFCC technique.

Here K represents the number of multitapers used. $W_j = [w_j(0)\ w_j(1)\ ...\ w_j(N-1)]^T$ is the multitaper weights and $j = 1, 2, ..., K,$ are used with corresponding weights $\lambda(j)$. A number of different tapers have been proposed for spectrum estimation. In [7], it was mentioned that the choice of multi-taper type was found less important than the choice of the number of tapers, K ($3 \leq k \leq 8$). In this work, SWCE multi-taper is used with K=6 windows. A mel-warping is then performed using 22 triangular band pass filters followed by a discrete cosine transform (DCT). A 13-dimensional MFCC feature vectors (excluding 0th coefficient) are finally obtained. Fig. 2 shows the conventional Hamming window and sine tapers with K=6 representation in the frequency domain.

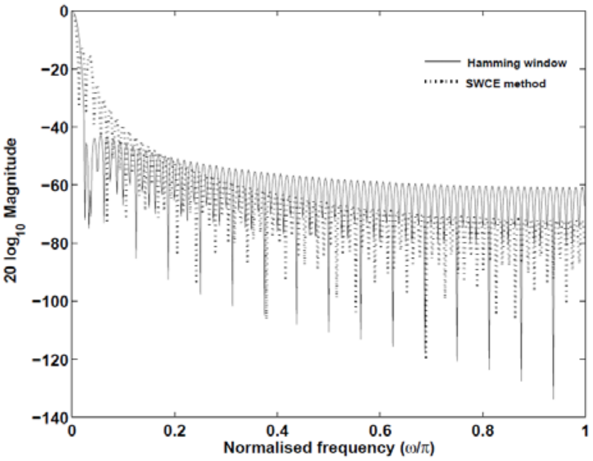

Figure 2. Frequency domain for a window of size 160 samples.

In GMM-UBM systems, speech data are pooled from many speakers to train a single independent model, known as UBM. The *k*-means algorithm was used to obtain the initial estimate for each cluster. UBM acts as a speaker independent model in the GMM-UBM system. In [11] and [3], it was mentioned that there are no criteria to select number of speakers and the amount of data to train the UBM. For building the gender independent UBM, we have used roughly two hours of speech data from 138 speakers of YOHO database. The speaker specific models were created by adapting only the mean vectors of the UBM using maximum a posteriori (MAP) adaptation algorithm [1]. The parameters of the GMM models (mean vector, covariance matrix and mixture weights) were estimated using expectation maximization (EM) algorithm. We have modeled speakers by using GMMs with 8, 16, 32, 64 and 128 mixtures.

IV. SPEAKER IDENTIFICATION RESULTS

In this section, mono, cross and multi-lingual speaker identification results are presented. In all our experiments, the speaker set (30 speakers) and amount of speech data (10 seconds) are kept constant to make a relative comparison of the performance of speaker identification. Note: X/Y indicates training with language X and testing with language Y.

A. *Monolingual Experiments*

The monolingual experimental results for different Gaussian mixtures are given in Fig. 3. The results show that the speaker identification system yields good performance of 66.66% for 64 and 128 Gaussian mixtures when trained and tested with English language. The performance of the speaker identification system trained and tested with Hindi language is 66.66% for 128 Gaussian mixtures. The speaker identification system trained and tested with Kannada language gives the highest performance of 56.66% for 128 Gaussian mixtures.

B. *Crosslingual Experiments*

The speaker identification system trained in English, and tested with Hindi and Kannada languages for different Gaussian mixtures are shown in Fig. 4. The highest performance obtained for E/H and E/K is 53.33% for 128 Gaussian mixtures. The speaker identification system trained in Hindi, and tested with the English and Kannada languages for different Gaussian mixtures are shown in Fig. 5. The speaker identification system trained in Hindi and tested with the English language (H/E) gives the highest performance of 50% for 64 and 128 Gaussian mixtures.

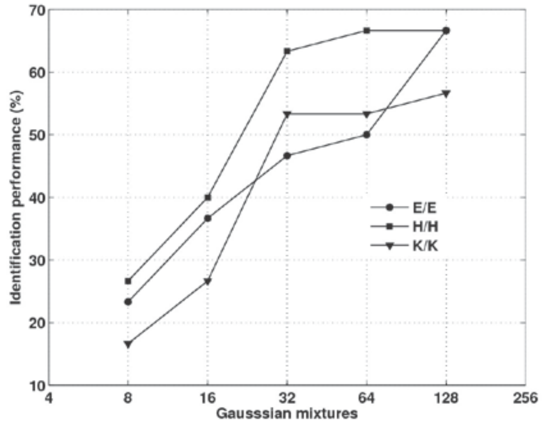

Figure 3. Monolingual speaker identification performance.

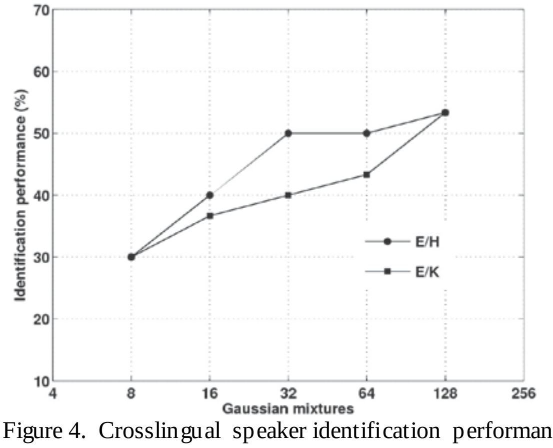

Figure 4. Crosslingual speaker identification performance (E/H and E/K).

The performance of the speaker identification system trained in Hindi language and tested with Kannada language (H/K) is 43.33% for 128 Gaussian mixtures. The speaker identification system trained in Kannada, and tested with English and Kannada languages for different Gaussian mixtures are shown in Fig. 6. The highest performance obtained for K/E is 50% for 128 Gaussian mixtures and K/H is 46.66% for 64 and 128 Gaussian mixtures, respectively.

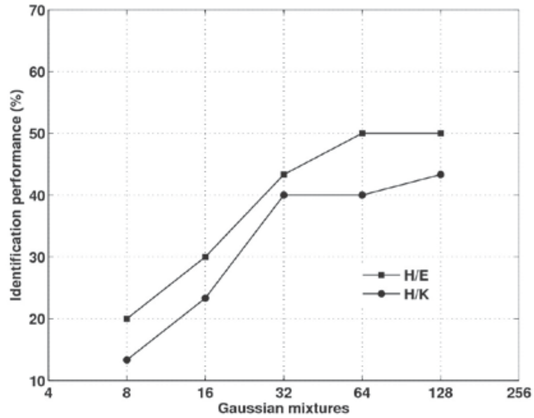

Figure 5. Crosslingual speaker identification performance (H/E and H/K).

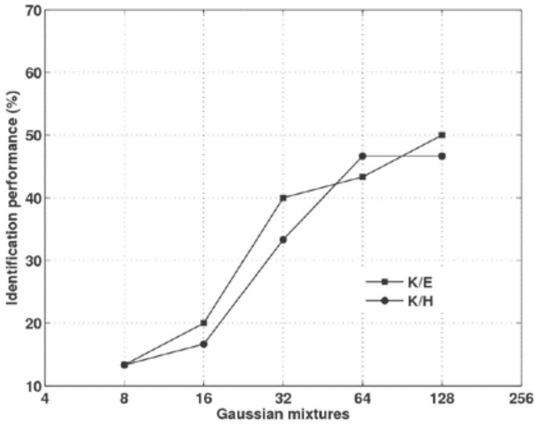

Figure 6. Crosslingual speaker identification performance (K/E and K/H).

C. Multilingual Experiments

In multilingual speaker identification, some speakers in database are trained and tested in language A, some speakers are in language B and so on. The multilingual experimental results for different Gaussian mixtures are given in Fig. 7. The multilingual speaker identification system yields a highest performance of 70% for 128 Gaussian Mixtures.

Some of the observations we made from the mono, cross and multi-lingual results are as follows:

(1) It was observed that the results are better for monolingual experiments than the crosslingual. This may be due to the variation in fluency and word stress when the same speaker speaks different languages and also due to different phonetic and prosodic patterns of the languages [12].

(2) Mono and cross-lingual experimental results show that the performance of speaker identification trained and/or tested with Kannada language is decreased.

(3) In crosslingual speaker identification, the use of the English language either in training or testing gives a better identification performance.

(4) The multilingual results are better than the monolingual and crosslingual experiments. This may be due to the better discrimination between the trained and testing models (multiple languages) in multilingual scenario.

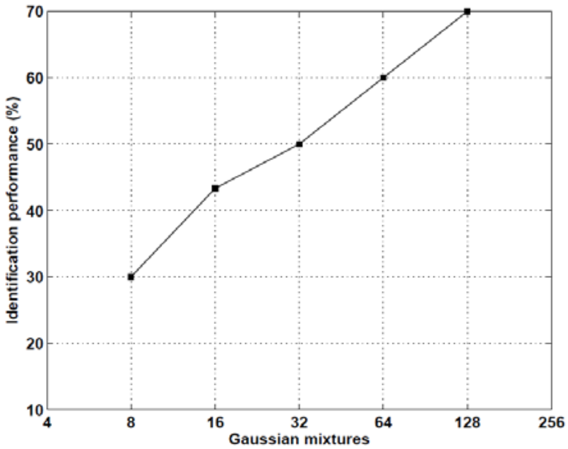

Figure 7. Multilingual speaker identification performance.

V. LANGUAGE PARAMETERS FOR KANNADA LANGUAGE

Kannada language is an alphasyllabary [13]. The majority of its words are bi and tri-syllabic, with four, five and six-syllable words also in the vocabulary [13]. Kannada has over 400 of the written symbols that are called akshara. We examined the speech sample on whole words as a function of the type of Akshara they contained. Words of the type, such as in (avanu ‘he’), (kelasa ‘work’), (kamala ‘lotus’), (galisu ‘earn’), etc., contained only consonant-vowel (CV) akshara with the inherent vowel. Words of the type that contain at least one consonant-consonant-vowel (CCV) akshara such as in (manassu ‘mind’), (tapassu ‘austerities’), (belli ‘silver’), (kabbina ‘iron’), (ukku ‘steel’), (tamma ‘younger brother’), (anna ‘elder brother’), etc., these could involve sequences of two identical consonants or sequences of different consonants such as as in, (vidye ‘knowledge’), (dwara ‘entry’), etc. Speaker’s found difficulty in reading complex akshara.

It is observed in Figs. 8-11 that when the same speaker utters a word in English, there is no much pause in the speech signal, but when he/she pronounces the corresponding word in Kannada there is a long pause in the speech signal. The voice activity detection test$\geq$ (0.06 times the average energy) performed on both English and Kannada words uttered by the same speaker denote that the energy frames in English words are more than Kannada words. More energy frames indicates that there is more speaker specific information in speech. Table 2 shows number of energy frames obtained for few English and its corresponding Kannada words uttered by the same speaker. The presence of ottakshara (CCV akshara like, /gga/ in /agga), arka (refers to a specific /r/ in consonant clusters) [13] and anukaranavyayagalu (/julujulu/) leads to long pause and hence less number of energy frames in Kannada words.

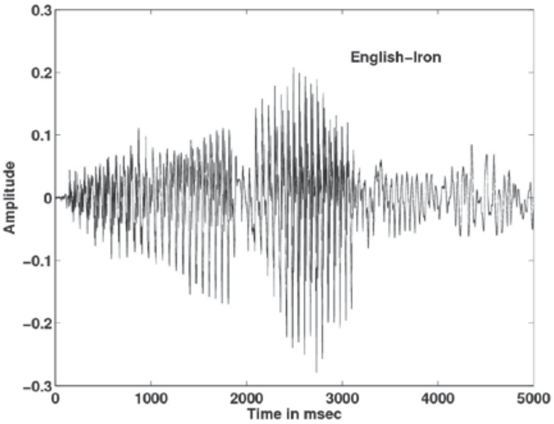

Figure 8. Speech signal of the word Iron in English.

In order to alleviate this problem, a new database was created using the same speakers in Kannada language where words which are free from ottakshara, arka and anukaranavyayagalu were considered. Mono and cross-lingual experiments were conducted for the new Kannada database created shown in Figs. 12-14. The monolingual results obtained are almost comparable to that of the English and Hindi languages (Fig. 12).

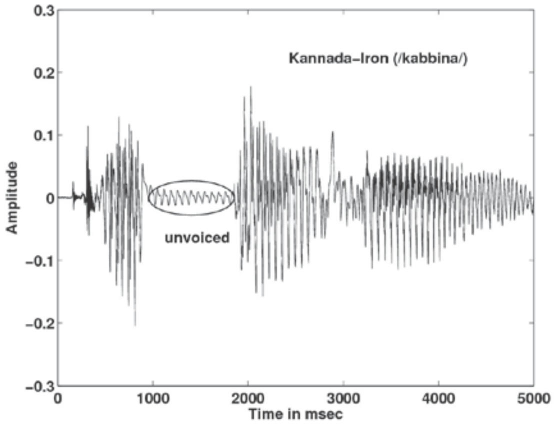

Figure 9. Speech signal of the word Iron in Kannada.

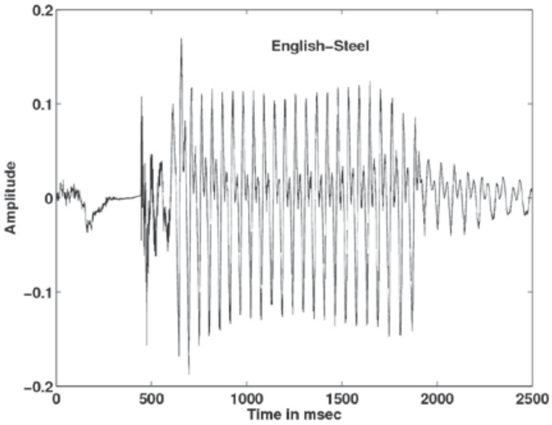

Figure 10. Speech signal of the word Steel in English.

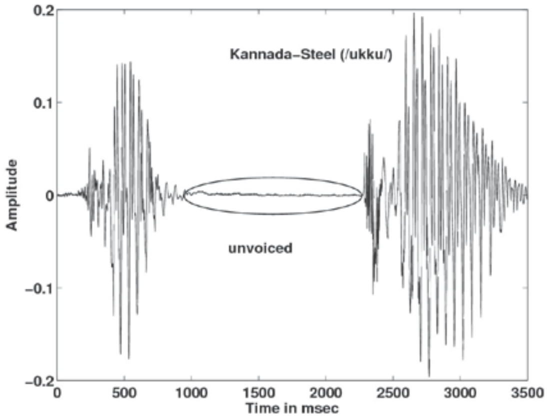

Figure 11. Speech signal of the word Steel in Kannada.

TABLE II. TOTAL NUMBER OF ENERGY FRAMES OBTAINED FOR FEW ENGLISH AND ITS CORRESPONDING KANNADA WORDS UTTERED BY THE SAME SPEAKER

English	Energy Frames	Kannada	Energy Frames
Iron	72	/kabbina/	62
Silver	79	/belli/	59
Steel	50	/ukku/	27
Simple	77	/sulabha/	68
Photo	48	/chitra/	40

The speaker identification system trained in Kannada, and tested with Hindi and English languages for different Gaussian mixtures are shown in Fig. 13. The highest performance obtained for K/E is 56.66% for 128 Gaussian mixtures and K/H is 50% for 128 Gaussian mixtures, respectively. The speaker identification system trained in English and Hindi, and tested with Kannada language is shown in Fig. 14. The highest performance obtained for E/K is 50% for 128 Gaussian mixtures and H/K is 43.33% for 64 and 128 Gaussian mixtures, respectively.

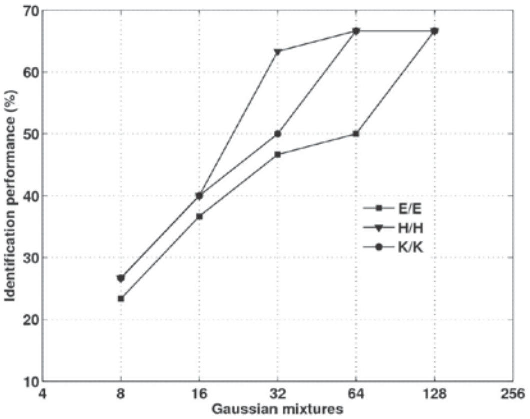

Figure 12. Monolingual speaker identification performance after considering the language parameters for Kannada.

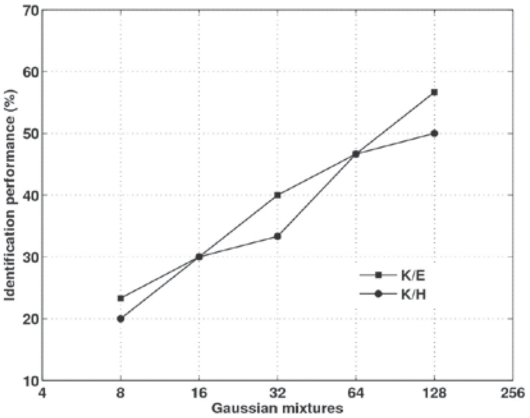

Figure 13. Crosslingual speaker identification performance (K/E and K/H) after considering the language parameters for Kannada.

VI. CONCLUSION

In this work, mono, cross and multi-lingual speaker identification with the constraint of limited data are demonstrated using Indian English, Hindi and Kannada languages. Experimental results showed that the language mismatch between training and testing data leads to a considerable performance degradation. It was also found out in the mono and cross-lingual study that the speaker identification system does not yield satisfactory performance with Kannada as training and/or testing language. The presence of ottakshara, arka and anukaranavyayagalu leads to long pause and hence the less number of energy frames (features) in Kannada words. Therefore, we suggest to avoid the use of complex akshara (CCV) while building the database in Kannada language for speaker identification. Future work includes in-depth study of the language parameters for speaker identification using different languages.

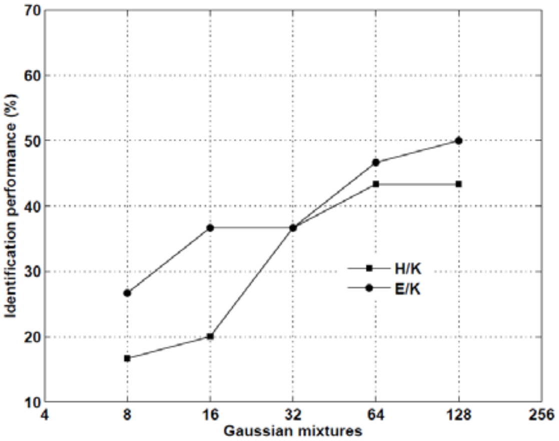

Figure 14. Crosslingual speaker identification performance (E/K and H/K) after considering the language parameters for Kannada.

ACKNOWLEDGMENT

This work is supported by Visvesvaraya Technological University (VTU), Belgaum-590018, Karnataka, India.

REFERENCES

[1] Reynolds D.A. and Rose R.C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. Speech and Audio Processing*. 72-83, 1995.

[2] Arjun P.H., "Speaker Recognition in Indian Languages: A Feature Based Approach", *Ph.D. dissertation*, Indian Institute of Technology, Kharagpur, India, 2005.

[3] Jayanna H.S., "Limited data Speaker Recognition", *Ph.D. dissertation*, Indian Institute of Technology, Guwahati, India, 2009.

[4] Zhao Jing, Gong Wei-guo and Yang Li-ping., "Mandarin-Sichuan dialect bilingual text-independent speaker verification using GMM", *Journal of Computer Applications*, vol. 28, no. 3, pp. 792-794, 2008.

[5] Bhattacharjee U. and Sarmah K., "A multilingual speech database for speaker recognition", *Proc. IEEE, ISPCC*, pp. 15-17, 2012.

[6] Nagaraja B.G. and Jayanna H.S., "Mono and Cross-lingual speaker identification with the constraint of limited data", *Proc. IEEE, PRIME*, pp. 439-449, 2012.

[7] Kinnunen T, Saeidi R, Sedlak F, Lee K.A, Sandberg J, Hansson-Sandsten M. and Li H., "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification", *IEEE Transaction on Audio, Speech and Language Processing 20*, pp. 1990-2001, 2012.

[8] Kinnunen T, Saeidi R, Sandberg J. and Hansson-Sandsten M., "What Else is New Than the Hamming- Window? Robust MFCCs for Speaker

Recognition via Multitapering", *Proc. Interspeech 2010*, pp. 2734-2737, 2010.

[9] Percival D.B. and Walden A.T., "Spectral Analysis for Physical Applications", Cambridge, *MA: Cam- bridge Univ. Press*, 1993.

[10] Thomson D.J., "Spectrum estimation and harmonic analysis", *Proc. IEEE*, pp. 1055-1096, 1982.

[11] Reynolds D.A., "Universal Background Models", Encyclopedia of Biometric Recognition, *Springer*, Journal Article, Feb. 2008.

[12] Geoffrey Durou., "Multilingual text-independent speaker identification", *Proc. MIST'99 Workshop, Leusden, Netherlands*. pp.115-118, 1999.

[13] Sonali Nag, "Early reading in Kannada: the pace of acquisition of orthographic knowledge and phonemic awareness", *Journal of Research in Reading*, vol. 30, Issue 1, pp. 7-22, 2007.

Mr. Nagaraja B.G. was born in India in the year 1982. He received the B.E. degree in Electronics and Communications Engineering from Bapuji Institute of Technology, Davangere, Karnataka, India, in the year 2004 and the M.Tech degree in Computer Science and Engineering from East-West Institute of Technology, Bangalore, India, in the year 2009. He is currently pursuing the Ph.D. degree at the Visvesvaraya Technological University, Belgaum, Karnataka, India. He has been working as a Research Assistant in the department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India. His research interests include speech, Multilingual Speaker Recognition.

Dr. H.S. Jayanna was born in India in the year 1970. He received the B.E. degree in instrumentation and electronics engineering from Dr. Ambedkar Institute of Technology, Bangalore University, Bangalore, India, in the year 1992 and the M.E. degree in electronics from University Visvesvaraya College of Engineering, Bangalore, India, in the year 1995 & his Ph.D. in Electronics and Communication Engineering from the prestigious Indian Institute of Technology (IIT), Guwahati, India in the year 2009. He has published a number of papers in various national, international journals & conferences apart from guiding a number of UG, PG & Research Scholars. Currently, he is working as Professor & Head of the department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India. His research interests are in the area of speech, Limited Data Speaker Recognition, Image Processing, Computer Networks and Computer Architecture.