Multilingual Speaker Recognition on Indian Languages

Sourjya Sarkar, K. Sreenivasa Rao, Dipanjan Nandi and Sunil Kumar S. B.
School of Information Technology
Indian Institute of Technology Kharagpur, India - 721302
sourjyasarkar@gmail.com, ksrao@iitkgp.ac.in, dipanjannandi@sit.iitkgp.ernet.in, sunil220552@gmail.com

***Abstract*—In this paper we explore the performance of multilingual speaker recognition systems developed on the IITKGP-MLILSC speech corpus. Closed-set speaker identification and speaker verification experiments are individually conducted on 13 widely spoken Indian languages. In particular, we focus on the effect of language mismatch in the speaker recognition performance of individual languages and all languages together. The standard GMM-based speaker recognition framework is used. While the average language-independent speaker identification rate is as high as 95.21%, an average equal error rate of 11.71% shows scope for further improvement in speaker verification performance.**

***Index Terms*—Multilingual Speaker Recognition, Speaker Verification, Gaussian Mixture Models, Indian Languages**

I. Introduction

Speaker recognition (SR) is the task of recognizing a person by processing his/her spoken utterances. An ideal SR system is expected to perform effectively irrespective of changes in session, emotion, health and language of a speaker. Prior work demonstrates the prominent effect of spoken languages in text-independent SR accuracy [1] [2] [3]. Multilingual speaker recognition has thus been a field of active research in recent years [4] [5]. Apart from the common issues concerning conventional SR systems, a major challenge in this field is the collection of adequate data for preparing a speech corpus. In the context of Indian languages, a suitable corpus should span most of the local languages spoken in the country. The distribution of speakers in each language should be made according to a recent census. Care must be taken to retain accent variations in each language.

Most of the recent studies in Indian languages are either focused on a particular geographical area [4] [6] or restricted within a few languages (typically 4) [7] of the country. The IITKGP-MLILSC (IITKGP-Multilingual Indian Language Speech Corpus) introduced in [8] addresses some of the aforementioned issues. The entire corpus spans 27 Indian languages including 16 most widely spoken ones. Though the primary motivation was to study implicit Indian language recognition, [8] [9] [10] availability of sufficient number of speakers in each language also provides scope for exploring multilingual SR.

The state-of-the-art speaker modelling techniques for SR include Gaussian mixture models (GMM) [11], Gaussian Mixture Model-Universal Background Model (GMM-UBM) [12] and combined Gaussian Mixture Model-Support Vector Machine (GMM-SVM) [13]. GMMs are used in all of these techniques to capture acoustic feature distribution and spectral shapes. The performance of the GMM-UBM and GMM-SVM-based methods depend significantly on the quality of a UBM. However, in the context of multilingual corpora, designing an appropriate UBM is a complicated task [2]. In the present study, we therefore restrict ourselves to the simple GMM-based approach for speaker identification and speaker verification. In particular, we study the effect of language mismatch in SR performance in which trial utterances of any speaker are evaluated against speaker models of various other languages.

The rest of the paper is organized as follows. The development of multilingual SR systems has been described in Section II. Results are discussed in Section III. Conclusion and future work plan are described in Section IV.

II. Development of Proposed Multilingual Speaker Recognition Systems

In this section we discuss the development of the multilingual speaker recognition systems in details.

A. Speech Corpus

The Indian Institute of Technology Kharagpur- Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) [8] was used for developing the speaker recognition systems. The corpus consists of data from 27 Indian languages. Out of these, the data for 16 languages was recorded from broadcasted TV talk shows, live shows, interviews and news bulletins while the remaining were collected from All India Radio (AIR) news bulletins of broadcasted radio channels. Each language comprises minimum 1 hour of speech data from at least 10 speakers which includes both male and female speakers. The speech data was collected at a sampling rate of 16 kHz with 16 bits per sample. All speech signals were down-sampled to 8 kHz for the present speaker recognition study. The broadcasted television channels were accessed using VentiTV software and the Pixelview TV tuner card. Audacity software was used for recording the speech data from TV channels. The language data of broadcasted radio channels was collected from the archives of Prasar Bharati and AIR website.

In order to avoid channel mismatch, only data from 13 languages recorded over TV broadcast channels were considered

978-1-4799-2275-8/13/$31.00 ©2013 IEEE

for the present work. Approximately 3-4 minutes of speech from each speaker was used for training the speaker models. 10 test utterances (approximately 10 secs each) from each speaker were used for evaluation. The distribution of speakers in each language has been summarized in Table I.

B. Feature Extraction

The human speech production system consists of a vocal tract and a source for exciting the vocal tract resonator. During speech production, vocal tract system behaves like a time varying resonator or may be treated as a time varying filter. This time varying filter characterizes the variations in the vocal tract shape in the form of resonances and anti-resonances that occur in the speech spectrum. Parameterization techniques like linear prediction cepstral coefficients (LPCCs) and mel-frequency cepstral coefficients (MFCCs) [14] are available for modeling vocal tract information. Since mel-filters are based on human perceptual nature, we have used MFCC feature for this speaker recognition study. The steps for calculating the MFCCs from the speech signal are discussed below.

(i) Pre-emphasis : It refers to a filtering technique that emphasizes the higher frequencies. Some voiced sounds have a steep roll-off in the high frequency region. So, to balance the speech spectrum of voiced sounds, high-frequency filtering is needed. The procedure for performing the pre-emphasis is shown in the equation 1.

$$H(z) = 1 - \alpha z^{-1} \tag{1}$$

where the value of α controls the slope of the filter and is usually between 0.9 to 1.0.

(ii) Windowing : The human speech signal is a quasi-stationary signal. The voiced sound units are quasi-periodic in nature, whereas, the unvoiced sound units are noise like signal. Therefore, the analysis of speech signal for any speech applications must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over the range of 10-30 ms frame size and frame shift of half of the frame size. The blocked frames are Hamming windowed. Hamming window is used to reduce the edge effect while taking the discrete Fourier transform (DFT) on the signal.

(iii) Discrete Fourier Transform (DFT) : Each windowed frame is converted into magnitude spectrum by applying DFT using the equation 2.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}, \quad 0 \leq k \leq N-1. \tag{2}$$

where $x(n)$ is the samples of the windowed speech signal. $X(k)$ is the magnitude spectrum of windowed speech signal and N is the number of points used to compute the DFT.

(iv) Mel-spectrum : The Mel-spectrum is computed by passing the DFT spectrum through a set of band-pass triangular filters known as mel-filter bank. A mel is a unit of perceived speech frequency or a unit of tone. The mel scale is therefore a mapping between the physical frequency scale (Hz) and the perceived frequency scale (Mels). The approximation of mel from physical frequency can be expressed by the following equation [14][15].

$$f_{mel} = 2595 \log(1 + \frac{f}{700}) \tag{3}$$

where f denotes the physical frequency and f_{mel} denotes the perceived mel-frequency. The mel-spectrum values or mel-frequency coefficients of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the triangular mel-weighting filters.

$$S(m) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k), \quad 0 \leq m \leq M-1. \tag{4}$$

where $S(m)$ is the mel-frequency coefficients and M is total number of triangular mel-weighting filters.

(v) Inverse discrete cosine transform (IDCT) : The $\log$ operation is performed on the Mel-frequency coefficients. The IDCT is then applied to obtain the cepstral coefficients. This yields a signal in the cepstral domain. MFCC is computed as follows :

$$c(n) = \sum_{m=0}^{M-1} \log(S(m)) \cos(\tfrac{\pi n(m-0.5)}{M}), \quad n = 0, 1, 2, ..., C-1 \tag{5}$$

where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. The zeroth coefficient represents the average log-energy of the input signal.

Mel-frequency cepstral coefficients excluding the 0th coefficient, were derived from a 26 channel Mel-scaled filter-bank constrained in the frequency band of 300-3500 Hz. The MFCC feature vector represents only the information present at power spectral envelope of a single frame. However, speech signal may also carry information in the dynamics i.e., the knowledge present in the trajectories of the MFCC coefficients over time. These are known as delta coefficients or velocity coefficients. The formula used for calculating the delta coefficients is given below :

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^{N} n^2} \tag{6}$$

where d_t is a vector of 13 dimensions consisting of delta coefficients or velocity coefficients computed from the frame t in terms of the static coefficients c_{t+N} to c_{t-N}. The delta-delta (acceleration) coefficients are calculated in the same way, but, calculated from the static coefficients, not from the deltas. The velocity and acceleration coefficients over a frame span

of 2 were appended to form a resultant 39 dimensional feature vector. All vectors were subjected to cepstral mean subtraction followed by cepstral variance normalization.

C. Development of Speaker Models

The acoustic features (as described in Section II-B) were modeled by Gaussian probability density functions (PDFs). Since a single PDF is unsuitable for modeling speaker-dependent spectral shapes, a mixture of single densities i.e., a Gaussian Mixture Model (GMM) was used for modeling the complex structure of its distribution. For a D-dimensional feature vector denoted as x_t, the mixture density for a language model λ is defined as weighted sum of M component Gaussian densities as given by the following equation [11]

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i p_i(x_t) \tag{7}$$

where w_i are the weights and $p_i(x_t)$ are the component densities. Each component density is a D-variate Gaussian function of the form

$$p_i(x_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}[(x_t-\mu_i)'\Sigma_i^{-1}(x_t-\mu_i)]} \tag{8}$$

where μ_i is a mean vector and Σ_i covariance matrix for i^th component. The mixture weights have to satisfy the constraint [11]

$$\sum_{i=1}^{M} w_i = 1 \tag{9}$$

The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities. These parameters are collectively represented by

$$\lambda = \{w_i, \mu_i, \Sigma_i\}; \quad i = 1, 2, ..., M \tag{10}$$

To determine the model parameters of a GMM for a particular speaker speaking a specific language, the model has to be trained. The maximum likelihood (ML) procedure was used for parameter estimation. The main objective of ML estimation is to derive the optimum model parameters that can maximize the likelihood of the training data for a given GMM. The likelihood value is however a higher order nonlinear function of the model parameters and therefore, direct maximization is not possible. Instead, maximization was carried out iteratively using an Expectation Maximization (EM) algorithm [16]. The EM algorithm begins with an initial model λ and estimates a new model such that the likelihood of the model increases with each iteration. This new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained or a certain predetermined number of iterations have been made. The performance of EM algorithm depends on the initialization. In the present work, GMM parameters were initialized using vector quantization and $M = 16$ was empirically found to perform best. In each iteration the posterior probabilities for the i^{th} mixture was computed as given by the following equation [11] :

$$Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{11}$$

The model parameters were updated according to the following expressions [11] :

The updated mixture *weight* is

$$\overline{w_i} = \frac{\sum_{t=1}^{T} Pr(i|x_t)}{T} \tag{12}$$

The updated *mean* vector is

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} Pr(i|x_t) x_t}{\sum_{t=1}^{T} Pr(i|x_t)} \tag{13}$$

The updated *covariance matrix* is

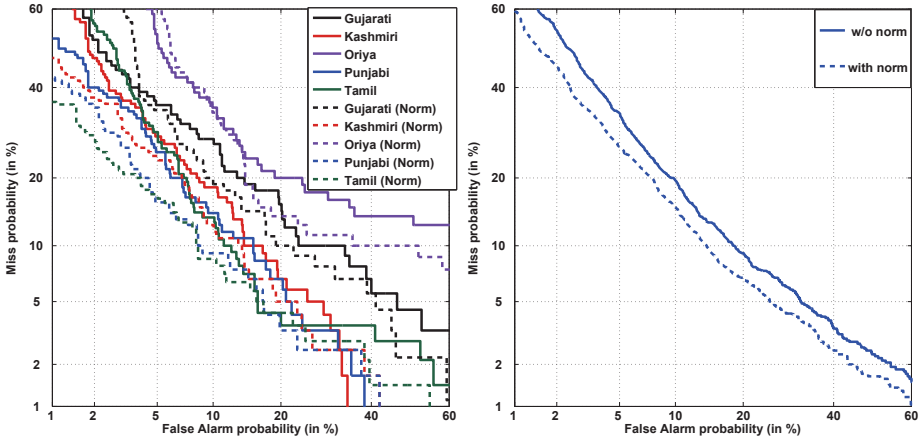
$$\overline{\sigma_i^2} = \frac{\sum_{t=1}^{T} Pr(i|x_t)|x_t - \overline{\mu_i}|^2}{\sum_{t=1}^{T} Pr(i|x_t)} \tag{14}$$

where, T denotes the total number of feature vectors in a speaker's training utterance. In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. It is more common to use diagonal covariance matrices for GMM, since linear combination of diagonal covariance Gaussians has the same modeling capability like full matrices. [17]. Another reason is that speech utterances were parameterized using MFCC features (Section II-B) which are nearly uncorrelated thus allowing diagonal covariance to be used by the GMMs. Around 50 iterations of the EM algorithm were carried out at which point the model parameters reached their optimum values with negligible changes in any subsequent iteration. The effect of language mismatch in SR performance was studied for both speaker identification and speaker verification, as discussed in Sections II-D and II-E respectively.

D. Speaker Identification

Ten test utterances from each speaker (approx. 10 secs in length) were used for evaluation. For language-mismatched SR, the utterances from each speaker were scored against all the speaker models irrespective of their languages. The identification rate (IR) is given by the ratio of the number of correctly classified test cases and the total number of trials. The overall IR for all speakers and the individual IRs for each language were calculated. The evaluation process has been summarized in Equations 15, 16 and 17 respectively.

Given a set of speaker models $\{\lambda_1, \lambda_2 ... \lambda_N\}$, the posterior probability of a speaker model λ_S given a set of i.i.d test



(a) Individual Languages (b) All Languages

Fig. 1. Det Plots showing the performance of SV systems for a) Individual Languages and b) All Languages in mismatched conditions. The solid and broken lines correspond to the unnormalized and normalized scores respectively

vectors $X = \{x_1, x_2...x_T\}$ is calculated by Bayes' theorem.

$$p(\lambda_S|X) = \frac{p(X|\lambda_S)p(\lambda_S)}{p(X)} \quad 1 \leq S \leq N \tag{15}$$

where

$$p(X|\lambda_S) = \prod_{i=1}^{T} p(x_i|\lambda_S) \tag{16}$$

Noting that the denominator in Eq. 15 is a normalizing factor and assuming all speakers to be equiprobable $(i.e., p(\lambda_S) = \frac{1}{N})$, the identified speaker model is the one producing the maximum likelihood score.

$$\lambda^* = \arg \max_{1 \leq i \leq N} p(X|\lambda_i) \tag{17}$$

E. Speaker Verification

For each enrolled speaker of a particular language, a set of 11 cohort speaker models from other languages in the entire corpus were selected as the anti-speaker models, as proposed in [18]. Each test utterance from the enrolled speaker was scored against 12 speaker models (11 cohorts + 1 true speaker) as given in Equation 16. The true score (true speaker) and false scores (impostors) obtained, were used to calculate the miss and false alarm probabilities, respectively. These probabilities were subsequently plotted as a Detection Error Tradeoff (DET) curve [19]. The Equal error rate (EER), determined from the DET curve was used as the evaluation metric for speaker verification. For the overall SR performance in language mismatched condition, a total 1770 (10 test cases $\times$ 177 enrolled speaker) true trails and 19470 false trials were evaluated. The SR performance of each language in the mismatched condition were also individually studied. In the latter case, the number of true and false trials depends on the number of speakers enrolled in each language.

F. Score Normalization

For an improved speaker verification performance we have used the geometric mean method for score normalization [18]. Given a test vector 'x' in any trial and the cohort set $\Lambda = \{\Lambda_1, \Lambda_2...\Lambda_{N-1}\}$, the mean log-likelihood score generated by the cohort models is subtracted from the target speaker (λ) score as shown in Equation 18.

$$Score_{norm} = log(p(x|\lambda)) - \frac{1}{N-1}\sum_{i=1}^{N-1} log(p(x|\Lambda_i)) \tag{18}$$

III. Results and Discussion

The performance of the multilingual SR systems has been summarized in Table I. Figure 1 shows the DET curves for the speaker verification (SV) systems for a) 5 individual languages and b) all languages together in mismatched conditions. To maintain clarity in Figure 1(a), we selectively plotted the curves for those languages which showed at least 2% EER improvement after score normalization. These languages have been highlighted in Table 1.

The speaker identification (SI) accuracy for most languages is above a threshold of 93% with Urdu and Oriya being the only exceptions. An overall average SI rate of 95.21% demonstrates the robustness of GMM classifiers towards misclassifications due to language mismatch.

Contrary to the high SI rate, an overall EER of 13.36% shows medium SV performance. An overall improvement of 1.67% is achieved with score normalization. The improvement is consistent in individual language-wise recognition performances. The most prominent impact of score normalization is seen in case of Oriya, Gujarati and Kasmiri, with an average EER reduction of 3.18%. The classifiers for Indian English, Hindi, Tamil, Punjabi and Urdu shows significant robustness towards language mismatch, with an average EER of 8.97%. However, the non-uniform distribution of speakers in different

TABLE I
SUMMARY OF PERFORMANCE OF THE SPEAKER RECOGNITION SYSTEMS FOR DIFFERENT INDIAN LANGUAGES IN LANGUAGE MISMATCHED CONDITION

Training Languages	No. of Speakers	Speaker Identification Rate (%)	Equal Error Rate (%)	
			w/o Normalization	with Normalization
Assamese	10	94.54	17.27	15.45
Bengali	17	96.47	15.29	13.53
Gujarati	10	93.33	**17.78**	**14.44**
Hindi	26	99.23	11.54	11.53
Ind English	17	100.00	08.00	07.00
Kashmiri	14	93.33	**13.33**	**10.88**
Malayalam	14	98.00	14.00	13.00
Marathi	11	98.75	15.00	13.75
Nepali	10	96.25	15.00	13.75
Oriya	10	80.00	**20.00**	**16.25**
Punjabi	12	97.50	**11.67**	**09.17**
Tamil	12	97.86	**10.71**	**08.57**
Urdu	14	89.29	08.57	08.57
Overall	177	**95.21**	**13.36**	**11.71**

languages prevents concluding any relation between the number of speakers and the SR performance in each language.

IV. CONCLUSION AND FUTURE WORK PLAN

In this work we presented a brief study of the multilingual SR systems developed using IITKGP-MLILSC. The SR performances of 13 Indian languages were demonstrated. The impact of language mismatch was found to be more pronounced in case of speaker verification compared to speaker identification. Future work shall include exploring the GMM-UBM [12] and GMM-SVM [13] frameworks for improved multilingual speaker verification.

REFERENCES

[1] R. Auckenthaler, M. Carey, and J. Mason, "Language dependency in text-independent speaker verification," in *Proc. IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP '01)*, 2001, pp. 441–444.

[2] N. Kleynhans and E. Barnard, "Language dependence in multilingual speaker verification," in *Proc. The 16th Annual Symposium of the Pattern Recognition Association of South Africa*, 2005, pp. 117–122.

[3] B. Nagaraja and H. Jayanna, "Multilingual Speaker Identification with the Constraint of Limited Data Using Multitaper MFCC," pp. 127–134, 2012.

[4] U. Bhattacharjee and A. Sarmah, "A multilingual speech database for speaker recognition," in *Proc. IEEE International Conference on Signal Processing, Computing and Control (ISPCC)*, 2012, pp. 1–5.

[5] L. Mary, K. S. Rao, and B.Yegnanarayana, "Neural network classifiers for language identification using syntactic and prosodic features," in *Proc. 2nd Int. Conf. Intelligent Sensing and Information Processing (ICISIP-2005)*, Chennai, India, January 2005.

[6] L. Mary, K. S. Rao, S. Gangashetty, and B.Yegnanarayana, "Neural network models for capturing duration and intonation knowledge for language and speaker identification," in *Proc. 8th Int. Conf. on Cognitive and Neural systems*, Boston, MA, USA., May 2004.

[7] K. S. Rao, S. Maity, and V. R. Reddy, "Pitch synchronous and glottal closure based speech analysis for language recognition," *International Journal of Speech Technology*, vol. Springer (Accepted, DOI: 10.1007/s10772-013-9193-5), 2013.

[8] S. Maity, A. Vuppala, K. S. Rao, and D. Nandi, "IITKGP-MLILSC Speech Database for Language Identification," in *Proc. IEEE 18th National Conference on Communications*, 2012, pp. 1–5.

[9] V. R. Reddy, S. Maity, and K. S. Rao, "Identification of Indian languages using multi-level spectral and prosodic features," *International Journal of Speech Technology (Springer)*, vol. DOI: 10.1007/s10772-013-9198-0, 2013.

[10] K. S. Rao, S. Maity, and V. R. Reddy, "Pitch synchronous and glottal closure based speech analysis for language recognition," *International Journal of Speech Technology (Springer)*, vol. DOI: 10.1007/s10772-013-9193-5, 2013.

[11] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[12] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing.*, vol. 10, no. 1, pp. 19–41, 2000.

[13] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, July 2006.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 28, no. 28, pp. 357–366, Aug. 1980.

[15] J. R. Deller jr., John H. L. Hansen and J G Proakis, "Discrete-time processing of speech signal," *IEEE press*, 2000.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[17] Q. Hong and S. Kwong, "A discriminative training approach for text-independent speaker recognition," *Signal Processing*, vol. 85, no. 7, pp. 1449 – 1463, 2005.

[18] C. Liu, H. Wang, and C. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, 1996.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. The European Conference on Speech Communication and Technology*, 1997, pp. 1895–1898.