Identification of the major language families of India and evaluation of their mutual influence

Debapriya Sengupta* and Goutam Saha

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721 302, India

A language family is a group of languages which have descended from a common mother language. Since the ancestor is common, these languages are expected to be similar in some respect and manifest the similarity in scientific experiments. In language identification, language-specific features are extracted from speech and a model is created which represents the language. This work extends the language identification framework to capture features common to language families and create models which can efficiently represent the language families. Mel frequency cepstral coefficient (MFCC) and speech signal-based frequency cepstral coefficient (SFCC) are used as primary feature extraction tools. A combination of these along with shifted delta coefficient (SDC) gives the final set of features. The work uses Gaussian mixture model (GMM) and support vector machines (SVM) as modelling tools. Different combinations of these feature extraction and modelling techniques are used to get four different systems: MFCC + SDC + GMM, SFCC + SDC + GMM, MFCC + SDC + SVM and SFCC + SDC + SVM. Experiments with these systems show that the language families can be identified with reasonable accuracy. Further, the work tests the influence of one language family on the other and finds that in most cases, the languages which are spoken in areas lying on the boundary of two families are more influenced by the other family. A deviation from it can relate to geopolitical isolation of two neighbouring regions and thus can give new insights or corroborate investigations of historians.

Keywords: Feature extraction, language family, modeling techniques, mutual influence.

A systematic study of language families is essential in order to understand the origin and development of languages. In a large country like India, where the number of languages spoken (official and unofficial) is more than 1500, knowledge about language families is important. The mother language from which different child languages are born is called the proto language of the family. With passage of time and increase in the number of speakers, the mother language splits into several pronunciations and dialects giving rise to different languages. Therefore, language families are a rich source of information to historians. A study about language families tells us how our modern colloquial languages sounded hundred years ago or even earlier. New historical facts may also be revealed as a result of this study.

Most of the Indian languages belong to two major language families, Indo-European and Dravidian. Indo-European is spoken by 73% of Indians and Dravidian by 24% of Indians[1,2]. The remaining 3% of Indians speak languages belonging to several minor families like Austro-Asiatic, Tibeto-Burman, etc. The Indo-European family can be further grouped into a number of sub-families. The largest among these sub-families is Indo-Aryan, which contains the Indian languages. About half of the languages in the Indo-European family belongs to this sub-family[3]. Among the Indian languages, Indo-European corresponds to Assamese, Bengali, Bhojpuri, Chhattisgarhi, Dogri, English, Gujarati, Hindi, Kashmiri, Konkani, Manipuri, Marathi, Nagamese, Odia, Punjabi, Sanskrit, Sindhi and Urdu, and Dravidian corresponds to Kannada, Malayalam, Tamil and Telugu[4].

Language identification (LID) is a popular research area and many research papers are available on the subject. Zissman[5] has discussed LID of telephone speech using four different approaches, namely Gaussian mixture model (GMM), phone recognition language modelling (PRLM), parallel PRLM and parallel phone recognition (PPR). He showed that phone recognizers give a better result than GMM, but availability of phonetically labelled speech is a drawback. An approach using GMM tokenization is discussed by Torres-Carrasquillo *et al.*[6], where system complexity is reduced compared to phone recognizers. Shifted delta coefficient (SDC) as features are used in another study[7], which shows the importance of broader temporal features for LID. Campbell *et al.*[8] have used support vector machines SVM with a generalized linear discriminant kernel (GLDS) for LID. Results are comparable to GMM classifiers. Li *et al.*[9] have discussed about all the tools and techniques of LID available in the literature.

The present work uses some of the popular techniques utilized for language, speaker or speech recognition, for experiments related to identification of language family. For feature extraction, mel frequency cepstral coefficient (MFCC) is considered as a standard tool in speech processing domain[10]. Speech signal-based frequency

*For correspondence. (e-mail: debapriya_20oct@yahoo.co.in)

cepstral coefficient (SFCC) is a newer method and is known to work well in speech recognition. SDC has been successfully used for language identification. GMM is extensively used in speaker or language identification for modelling purpose. SVM is a powerful classification tool, particularly suitable for two-class problems.

The rest of the article is arranged as follows: the next section gives a brief description of the feature extraction and modelling techniques that have been used in the study. Next, the database and experimental set-up are described. The results of the experiments are then presented and discussed, followed by conclusion.

Brief description of the feature extraction and modelling techniques

Feature extraction

We have used MFCC and SFCC as language-dependent features. They are concatenated with SDC to give the final feature vector. Thus the two sets of feature vectors that we have used are MFCC + SDC and SFCC + SDC.

Mel frequency cepstral coefficients: The mel scale is designed to match the human auditory spectrum. The filters in this scale are equally spaced from 0 to 1000 Hz. But their spacing increases continuously beyond 1000 Hz. The mel scale and the 'perpetual' frequency scale are related by the equation:

$$\text{Mel frequency} = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \quad (1)$$

where f is the perpetual frequency.

Figure 1 shows the mel scale filter bank. Speech signal is pre-emphasized, broken into smaller segments of 20 ms, passed through Hamming window and then its power spectral density (PSD) is calculated. The resulting signal is passed through the mel scale filter bank. Logarithm of the output of the mel filter bank is taken and its discrete cosine transform (DCT) is evaluated. These constitute the mel coefficients[11]. We have used frame size of 20 ms with an overlap of 10 ms. The number of filter banks is 20.

Speech signal-based frequency cepstral coefficients: SFCC is a frequency warping technique based purely on the properties of the acoustic speech signal. The filter bank width depends on the data. Train data are divided

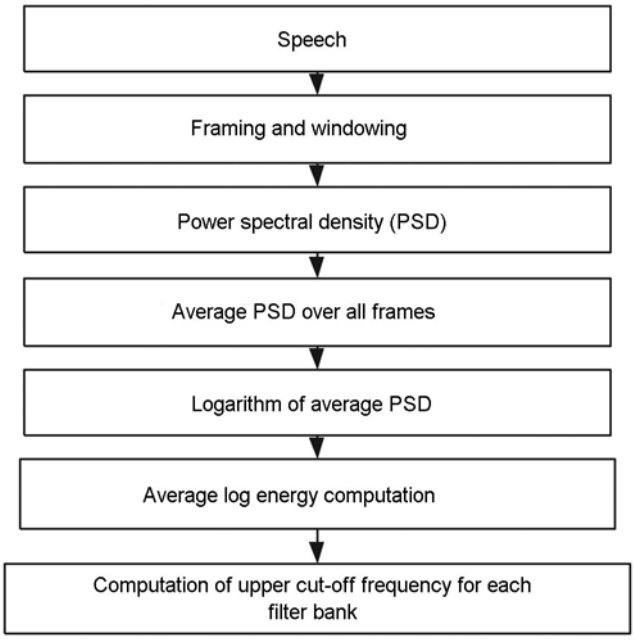

Figure 2. Steps required to compute signal-based frequency cepstral coefficient (SFCC) filter bank.

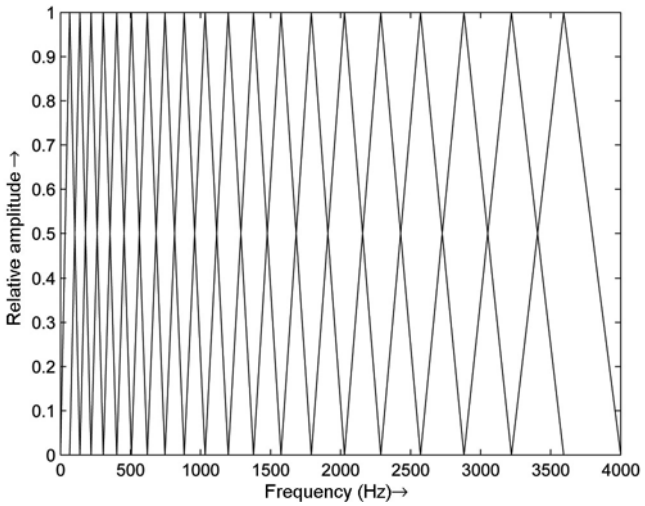

Figure 1. Mel frequency cepstral coefficient filter bank.

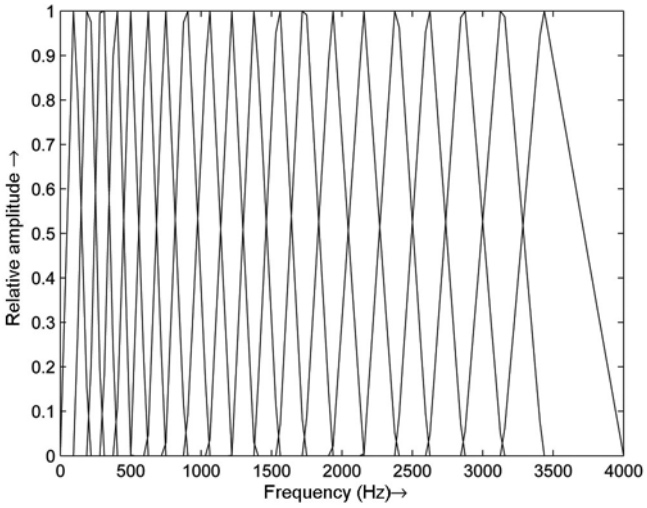

Figure 3. SFCC filter bank.

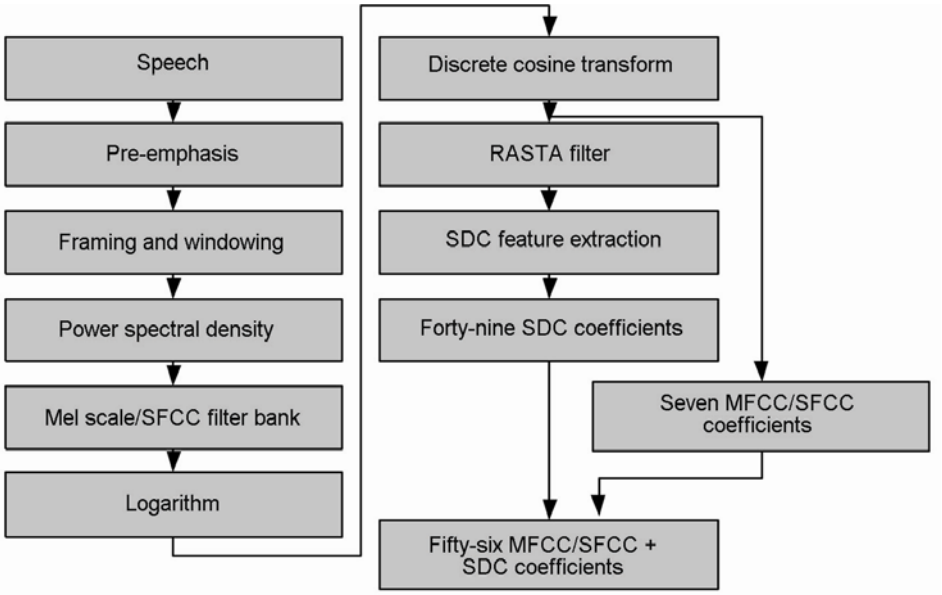

Figure 4. Steps required to derive MFCC/SFCC + shifted delta coefficient (SDC) features of speech.

into frames of 20 ms with 10 ms overlap and passed through Hamming window. PSD of each frame is calculated and averaged over all frames, and its logarithm is computed. Average energy is computed by summing up the log PSD and dividing it by the number of filter banks. We have used 20 filter banks for our experiments. The upper cut-offs for each filter bank are chosen to be such that the log energy of the filter banks is equal to the average energy[12]. The rest of the procedure is the same as that used in case of MFCC. Figure 2 shows the steps involved in computing SFCC filter banks, whereas Figure 3 shows the SFCC filter bank.

Shifted delta coefficients: Delta cepstral features are computed across multiple speech frames and stacked together to give shifted delta cepstral features. Four parameters are used to specify these features: N, d, P and k. N is the number of cepstral coefficients computed at each frame, d is the advance or delay for delta computation, delta features of k number of blocks are concatenated to form the final feature vector and P is the shift in time between two consecutive blocks. SDC coefficients can be represented by the following equation:

$$\Delta c_i(t) = c_i(t + iP + d) - c_i(t + iP - \mathrm{d}), \tag{2}$$

where $c_i(t)$ is the ith block of delta cepstral feature, $i = 0, 1, 2,\ldots, (k - 1)$, and t is the time.

Values of N, d, P and k in this case are 7, 1, 3, and 7 respectively.

We take the first seven MFCC or SFCC coefficients. SDC constitutes 49 coefficients (seven sets of delta coefficients, each set having seven coefficients). So the resulting feature vector has 56 coefficients.

Figure 4 shows the steps required to get MFCC/SFCC + SDC features from speech. SDC is discussed in the literature[7,13,14].

Modelling

For modelling the feature vectors, we have used GMM and SVM.

Gaussian mixture models: In GMM, given the feature vector of train data, the aim is to estimate the parameters of the GMM λ, which best represents the distribution of the feature vectors. A Gaussian mixture density is a weighted sum of M component densities. It can be represented by

$$\mathrm{pr}(\boldsymbol{o}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{o}), \tag{3}$$

where $\boldsymbol{o}$ is the D-dimensional feature vector, p_i, $i = 1, 2,\ldots, M$ are the mixture weights and $b_i(\boldsymbol{o})$, $i = 1, 2,\ldots,M$ are the component densities. Each component density is a Gaussian function of D dimensions of the form

$$b_i(\boldsymbol{o}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu}_i)^t \sum_i^{-1}(\boldsymbol{o} - \boldsymbol{\mu}_i)\right\}, \tag{4}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\sum_i$. We assume diagonal covariance because it gives a better result. The constraint on p_i is $\sum_{i=1}^{M} p_i = 1$. The complete GMM can be represented by $\lambda = \{p_i, \boldsymbol{\mu}_i, \sum_i\}$, $i = 1, 2,\ldots, M$.

The parameters p_i, $\boldsymbol{\mu}_i$ and $\sum_i$, $i = 1, 2, \ldots, M$ are estimated using maximum likelihood method. This is done using expectation maximization (EM) algorithm[15].

The number of Gaussians M in a GMM is called the model order of GMM and is determined experimentally. In our experiments we have used two values of M – 64 and 128. Every experiment is done twice, once for $M = 64$ and then for $M = 128$.

Support vector machines: SVM is a two-class classifier which maps the input space to a high dimensional space and then creates a hyperplane which separates the two classes. The hyperplane should be such that it ensures maximum separation between the two classes (Figure 5). This leads to an optimization problem. Solving this problem using Lagrangian multipliers with suitable optimality conditions[16], we get

$$f(\boldsymbol{x}) = \sum_{i=1}^{I} \alpha_i t_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + \beta, \tag{5}$$

where t_i are the ideal outputs, $\sum_{i=1}^{I} \alpha_i t_i = 0$, the weights $\alpha_i > 0$, β is bias, $\boldsymbol{x}$ is any vector and $\boldsymbol{x}_i$ are the support vectors. Ideally outputs should be either +1 or –1 depending upon which class the corresponding support vectors belong to. $\kappa(\boldsymbol{x}, \boldsymbol{y})$ is the kernel function which is constrained to fulfil certain properties called Mercer condition, so that it can be expressed as

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^t \phi(\boldsymbol{y}), \tag{6}$$

where $\phi(\boldsymbol{x})$ is a mapping from input space to a high-dimensional space. The Mercer condition ensures that the margin concept is valid, and the optimization of the SVM is bounded[8].

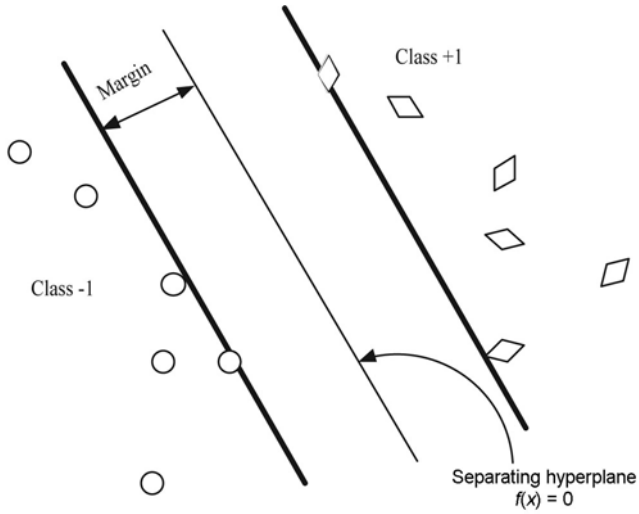

Figure 5. The hyperplane which ensures maximum separation between the two classes.

We have used a GMM supervector linear kernel in our experiments. The means of GMM are stacked to form a GMM mean supervector[17] (Figure 6). The supervectors can be thought of as a mapping between an utterance and a high-dimensional vector. We use maximum-a-posteriori (MAP) adaptation to compute the means of GMM. For this, a universal background model (UBM) is developed using the entire train data[18]. UBM is a large GMM trained to represent the language-independent distribution of features. For a good model it is required to be trained with languages, tones and speakers of all variety which the model is expected to encounter during classification.

The linear kernel used in our experiments is represented by

$$\kappa(\varphi_a, \varphi_b) = \sum_{i=1}^{M} p_i (\boldsymbol{\mu}_i^a)^t \sum_i^{-1} \boldsymbol{\mu}_i^b$$

$$= \sum_{i=1}^{M} (\sqrt{p_i}\, \Sigma_i^{-1/2} \boldsymbol{\mu}_i^a)^t (\sqrt{p_i}\, \Sigma_i^{-1/2} \boldsymbol{\mu}_i^b), \tag{7}$$

where φ_a and φ_b are two utterances under consideration and $\boldsymbol{\mu}^a$ and $\boldsymbol{\mu}^b$ are the adapted supervector of means. Detailed description of this method is given by Campbell *et al.*[17].

Two feature extraction and two modelling techniques are described in this section. With these four techniques we get four systems, namely MFCC + SDC + GMM, SFCC + SDC + GMM, MFCC + SDC + SVM and SFCC + SDC + SVM. Our experiments are performed with all these four systems in order to ensure that the results are consistent irrespective of the system used.

Experimental set-up

Database preparation

The database used is prepared from the All India Radio website[19]. Daily news bulletins of all major Indian languages are available in its repository.

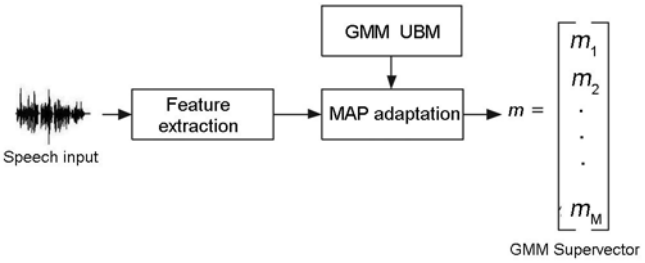

Figure 6. Formation of Gaussian mixture model (GMM) supervectors.

There are various reasons for which we have selected this website as our data source. The quality of speech is good as it contains very little noise. Speech is sufficiently loud and is of sufficiently long duration. A large number of speakers (news readers) are available in each language. This reduces speaker bias. Both male and female speakers are present in almost all the available languages, ensuring little or no gender bias. Speech covers a large variety of topics that helps capture details of acoustic information.

To prepare the database, we have downloaded all news files that are available in the 22 languages mentioned earlier (18 Indo-European and 4 Dravidian languages). After listening to each of the news files, we have rejected those of poor quality. Almost all news bulletins have a short music of 10 sec at the beginning and the end. These have been removed in our experimental database. The speech is originally in .mp3 format. It is converted to .wav format to enable further processing.

Language family identification

We train the systems to identify speech from each of the two families (Indo-European and Dravidian). Then we test them by giving speech samples from both the families as input and find the percentage of correct identification.

Training: For each family, 4 h of speech is taken for training. This 4 h contains speech from all the languages belonging to the families, i.e. Indo-European training set contains news files from all the 18 languages in equal proportion. Since each news file is nearly of 10 min duration, we have taken two files from each language on an average. Similarly, since Dravidian family has four languages, eight files from each language have been taken on an average. This is done so that the systems do not get biased towards any particular language. Also, duration of male and female speech is nearly equal in each family to reduce gender bias. After the train data are arranged in this format, each speech file is broken into segments of 30 sec duration. Each segment is listened to and segments containing music, long duration of silence or unwanted sounds are deleted. At the end, we have two sets of speech files, one for each family, each having 4 h of speech segmented into 30 sec duration. This is the training set.

Testing: This is done for three different utterance durations, short (3 sec), medium (10 sec) and long (30 sec). For each utterance duration, we have 100 test utterances. The train and test speaker sets are mutually exclusive. News files from test speakers are chosen taking care of the fact that the number of male and female-speaking files is nearly equal. These news files are then broken into 3 sec, 10 sec and 30 sec segments. Music, long duration of silence and unwanted voices are deleted as is done for training speech. From these segments, we select 100 utterances for each family and for each duration. These come from all the languages belonging to the particular family. For example, since Dravidian family has 4 languages, there are 25 utterances from each language and Indo-European family has nearly 6 utterances from each language. These 25 or 6 utterances from each language are equally divided between male and female speakers (where exact equal division is not possible, nearly equal division is taken). This is how one test set is prepared. We have prepared three such test sets to verify our results. These three test sets are again mutually exclusive, i.e. the 100 utterances contained in set A are different from the 100 utterances in sets B or C. Figure 7 shows the structure of train and test sets.

Influence of Dravidian family on Indo-European languages

Two languages spoken in the neighbouring regions may be influenced by each other. We have tried to test this experimentally. Also, the extent of influence can be indirectly measured from the accuracy of correct identification.

Since languages belonging to the same family sound similar (have similar acoustic features), it is expected that a system trained with a language family can identify whether a language belongs to that family or not even if that language is not used during training of the system. Based on this idea, we trained the systems by leaving out one of the languages of Indo-European family. Then when we test these systems with the language left out, languages having more Dravidian influence are expected to give less identification accuracy compared to those having less Dravidian influence.

Influence on neighbouring Indo-European languages: Indo-European languages are spoken by most of Central and North Indian states. Dravidian languages are spoken in the southern states of Karnataka (Kannada), Kerala (Malayalam), Tamil Nadu (Tamil) and Andhra Pradesh (Telugu). The neighbouring Indo-European speaking states are Chhattisgarh, Goa, Maharashtra and Odisha speaking Chhattisgarhi, Konkani, Marathi and Odia respectively. The neighbouring languages are expected to have more Dravidian influence than non-neighbouring languages. We tested the influence of Dravidian languages on each of the neighbouring languages. Figure 8 shows the Dravidian speaking states and their neighbouring Indo-European speaking states.

Training: In order to test the influence of Dravidian family on neighbouring Indo-European languages, we have prepared a train dataset which is similar to that

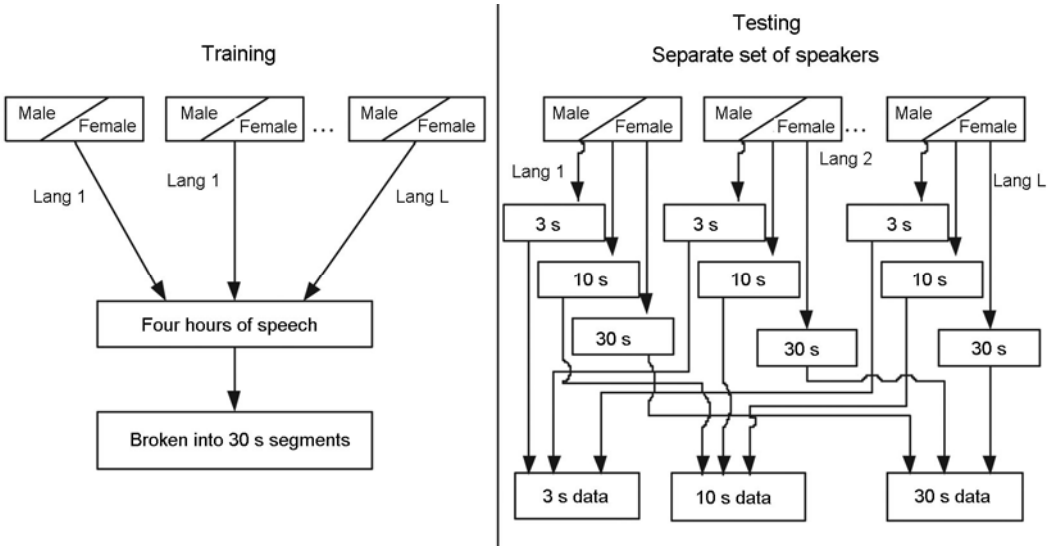

Figure 7. Train and test dataset.

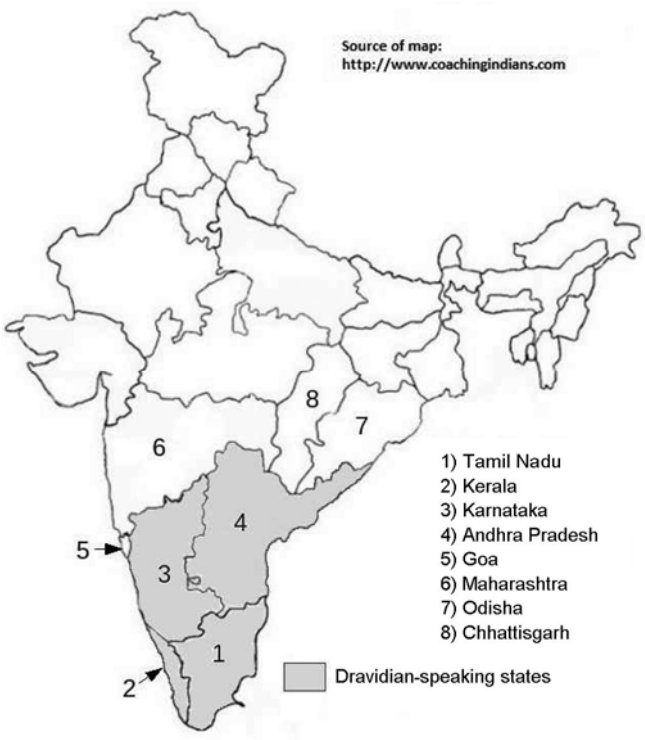

Figure 8. Dravidian-speaking states and their neighbours.

described in the section on 'training'; but with a little change. As in the earlier case, we take 4 h of speech for each language family. But here the Indo-European dataset contains 17 languages, unlike 18 in the previous case. We leave out one neighbouring Indo-European language each time. For example, to test the influence of Dravidian on Chhattisgarhi, we train the system with all 17 Indo-European languages, except Chhattisgarhi. So, Chhattisgarhi news files are removed from the training set of the earlier section and replaced by news of any other Indo-European language. Care is taken so that total duration and male–female proportion of the dataset are disturbed as little as possible. It is not necessary that the replaced speech should be from one language. For example, Chhattisgarhi speech can be replaced by one male speech in Hindi and one female speech in Gujarati. The Dravidian dataset remains unchanged. In this way, the systems are trained four times, each time leaving out one neighbouring Indo-European language.

Testing: The test set contains speech of only the left out language. We select some news files from the left out language keeping nearly equal proportion of male and female speakers. These files are broken into segments of 3 sec, 10 sec and 30 sec duration. We select 100 segments from each duration to prepare the test dataset. Similar procedure is carried out for all the four neighbouring Indo-European languages. The only exception is Chhattisgarhi, where 100 test utterances could not be collected for 30 sec utterance duration due to shortage of data. So tests are done with 86 segments.

Influence on non-neighbouring Indo-European languages: In order to compare the results of Dravidian influence on neighbouring Indo-European languages, we have done similar experiments on non-neighbouring Indo-European languages as well. For this, we have randomly selected four Indo-European languages which do not have a Dravidian-speaking neighbour. The languages we selected are Bengali, Gujarati, Hindi and Punjabi.

Training: The data are prepared in the same way as described earlier. Each system is trained four times, each time leaving out one of the four selected Indo-European languages and compensating it with other Indo-European languages of equal duration and equal male–female proportion.

Table 1. Language family identification accuracy of three test datasets using MFCC + SDC + GMM and SFCC + SDC + GMM

	MFCC + SDC + GMM						SFCC + SDC + GMM					
	Set A		Set B		Set C		Set A		Set B		Set C	
Test duration (sec)	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	74	74.5	80	79	76.5	78.5	75.5	77	78	76.5	77.5	78.5
10	81.5	84.5	83	82.5	83.5	85	84.5	82.5	85	83.5	86	85.5
30	88.5	88.5	87	87	89.5	89.5	90.5	90	88.5	87.5	91	88.5

MFCC, Mel frequency cepstral coefficient; SDS, Shifted delta coefficient; GMM, Gaussian mixture model; SFCC, Signal-based frequency cepstral coefficient; M.O., Model order.

Table 2. Language family identification accuracy of three test datasets using MFCC + SDC + SVM and SFCC + SDC + SVM

	MFCC + SDC + SVM						SFCC + SDC + SVM					
	Set A		Set B		Set C		Set A		Set B		Set C	
Test duration (sec)	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	74	75.5	78	78.5	77	78	77	77	78	80	78.5	80.5
10	86.5	88	84	87	84.5	88	90	89.5	87.5	89.5	86	87
30	90	88.5	92	92.5	94	93.5	93.5	92.5	91.5	93.5	95	97

SVM, Support vector machines.

Testing: Data preparation is similar to that mentioned earlier. Four sets of test data are prepared, one each for Bengali, Gujarati, Hindi and Punjabi, each set having segments of 3 sec, 10 sec and 30 sec duration and each duration having 100 test utterances.

Influence of Indo-European family on Dravidian languages

We have tested the influence of Dravidian family on neighbouring and non-neighbouring Indo-European languages. Similarly, we now test the influence of Indo-European family on neighbouring and non-neighbouring Dravidian languages. Figure 8 shows that the Dravidian-speaking states lying on the boundary of Indo-European-speaking states are Karnataka and Andhra Pradesh. The remaining two Dravidian-speaking states, Kerala and Tamil Nadu, which do not have a Indo-European speaking boundary, can be considered as non-neighbouring states.

Influence on neighbouring Dravidian languages: Here we test the influence of Indo-European family on Kannada and Telugu.

Training: Data are prepared as mentioned earlier. Two sets of data are prepared, leaving out Kannada in the first and Telugu in the next. This is compensated by speech from other Dravidian languages. This time only Dravidian dataset is modified and Indo-European data are kept untouched.

Testing: The process is similar to that mentioned earlier. Two sets of test data are prepared, one having Kannada speech segments and the other Telugu. The structure of the datasets is the same as described earlier.

Influence on non-neighbouring Dravidian languages: Here influence of Indo-European family on Malayalam and Tamil is tested.

Training: This process is the same as the earlier processes. Two sets of train data are prepared leaving out Malayalam in one and Tamil in the other. This is compensated by speech from other Dravidian languages.

Testing: The process of data preparation and structure of test datasets are similar to the above-mentioned cases. Test data are prepared with Malayalam and Tamil speech segments.

Results and discussion

The results have been divided into three sub-parts, namely results showing language family identification accuracy,

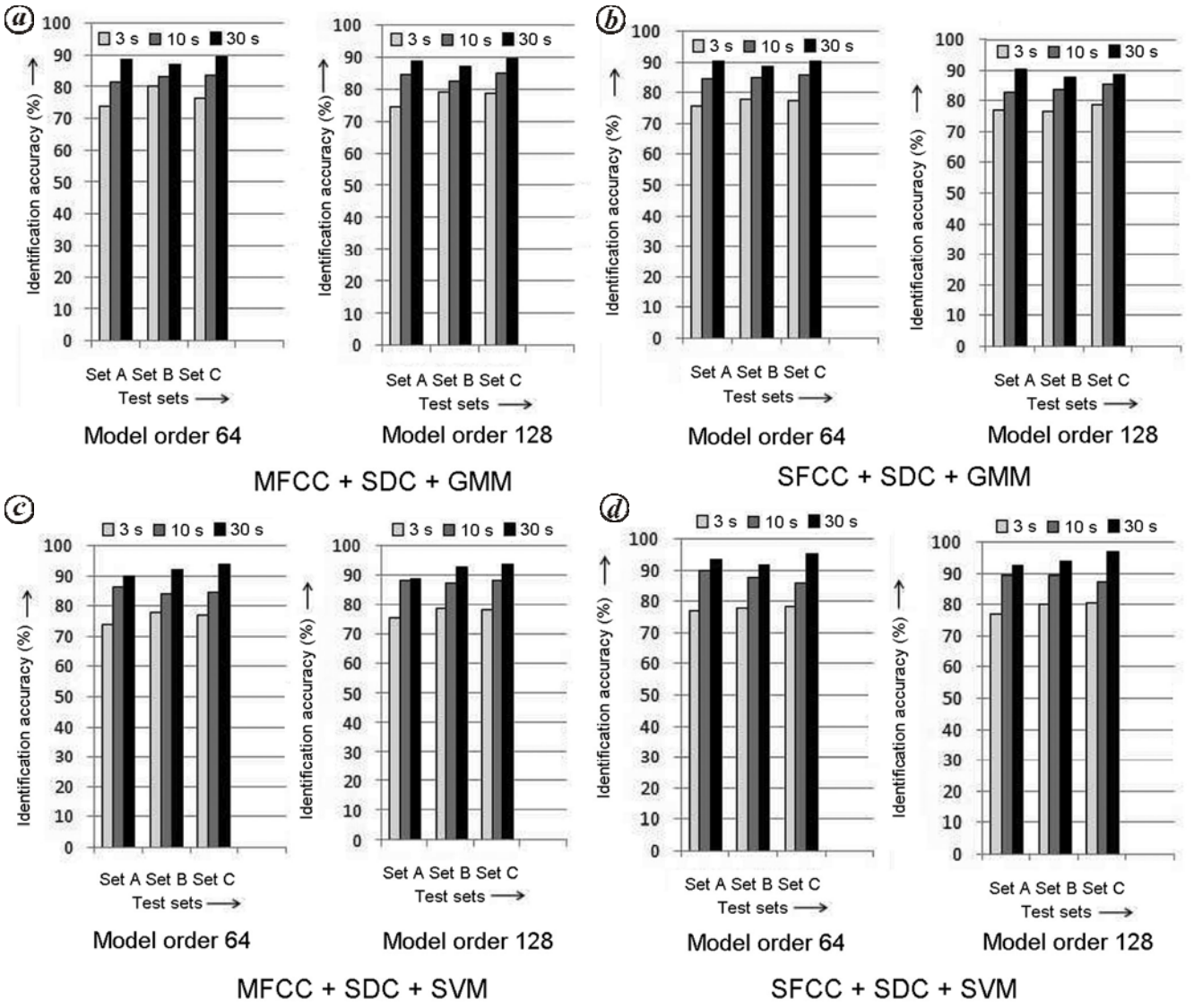

Figure 9 ***a–d.*** Graphical representation of language family identification accuracy of three sets of test data using the four systems.

influence of Dravidian family on Indo-European languages, and influence of Indo-European family on Dravidian languages.

Language family identification

Tables 1 and 2 show the results of language family identification tests using all the four systems, for three sets of test data. The experiments are done using GMM model order 64 and 128. Figure 9 is a graphical representation of the results.

All the four systems can identify the language families with high rate of accuracy. The accuracy ranges from 74% to 97%. Accuracy percentage increases with test utterance duration, i.e. 30 sec utterances give higher accuracy than 10 sec utterances which again give higher accuracy than 3 sec utterances. This is justified because higher duration implies more test data. Results are consistent across all the systems. No system gives abrupt high or low accuracy. Model order 64 and 128 do not show much variation in the results.

Influence of Dravidian family on Indo-European languages

Tables 3–6 show identification accuracy of Indo-European languages which are neighbouring and non-neighbouring to Dravidian-speaking states using the four systems respectively. Figures 10 and 11 show a graphical representation of the same.

As expected, the neighbouring Indo-European languages give less accuracy than the non-neighbouring languages which implies that neighbouring languages have higher Dravidian influence. Also, among the

Table 3. Identification accuracy of Indo-European languages which are neighbouring to Dravidian family, using MFCC + SDC + GMM and SFCC + SDC + GMM

	MFCC + SDC + GMM								SFCC + SDC + GMM							
	Chhattisgarhi		Konkani		Marathi		Odia		Chhattisgarhi		Konkani		Marathi		Odia	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	42	41	69	72	75	79	32	32	49	41	75	78	78	74	26	28
10	41	37	69	76	90	90	34	39	52	43	78	78	89	87	35	38
30	47.6744	36.0465	63	73	94	95	25	29	48.8372	43.0233	67	69	92	93	23	25

Table 4. Identification accuracy of Indo-European languages which are neighbouring to Dravidian Family, using MFCC + SDC + SVM and SFCC + SDC + SVM

	MFCC + SDC + SVM								SFCC + SDC + SVM							
	Chhattisgarhi		Konkani		Marathi		Odia		Chhattisgarhi		Konkani		Marathi		Odia	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	38	28	68	67	81	80	61	64	32	27	62	65	80	77	58	51
10	38	37	72	80	91	88	65	73	39	40	80	84	89	90	55	64
30	40.6977	48.8372	63	72	95	95	75	77	45.3488	45.3488	69	77	95	97	60	64

Table 5. Identification accuracy of Indo-European languages which are non-neighbouring to Dravidian family, using MFCC + SDC + GMM and SFCC + SDC + GMM

	MFCC + SDC + GMM								SFCC + SDC + GMM							
	Bengali		Gujarati		Hindi		Punjabi		Bengali		Gujarati		Hindi		Punjabi	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	89	90	59	63	94	91	81	79	88	87	63	64	90	95	76	74
10	95	95	74	72	98	99	89	88	94	94	77	75	99	100	79	83
30	99	100	64	64	100	100	93	91	100	100	62	64	100	100	75	71

Table 6. Identification accuracy of Indo-European languages which are non-neighbouring to Dravidian family, using MFCC + SDC + SVM and SFCC + SDC + SVM

	MFCC + SDC + SVM								SFCC + SDC + SVM							
	Bengali		Gujarati		Hindi		Punjabi		Bengali		Gujarati		Hindi		Punjabi	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	76	72	79	77	83	90	74	77	76	79	73	70	83	83	73	65
10	90	91	91	90	94	97	89	88	87	91	87	87	98	97	84	82
30	98	95	90	92	98	99	98	94	93	95	84	81	99	99	89	83

neighbouring languages, Chhattisgarhi and Odia show higher Dravidian influence than Marathi and Konkani. This should be due to historical reasons.

Odisha is the modern name of Kalinga. Kalinga comprised of most parts of modern-day Odisha and the Andhra region of Andhra Pradesh (Dravidian-speaking state)[20]. Odisha has been ruled by several rulers from ancient times. Its boundaries have been reformed and reshaped from time to time. During the reign of rulers like Anantavarma Chodagangadeva of the Ganga dynasty, boundaries of Odisha extended from River Ganga in the north to River Godavari in the south. Godavari is in the

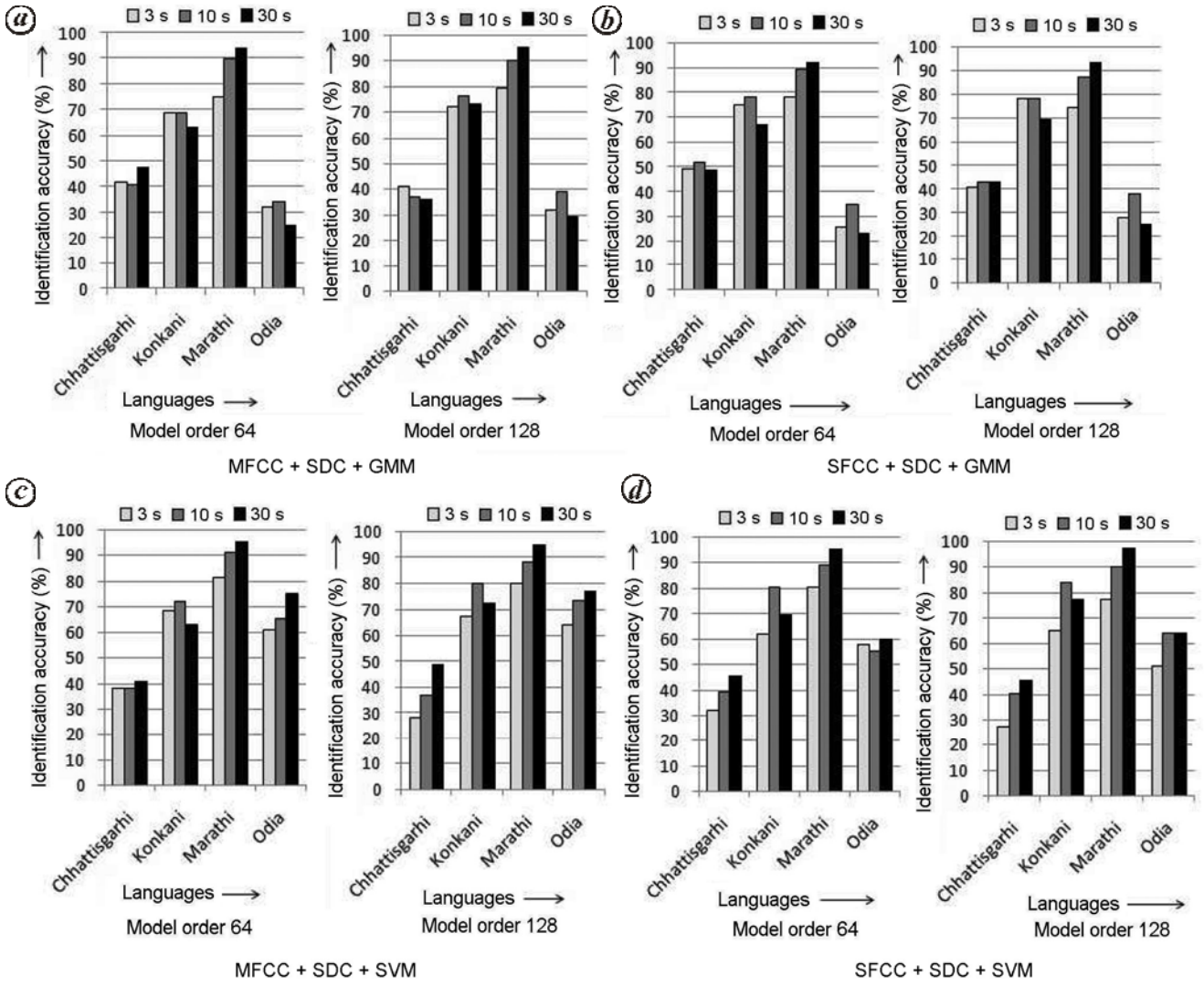

Figure 10 ***a–d.*** Graphical representation of identification accuracy of Indo-European languages which are neighbouring to Dravidian family, using the four systems.

present-day Andhra Pradesh. Also, the first king of the Surya dynasty, Gajapati Kapilendradeva, extended his empire from River Ganga to River Kaveri which includes regions of present-day Tamil Nadu (Dravidian-speaking state). All these might have resulted in the intermingling of Oriya with the Dravidian languages[21].

In Chhattisgarh, the Chalukya dynasty established its rule during the middle ages[22]. This dynasty ruled parts of southern and Central India covering modern-day Karnataka, Andhra Pradesh and parts of Maharashtra (Karnataka and Andhra Pradesh are Dravidian-speaking states). This can be a possible reason behind high Dravidian influence in Chhattisgarhi language.

On the other hand, Konkani, though spoken in a neighbouring state (Goa), has little Dravidian influence. There could be several reasons for this. The most important reason possibly is the Portuguese colonization over Goa shortly after Vasco da Gama entered into India[23], and gradually became the centre of Portuguese India[24]. Portugal is an Indo-European language[3]. Even prior to Portuguese control, Goa had trade connections with foreign land from early times. Influence of all these foreign languages must have been more than the Dravidian influence on Konkani. Earlier during the Satavahana dynasty, Konkani language was highly influenced by Maharashtri Prakrit, which was the administrative language of the period[25]. Prakrit also belongs to the Indo-European group of languages. Goa had a large Persian and Greek Buddhist population. During the Kadamba rule, the port of Goapakapattna in Goa became a centre for trade. It had trade contacts with several Indian states and foreign countries.

Maharashtra was ruled by the Mauryas, the Satavahanas, the Rashtrakutas, the Chalukyas and several other Indian dynasties. But all these do not seem to have influenced the language of the region much. In Maharashtra,

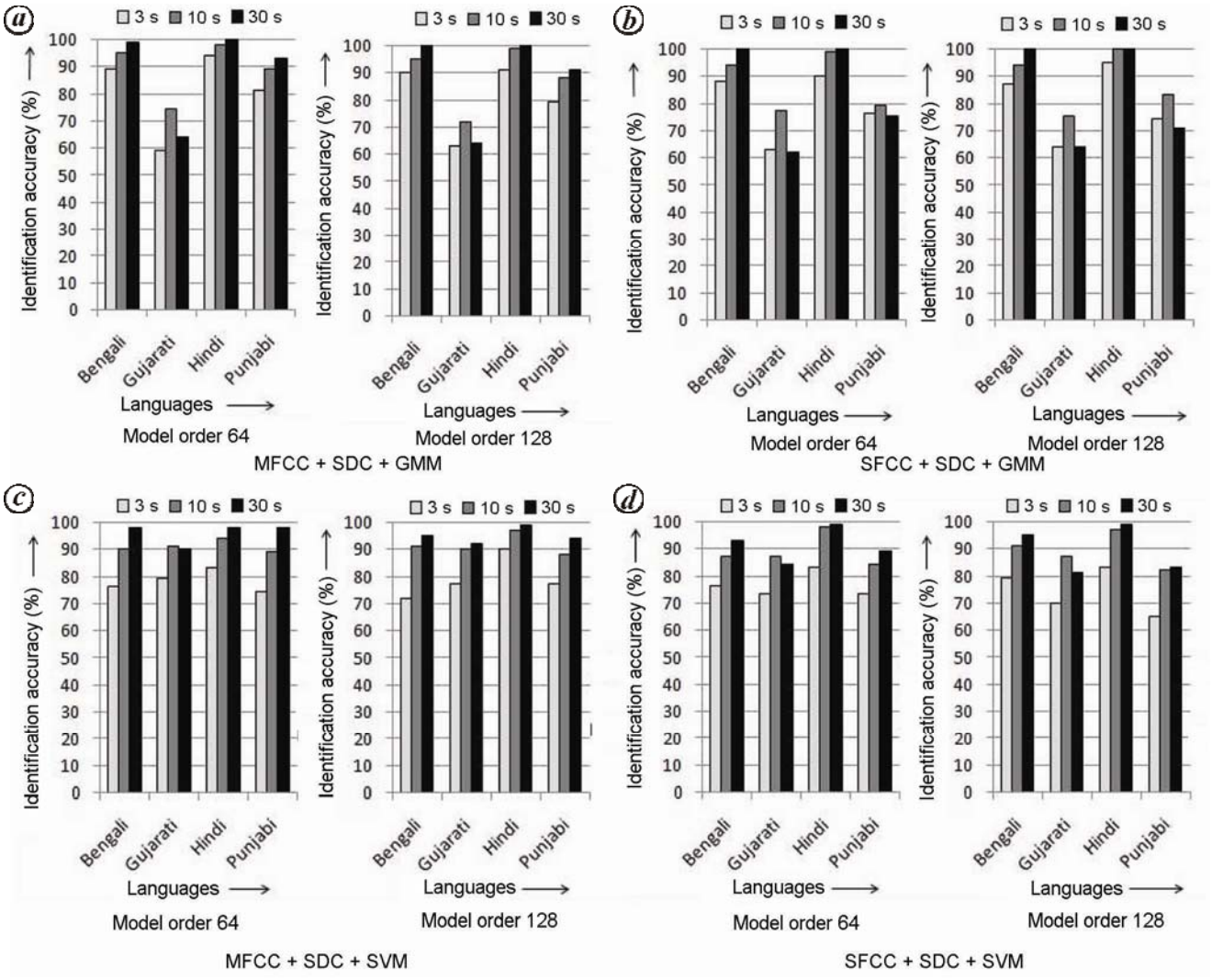

Figure 11 ***a–d.*** Graphical representation of identification accuracy of Indo-European languages which are non-neighbouring to Dravidian family, using the four systems.

the dominant community is Maratha, which is a result of Aryan penetration from north and northeast[26]. The language has maintained its originality, though small variations have taken place with time.

A look at the four systems shows that Odia accuracy increases when SVM is used. The rest of the languages show consistent results with all the systems.

In case of non-neighbouring languages, Gujarati results are comparatively lower using GMM. This is overcome using SVM. For the rest of the languages, overall accuracy does not show much variation from system to system. Accuracy is much higher than the neighbouring languages (in the range 62–100%).

Influence of Indo-European family on Dravidian languages

Tables 7 and 8 show the identification accuracy of Dravidian languages which are neighbouring and non-neighbouring to Indo-European speaking states, using all the four systems respectively. Figures 12 and 13 show a graphical representation of the same.

The results suggest that neighbouring and non-neighbouring classification is not suitable for these languages because there are only four languages and Malayalam, though a non-neighbouring language, is expected to have Indo-European influence as suggested by historical evidence and verified by experimental results. Kerala had been in contact with foreign land since the 15th century, when Vasco da Gama arrived in present-day Kozhikode in 1498 in order to trade spices[27–29]. Gradually, the Portuguese defeated the local rulers and started ruling over Kerala. It is to be noted that similar Portuguese colonization also happened in Goa, which resulted in low Dravidian influence (better identification accuracy of Konkani). After the Portuguese, Kerala came under Dutch rule. Finally by the end of 18th century, the whole of Kerala came under British control.

Table 7. Identification accuracy of Dravidian languages which are neighbouring to Indo-European family, using the four systems

	MFCC + SDC + GMM				SFCC + SDC + GMM				MFCC + SDC + SVM				SFCC + SDC + SVM			
	Kannada		Telugu		Kannada		Telugu		Kannada		Telugu		Kannada		Telugu	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	40	44	42	59	54	51	53	51	77	69	54	58	78	77	49	55
10	43	47	52	56	55	47	54	56	83	76	42	42	88	84	45	41
30	42	50	54	57	64	57	52	52	85	83	42	38	90	88	39	34

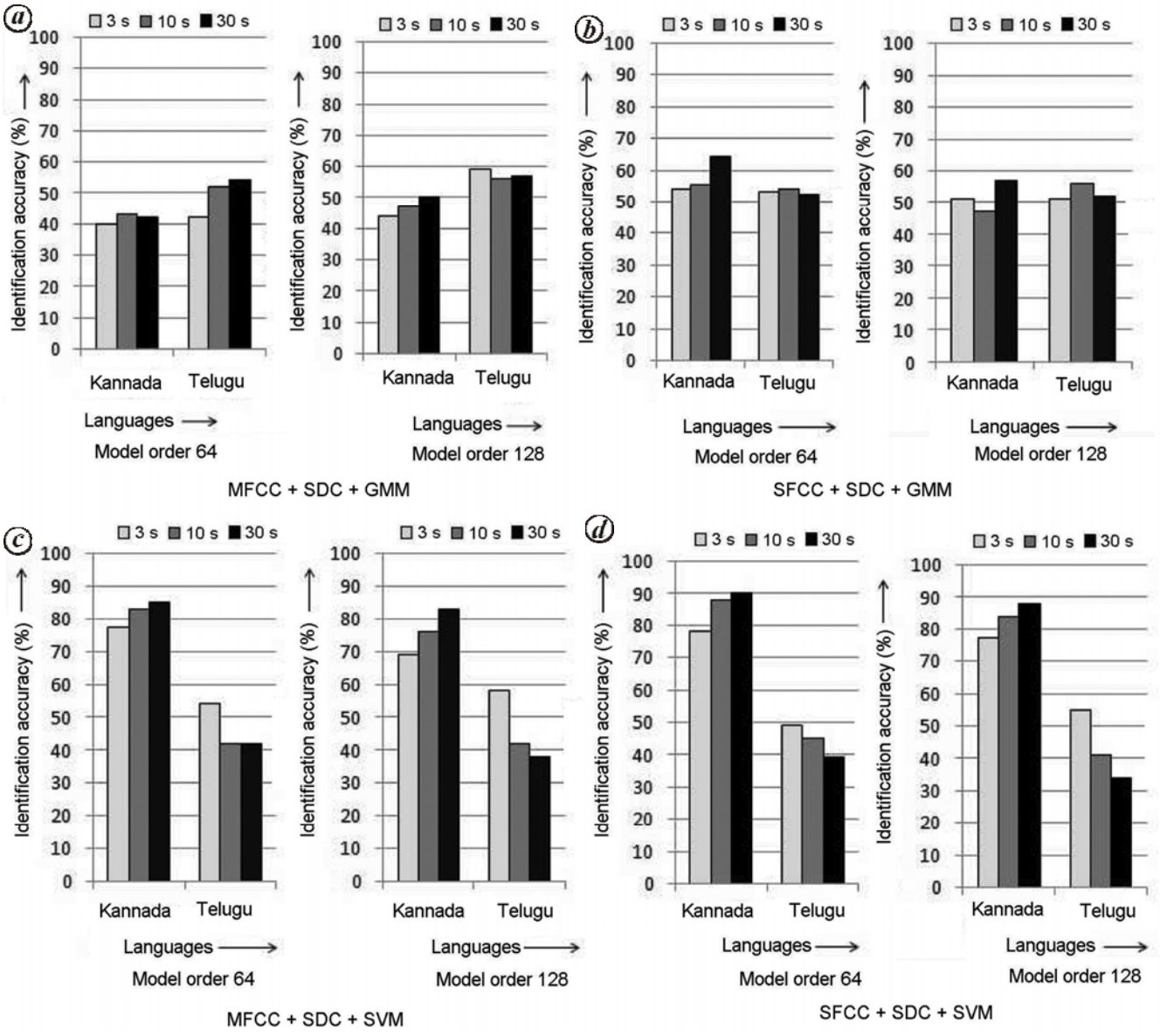

Figure 12*a–d*. Graphical representation of identification accuracy of Dravidian languages which are neighbouring to Indo-European family, using the four systems.

Unlike Kerala, Tamil Nadu was ruled mostly by Indian rulers like the Pallavas, the Rashtrakutas, the Cholas and the Pandyas. So colonial influence is not expected in Tamil. The Tamil results show better accuracy than Malayalam, which is in accordance with historical evidence.

Telugu is the only Dravidian language which consistently gives low accuracy. This is because Andhra Pradesh

Table 8. Identification accuracy of Dravidian languages which are non-neighbouring to Indo-European family, using the four systems

	MFCC + SDC + GMM				SFCC + SDC + GMM				MFCC + SDC + SVM				SFCC + SDC + SVM			
	Malayalam		Tamil		Malayalam		Tamil		Malayalam		Tamil		Malayalam		Tamil	
Test duration (sec)	M.O. 64	M.O. 128	M.O 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128	M.O. 64	M.O. 128
3	65	65	73	76	58	61	67	73	80	81	77	80	70	79	79	79
10	59	63	79	82	69	60	76	85	77	72	79	82	74	71	81	81
30	66	74	87	95	73	63	80	86	76	76	77	86	76	71	83	83

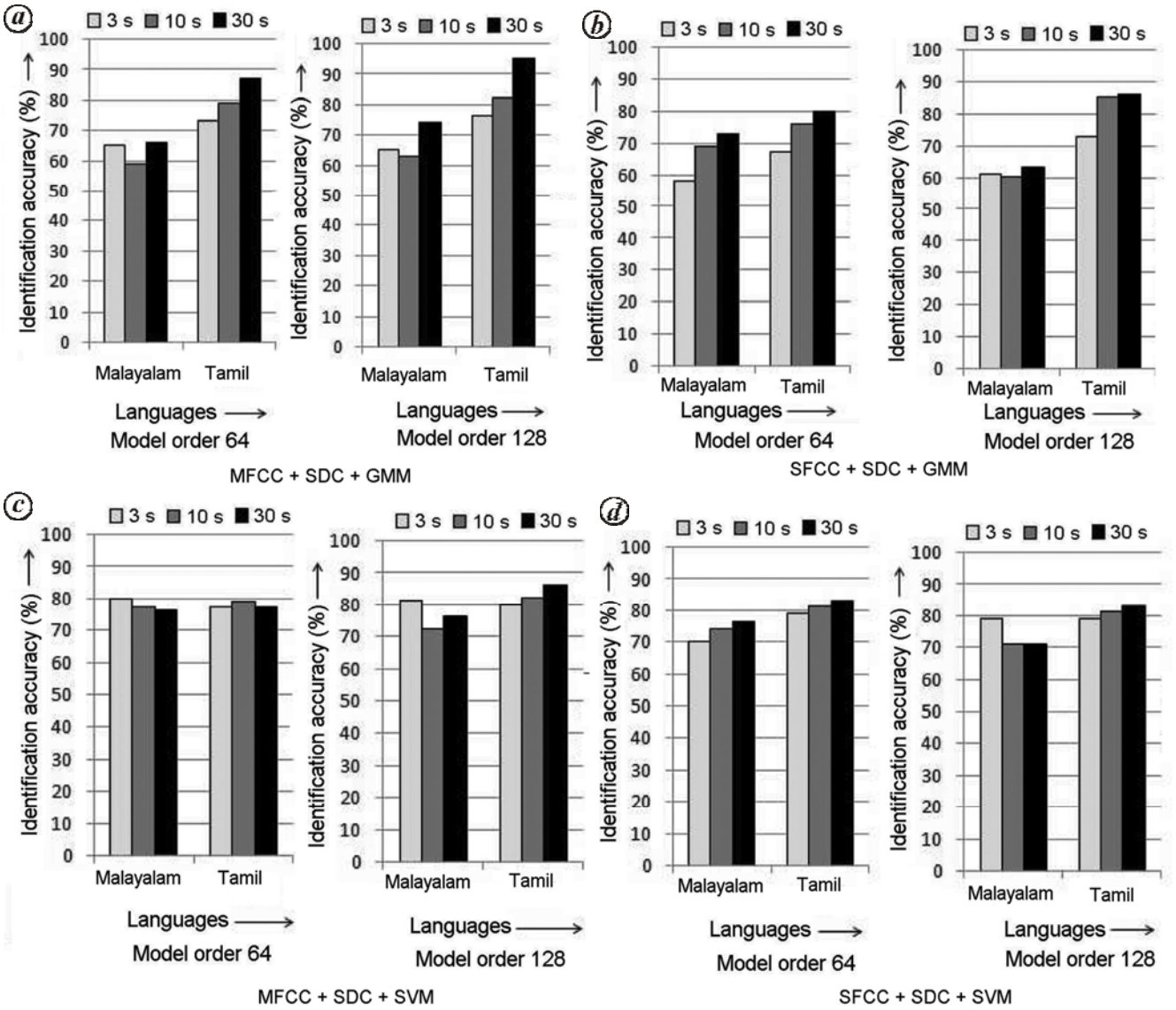

Figure 13*a–d*. Graphical representation of identification accuracy of Dravidian languages which are non-neighbouring to Indo-European family, using the four systems.

was under Mughal rule from 14th to 18th century. During this time, Telugu was highly influenced by Urdu. In the latter half of the 17th century, Mughal rule extended further south culminating in the establishment of the princely state of Hyderabad. This heralded an era of Persian/Arabic influence on Telugu[30]. Figure 14 shows that Telugu has descended from the Proto-South-Central-Dravidian branch, whereas the other three Dravidian

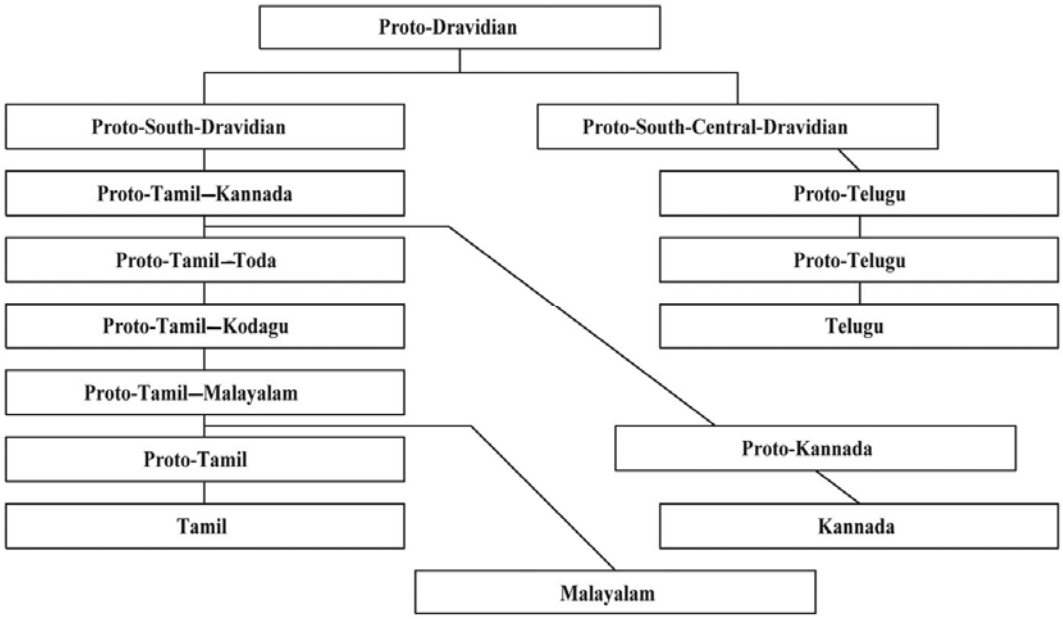

Figure 14. The Dravidian family tree.

languages have descended from Proto-South-Dravidian branch[3]. This separates out Telugu from the other three Dravidian languages.

A look at all the systems indicates that Kannada results show significant improvement when SVM is used. For the rest of the languages, results do not show much variation from system to system.

Analysis

There are certain notable facts in the results. Marathi, in spite of being a neighbouring Indo-European language, shows high accuracy indicating low Dravidian influence. This can be possibly due to non-porosity among the border states. Chhattisgarhi, Odia and Telugu give consistent low results. A look at the geographical location of the three states speaking these languages shows that they are neighbouring states lying on the eastern region of India. There should be some specific historical reason behind this high intermingling of languages in these regions apart from the ones stated above, which is yet to be found out.

Overall accuracy of non-neighbouring Indo-European languages is more than overall accuracy of non-neighbouring Dravidian languages. One reason for this could be the geographical extent of the Indo-European-speaking states. For example, Punjabi or Bengali (even Hindi in a few cases) is spoken in states which are far away from the Dravidian states. So it is unlikely that Dravidian influence would reach these states. Whereas even the farthest Dravidian-speaking states are geographically more closer to the Indo-European-speaking states.

Conclusion

This study uses machine learning for automatic recognition of the major Indian language families. MFCC and SFCC in conjunction with SDC have been used as feature extraction tools. For modelling of feature vectors, GMM and SVM are used. It is found that all the four systems can identify the language families with high accuracy. We have evaluated the influence of one language family on the other. We see that in most of the cases the neighbouring languages are influenced more by the other family. Also, among the neighbouring languages, some show higher or lower influence than others. This can be linked to certain known historical facts. The work opens new scope of study which will enable us to know our history better.

1. Ishtiaq, M., *Language Shifts Among the Scheduled Tribes in India: A Geographical Study*, Motilal Banarsidass Publ., 1999.
2. *CIA World Factbook*; https://www.cia.gov.
3. *Ethnologue: Languages of the World*; http://www.ethnologue.com
4. Encyclopedia Britannica; http://www.britannica.com
5. Zissman, M. A., Automatic language identification of telephone speech. *Lincoln Lab. J.*, 1995, **8**(2), 115–144.
6. Torres-Carrasquillo, P. A., Reynolds, D. A. and Deller Jr, J. R., Language identification using Gaussian mixture model tokenization. In International Conference Spoken Language Processing, Denver, Colorado, United States, September 2002.
7. Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A. and Deller Jr, J. R.. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. International Conference Spoken Language Processing, Denver, Colorado, United States, 2002.
8. Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A., Support vector machines for speaker

and language recognition. Computer Speech and Language, Elsevier, 2006, pp. 210–229.
9. Li, H., Ma, B. and Lee, K. A., Spoken language recognition: from fundamentals to practice. *Spoken Language Recogn., Proc. IEEE*, December 2012.
10. Davis, S. B. and Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech Signal Process.*, 1980, **28**(4), 357–366.
11. Vergin, R., Shaughnessy, D. O. and Farhat, A., Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Trans. Speech Audio Process.*, 1999, **7**(5), 525–532.
12. Paliwal, K., Shannon, B., Lyons, J. and Wojcicki, K., Speech-signal-based frequency warping. *IEEE Signal Process. Lett.*, 2009, **16**(4), 319–322.
13. Matejka, P., Burget, L., Schwarz, P. and Cernocky, J., Brno University of Technology System for NIST 2005 Language Recognition Evaluation. In IEEE Odyssey – The Speaker and Language Recognition Workshop, San Juan, Puerto Rico, 28–30 June 2006.
14. Kohler, M. A. and Kennedy, M., Language Identification Using Shifted Delta Cepstra, In Circuits and Systems Conference, IEEE, 4–7 August 2002.
15. Reynolds, D. A. and Rose, R. C., Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, 1995, **3**(1), 72–83.
16. Haykin, S., *Neural Networks and Learning Machines*, Pearson Education Inc., 2011.
17. Campbell, W. M., Sturim, D. E. and Reynolds, D. A., Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.*, 2006, **13**(5), 308–311.
18. Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., Speaker verification using adapted Gaussian mixture models. *Digital Signal Process.*, 2000, **10**(1–3), 19–41.
19. Prasar Bharati; http://newsonair.nic.in
20. Majumdar, R. C., Raychaudhuri, H. C. and Datta, K., *An Advanced History of India*, Macmillan, 1946.
21. http://orissa.gov.in
22. http://chhattisgarh.nic.in
23. Diffie, B. W. and Winius, G. D., *Foundations of the Portuguese Empire, 1415–1850*, University of Minnesota Press, Minnesota Archive Editions, 1977.
24. Shastry, B. S. and Borges, C. J., *Goa–Kanara Portuguese Relations, 1498–1763*, The Xavier Centre of Historical Research, 2000.
25. http://www.goakonkaniakademi.org
26. http://www.mu.ac.in
27. Ravindran, P. N.. *Black Pepper: Piper Nigrum*, CRC Press, 2004.
28. Curtin, P. D., *Cross-Cultural Trade in World History*, Cambridge University Press, 1984.
29. Mathias Mundadan, A., *From the Beginning up to the Middle of the Sixteenth Century (up to 1542) (History of Christianity in India)*, Church History Association of India, 1989.
30. http://www.aponline.gov.in

Received 8 April 2014; accepted 30 August 2015

doi: 10.18520/cs/v110/i4/667-681