Research Article

Study on Similarity among Indian Languages Using Language Verification Framework

Debapriya Sengupta and Goutam Saha

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721302, India

Correspondence should be addressed to Debapriya Sengupta; debapriya_20oct@yahoo.co.in

Majority of Indian languages have originated from two language families, namely, Indo-European and Dravidian. Therefore, certain kind of similarity among languages of a particular family can be expected to exist. Also, languages spoken in neighboring regions show certain similarity since there happens to be a lot of intermingling between population of neighboring regions. This paper develops a technique to measure similarity among Indian languages in a novel way, using language verification framework. Four verification systems are designed for each language. Acceptance of one language as another, which relates to false acceptance in language verification framework, is used as a measure of similarity. If language *A* shows false acceptance more than a predefined threshold with language *B*, in at least three out of the four systems, then languages *A* and *B* are considered to be similar in this work. It is expected that the languages belonging to the same family should manifest their similarity in experimental results. Also, similarity between neighboring languages should be detected through experiments. Any deviation from such fact should be due to specific linguistic or historical reasons. This work analyzes any such scenario.

1. Introduction

In India, the number of languages spoken is more than 1500 including official and unofficial languages (http://www.censusindia.gov.in/, accessed: June 2014). Majority of these languages have descended from two language families: Indo-European and Dravidian. A language family is a group of languages which have descended from a common mother language. With passage of time and increase in number of speakers, the mother language splits into several pronunciations and dialects giving rise to different languages. Therefore, it is obvious that the several Indian languages will have similarity of features if they belong to the same language family. Besides this, there are other factors as well, which influence similarity among languages, for example, geographical location, trade connections, political invasions, and tourist visits. Therefore, study of similarity among languages is important because it throws light in multiple directions. Apart from unfolding mysteries regarding the development of languages and the paths taken by them while transforming from the ancient forms to the modern day colloquial forms, study of similarity among languages gives information about ancient trade connections, routes followed by traders, political relations between states, and geographical conditions prevailing in ancient times to mention a few.

Similarity among languages is a well researched topic and many related articles are found in the literature. The work in [1] finds similarity and dissimilarity among the European languages from textual data using a file compression method. In the work of [2], language similarity is measured using a similarity metric based on ALINE algorithm [3] which examines shared-meaning word pairs and generates a similarity score. In [4], a metric to compute similarity among languages is developed using trigram profiles (list of three letter slice of words and their frequency). The process involves counting the number of trigram profiles between two languages. This is done using a coefficient called Dice's coefficient. All these works compute similarity based on textual data. Works on detecting similarity among languages from spoken utterances are relatively less. In [5], language similarity has been detected from spoken utterance by representing languages in a perceptual similarity space based on their overall phonetic similarity.

The work in this paper uses language verification framework to find similarity among languages. Four language verification systems are designed for each language using the state-of-the-art feature extraction and modeling techniques. Mel Frequency Cepstral Coefficient (MFCC) appended with Shifted Delta Coefficient (SDC) and Speech Signal Based Frequency Cepstral Coefficient (SFCC) appended with Shifted Delta Coefficient (SDC) are used as language depended features. For modeling of feature vectors, Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) are used. All possible combinations of these two feature extraction and two modeling techniques result in four language verification systems. Equal error rate (EER) of each system is evaluated and the systems are analyzed. These systems are used to find similarity among languages.

The rest of the paper is organized as follows. Feature extraction and modeling techniques used in the work are described in Section 2. The method of database preparation and the structure of database are detailed in Section 3. Performance measure used to evaluate the language verification systems is described in Section 4. Section 5 describes how similarity among languages is evaluated using the verification systems, followed by results and discussion in Section 6, computational complexity in Section 7 and finally, conclusion in Section 8.

2. Feature Extraction and Modeling Techniques Used in the Work

Using the two feature extraction techniques and two modeling methods mentioned above, four language verification systems are developed, namely, MFCC + SDC + GMM, MFCC + SDC + SVM, SFCC + SDC + GMM, and SFCC + SDC + SVM. Each language is trained and tested using all these four systems. Brief description of the feature extraction and modeling techniques is given below.

2.1. Feature Extraction

2.1.1. Mel Frequency Cepstral Coefficient (MFCC). According to psychophysical studies, human perception of the frequency content of sounds follows a nonlinear scale called the mel scale defined as

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

where f = perpetual frequency expressed in Hertz. This leads to the definition of MFCC which can be calculated as follows. Speech signal is preemphasized, broken into frames of 20 ms with 10 ms overlap and Hamming windowed. It is then converted to frequency domain which gives the energy spectrum of speech. The resulting signal is passed through mel scale filter bank which is a sequence of 20 triangular filters uniformly spaced in mel scale. Finally Discreet Cosine Transform (DCT) of the logarithm of the result is evaluated which gives the MFCC coefficients [6]. MFCC is also discussed in [7]. MFCC filter bank is shown in Figure 1.
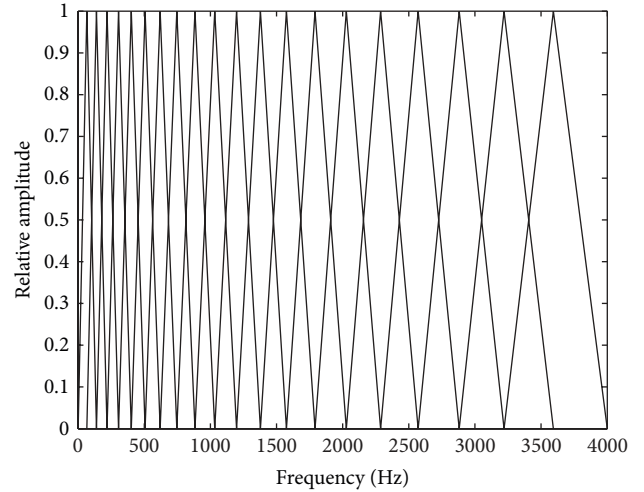

FIGURE 1: MFCC filter bank.

2.1.2. Speech Signal Based Frequency Cepstral Coefficients (SFCC). SFCC has been used for speech recognition and has shown comparable performance to MFCC. It is a frequency warping technique purely based on the speech signal. Like MFCC speech is broken into frames of 20 ms with 10 ms overlap and passed through Hamming window. PSD of each frame is calculated. Logarithm of average PSD over all frames is computed. Average energy is computed by summing up the log PSD's and dividing it by the number of filters in the filter bank. 20 filters have been used in the filter bank. The upper cutoffs for each filter are chosen to be such that the log energy of the filters is equal to the average energy [8]. Once filter bank computation is over, the rest of the procedure is same as that of MFCC computation. The steps involved in computing SFCC filter bank are shown in Figure 2 and the SFCC filter bank is shown in Figure 3.

2.1.3. Shifted Delta Coefficients (SDC). SDC are particularly suitable for language recognition because they capture broader temporal features over a wide range of time. SDC features are specified by four parameters N, d, P, and k. N is the number of cepstral coefficients computed at each frame, d represents the advance and delay in time for delta computation, k is the number of delta-cepstral blocks whose delta-cepstral coefficients are stacked to form the final feature vector, and P is the shift in time between consecutive blocks. The SDC coefficients can be represented by the following equation [9]:

$$\Delta\mathbf{c}(t,i) = \mathbf{c}(t + iP + d) - \mathbf{c}(t + iP - d), \qquad (2)$$

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

Values of N, d, P, and k used for this work are 7, 1, 3, and 7, respectively.

The first 7 MFCC or SFCC coefficients are taken. SDC constitute 49 coefficients (7 sets of delta coefficients, each set having 7 coefficients). So the resulting feature vector has 56 coefficients. Figure 4 shows the steps required to get
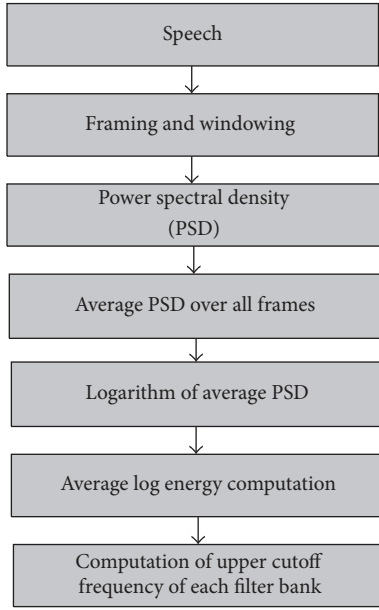
FIGURE 2: Steps required to compute SFCC filter bank.

FIGURE 3: SFCC filter bank.

MFCC/SFCC + SDC features from speech. SDC is discussed in [10, 11].

2.2. Modeling

2.2.1. Gaussian Mixture Model (GMM). For a D dimensional feature vector $\mathbf{o}$, with M being number of Gaussian mixture density functions, the likelihood function of a GMM is the weighted sum

$$\text{pr}(\mathbf{o} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{o}), \qquad (3)$$

where λ represents the complete GMM, p_i, $i = 1, 2, \ldots, M$, are the mixture weights and $b_i(\mathbf{o})$, $i = 1, 2, \ldots, M$, are the component densities. Each component density is a Gaussian function of D dimension of the form

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

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{\Sigma}_i$. We assume diagonal covariance for computational simplicity. The constraint on p_i is $\sum_{i=1}^{M} p_i = 1$. The complete GMM can be represented by $\lambda = \{p_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}, i = 1, 2, \ldots M$.

Given a training data set, a GMM is trained using an Expectation Maximization (EM) algorithm where the model parameters are estimated using Maximum Likelihood (ML) criterion [12].

M is determined experimentally. Value of M used for this work is 64.
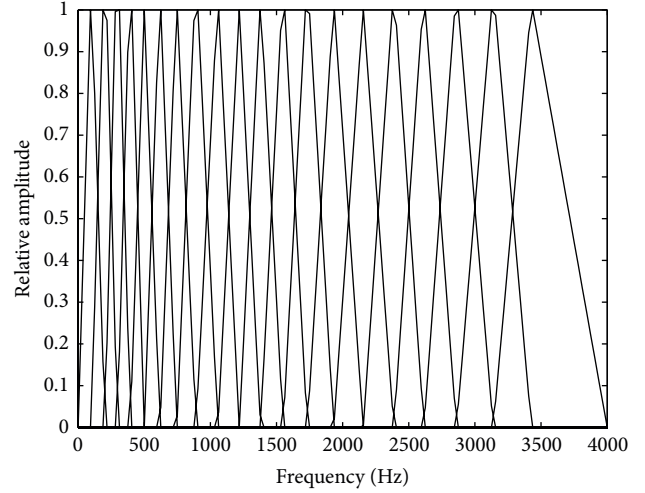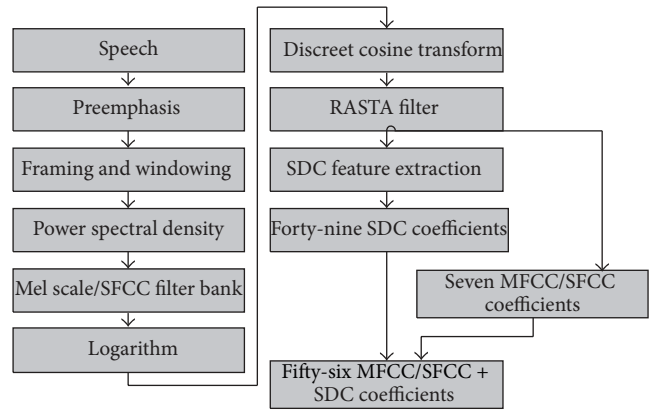

FIGURE 4: Steps required to derive MFCC/SFCC + SDC features from speech.

2.2.2. Support Vector Machine (SVM). SVM is particularly suitable for binary classification. It projects an input vector $\mathbf{x}$ into a scalar value $f(\mathbf{x})$ such that

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

where t_i are the ideal outputs (either +1 or −1), $\sum_{i=1}^{I} \alpha_i t_i = 0$, the weights $\alpha_i > 0$, β is bias, I is the number of support vectors, and $\mathbf{x}_i$ are the support vectors obtained from the training set by an optimization process. The optimization is based on a maximum margin concept. The hyperplane separating the two classes ensures maximum separation between the two classes (Figure 5). $\kappa(\cdot)$ is the kernel function which is constrained to fulfill certain properties called Mercer condition so that it can be expressed as

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

where $\boldsymbol{\phi}(\mathbf{x})$ is a mapping from input space to a high dimensional space. The Mercer condition ensures that the margin
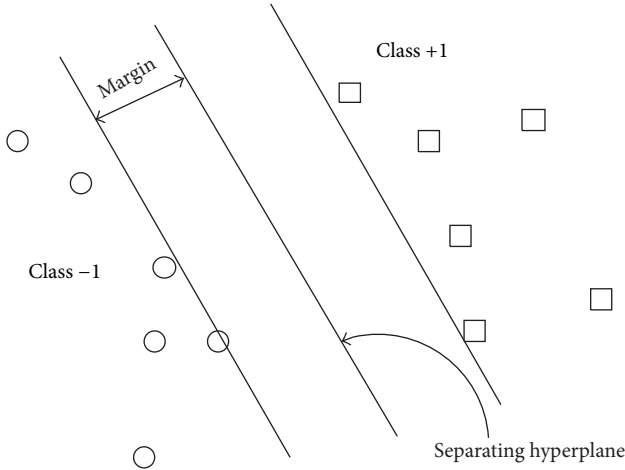
FIGURE 5: The hyperplane ensures maximum separation between the two classes.

FIGURE 6: Formation of GMM Supervectors.

concept is valid, and the optimization of the SVM is bounded [13].

The kernel used for this work is a GMM Supervector Linear Kernel. The means of GMM are stacked to form a GMM mean supervector [14] (Figure 6). The supervectors are a mapping between an utterance and a high dimensional vector. *Maximum-a-posteriori* (MAP) adaptation is used to compute the means of GMM. For this, a Universal Background Model (UBM) is created [15]. A UBM is a large GMM trained to represent the language independent distribution of features.

The GMM Supervector Linear Kernel can be represented by

$$
\kappa\left(\varphi_a, \varphi_b\right) = \sum_{i=1}^{M} p_i \left(\boldsymbol{\mu}_i^a\right)^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i^b
$$
$$
= \sum_{i=1}^{M} \left(\sqrt{p_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_i^a\right)^t \left(\sqrt{p_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_i^b\right), \tag{7}
$$

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. p_i and $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, M$, are the weight and covariance matrix of ith Gaussian as mentioned earlier. Detailed description of this method is given in [14].

3. Database Preparation

The database used for this work is prepared from All India Radio (PRASAR BHARATI, http://newsonair.nic.in/, accessed: August 2011) website repository. Daily news bulletins of all major Indian languages are available here.

Main reasons for selecting this website as the data source are mentioned below.

(i) The quality of speech is good as it contains very little noise.

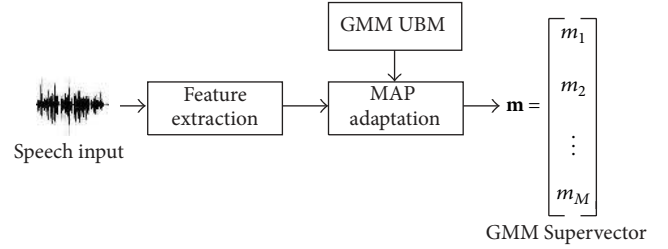(ii) Speech is sufficiently loud and is of sufficiently long duration.

(iii) A large number of speakers (news readers) are available in each language. This reduces speaker bias.

(iv) Both male and female speakers are present in almost all the available languages ensuring little or no gender bias.

(v) Speech covers a large variety of topics that helps to capture details of acoustic information.

16 languages have been used for this work: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. To prepare the database, all news files available in these 16 languages on a particular date were downloaded. Each of the news files was listened to and files of poor quality were rejected. The speech was originally in.mp3 format. It was converted to.wav format to enable further processing.

The organization of the train and test data sets is shown in the Figure 7. For training, four and half hours of speech is taken from the master database for each language. These four and half hours contain half the duration of male speech and half female speech. English, Sanskrit, Kashmiri, and Urdu do not have equal proportion of male and female data in the master database. For these languages exact equal division of male and female speech is not possible. Therefore, nearly equal proportion is taken for these languages except for Kashmiri, which does not contain any female files. Speaker variability is obtained by taking speech from a variety of speakers. This is done so that the system does not get biased towards a particular speaker.

The data thus collected (in the form of news files) are broken into segments of 30 seconds. The segments are listened to and those containing music, long duration of silence, and other unwanted voices are deleted. Only segments containing clean speech are retained. This results in a decrease in the total duration of data. So initially four and half hours of speech is taken so that total duration is never less than four hours after removing noise. The resulting data set containing 16 languages, each language containing four hours of clean speech, constitutes the train data set.

In order to test a verification system, two kinds of test data are required, namely, target test data and nontarget test data. Test data belonging to the target language is called target test data and test data belonging to all other languages are called nontarget test data. For example, if a verification system is designed for language *A*, then test utterances of language *A* are called target test utterances and test utterances belonging
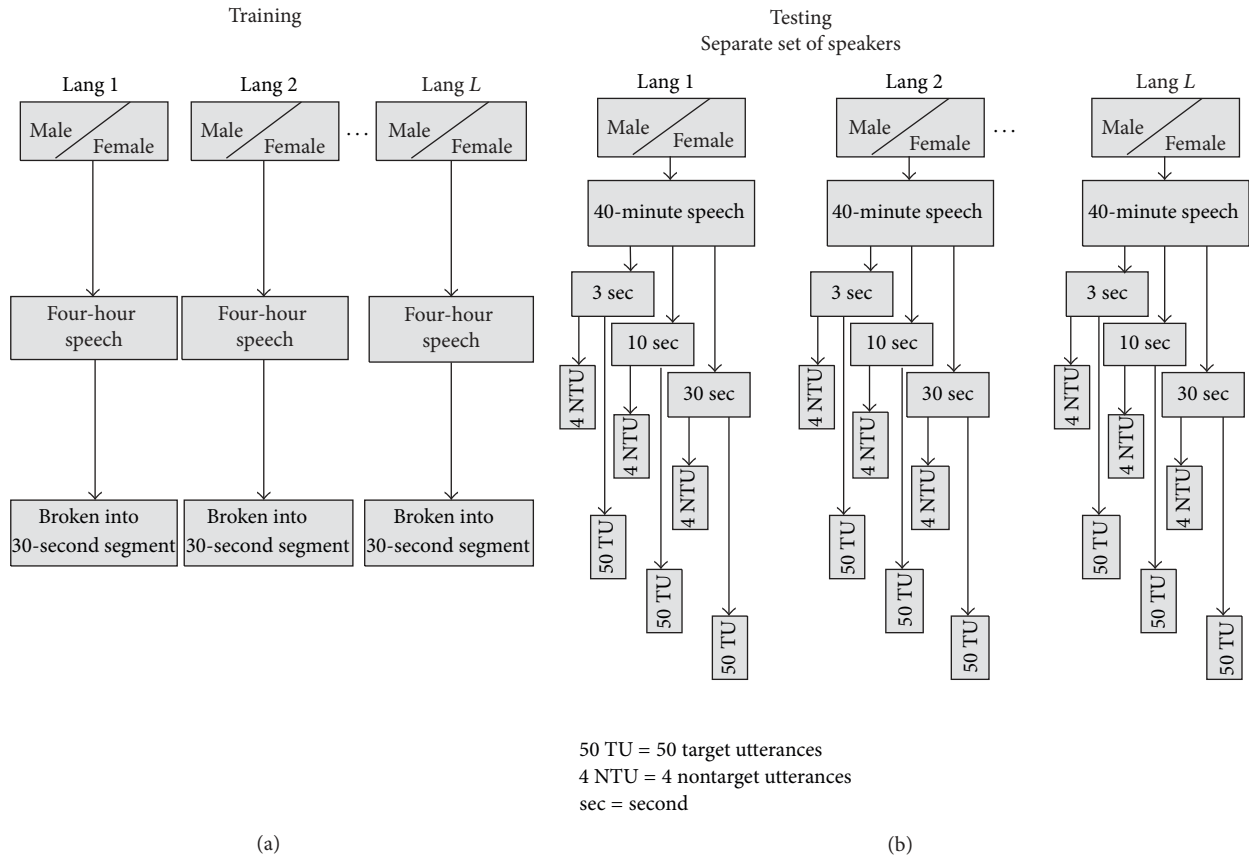
FIGURE 7: Train and test data set for language verification.

to all other languages apart from language *A* are called nontarget test utterances. For an ideal language verification system, all target test utterances should give score greater than the threshold score whereas all nontarget test utterances should give score less than the threshold score.

Testing is done for three different test duration lengths, short (3 seconds), medium (10 seconds), and long (30 seconds). For preparing the target test data of each language, 40 minutes of speech is taken from each language (male female inclusive). For Kashmiri, which does not have any female data, 40 minutes of male speech is used. The train and test speaker sets are mutually exclusive. This is done so that the verification results do not get biased by speakers.

The selected data (news files) are broken into segments of 30-second duration. These segments are listened to and segments containing music, long duration of silence, and other unwanted voices are removed as in case of training. Only segments containing clean speech are retained. Each of these clean 30-second segments are further split into 10-second and 3-second segments resulting in 3-second, 10-second, and 30-second segments of clean speech. Now 25 male and 25 female segments are selected randomly from each test segment duration. So, for each test utterance duration, there are 50 utterances, 25 male and 25 female (except Kashmiri where 50 male segments are selected). This is the target test data set.

Test utterances from all 15 languages apart from the target language are nontarget test utterances. Therefore, the data set from which 25 male and 25 female utterances are selected as target utterances for a language can serve as the data set of nontarget utterances for all other languages. For each target language, two male and two female test utterances are selected from this data set from each of the 15 nontarget languages. For example, when language *A* is the target language, two male and two female test utterances are selected from each of the other 15 languages (languages *B*, *C*, ..., excluding language *A*). Four test utterances from each nontarget language results in 60 nontarget test utterances. Similar nontarget test utterances are prepared for each of the 16 languages for each test duration (3 seconds, 10 seconds, and 30 seconds). This results in one set of nontarget test utterances. Five similar mutually exclusive nontarget test utterance sets are prepared in order to check the consistency of the result.

4. Performance Measure

The performance of a language verification system is measured by two types of errors.

  (i) False acceptance (FA): This error occurs when a nontarget utterance is accepted as target utterance.

  (ii) False rejection (FR): This error occurs when a target utterance is rejected as nontarget utterance.

For performance measurement, the rate of false acceptance R_{FA} and the rate of false rejection R_{FR} are calculated. They are computed using the formulae

$$R_{FA} = \frac{\text{Number of False Accepted Samples}}{\text{Total Number of Non-Target Samples}},$$
$$R_{FR} = \frac{\text{Number of False Rejected Samples}}{\text{Total Number of Target Samples}}.$$
(8)

The R_{FA} and R_{FR} values depend on the decision threshold of the verification system. If threshold is set very high, R_{FA} would be minimized, but many target samples would be rejected hence increasing R_{FR}. On the other hand, if threshold is set low, target samples would be accepted (R_{FR} reduced), but many nontarget samples would also be accepted at the same time (R_{FA} increased). So the threshold is set experimentally to such a value so as to have acceptable R_{FA} and R_{FR}.

4.1. Equal Error Rate (EER). EER is a performance criterion used to measure the performance of a language verification system. It refers to the operating point where R_{FA} is equal to R_{FR}.

4.2. Detection Error Tradeoff (DET) Curve. Unlike EER which measures the performance of a verification system at a particular operating point, the DET curve can be used to view the performance at various points [16]. This curve plots the variation of R_{FA} to R_{FR} according to varying decision thresholds. The point on the graph where R_{FA} and R_{FR} have equal values represents the EER [17]. Figure 8 depicts a typical DET plot.

5. Language Verification and Detection of Similarity

The work is divided into two parts. In the first part, language verification systems are designed for the 16 languages. Each system is tested five times with five sets of nontarget data. The target data set remains the same for the five tests (five different target data sets for each target language could not be taken due to shortage of data). This results in five test data sets for each language. DET curves are plotted for each test data set. Each curve gives one EER value. The average of these five EER values is noted. Similar testing is done for 3-second, 10-second, and 30-second test utterances and the average EER values are noted in each case and analyzed.

The second part of the work finds out similarity among languages. The focus of this part is on the nontarget test utterances only. A certain false acceptance percentage is fixed and the languages of the false accepted samples at that percentage are noted. For example, let false acceptance be fixed at X%. Let X% of 60 (60 is the number of nontarget test utterances in each set) be Y. So the first Y false accepted samples are found out and their languages are noted. These are the nontarget samples with the highest scores, that is, the most confusing ones. Among these Y samples, let Y_{A1} number of samples belong to nontarget language A, let Y_{B1} number of samples belong to nontarget language B, and so on until Y_{O1}
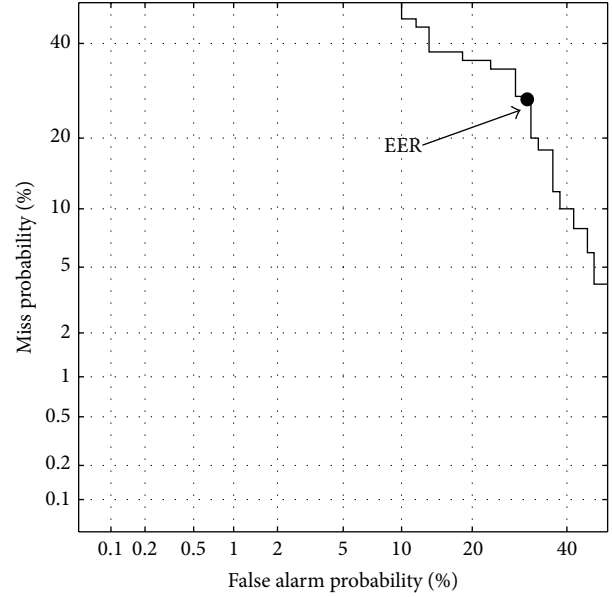

FIGURE 8: DET plot of a representative language verification system. x-axis represents R_{FA} and y-axis represents R_{FR}.

number of samples belong to nontarget language O (all from the 1st nontarget utterance set) such that

$$Y_{A1} + Y_{B1} + \cdots + Y_{O1} = Y_1 = Y,$$
(9)

where O is the 15th nontarget language. Similarly for the four other sets of nontarget test data, $Y_{A2}, Y_{B2} \cdots Y_{O2}, Y_{A3}, Y_{B3} \cdots Y_{O3}, Y_{A4}, Y_{B4} \cdots Y_{O4}$, and $Y_{A5}, Y_{B5} \cdots Y_{O5}$ are found out. Equation (9) can be represented in a more compact mathematical form as

$$\sum_{j=1}^{15} Y_{ji} = Y_i = Y,$$
(10)

where Y_{ji} represents the number of false accepted samples from language j coming from the ith nontarget set. It is to be noted that any of these values Y_{ji} can be 0. Only constraint is that their sum should be equal to Y as shown in (9) or (10). The percentage of false acceptance from nontarget language A is calculated as

$$Y_A\% = \frac{(Y_{A1} + Y_{A2} + Y_{A3} + Y_{A4} + Y_{A5})}{20} \times 100\%.$$
(11)

A mathematically compact representation of the above is

$$Y_A\% = \frac{\sum_{i=1}^{5} Y_{Ai}}{20} \times 100\%$$
(12)

since there are 20 utterances from nontarget language A in total (there are five nontarget sets with four utterances of language A in each set). Similar calculation is done for every nontarget language for each test utterance duration. The process is repeated for each of the 16 target languages.

The percentage of false acceptance of each nontarget language against each target language is plotted. This percentage

gives a measure of similarity between the two languages. For example if nontarget language A shows high percentage of false acceptance against target language T, it can be concluded that language A is similar to language T since the system cannot separate A from T. In such a case, language T should also shows high percentage of false acceptance against target language A.

Four language verification systems are prepared for each target language. Experiments are done with all the four systems (MFCC + SDC + GMM, MFCC + SDC + SVM, SFCC + SDC + GMM, and SFCC + SDC + SVM) so that any abrupt behavior by any one system can be ruled out and results are reliable.

In case of GMM, training is done with four hours of data (method of data preparation is described in Section 3). But, in case of SVM, training is done with 40 minutes of data. This is because SVM is a discriminative modeling technique and it requires target (positive) as well as nontarget (negative) data for training. Therefore, four hours of positive and negative data from the 16 languages would be huge and computation time would be high. Instead, 40 minutes of data is used, both for target and nontarget languages. This 40 minutes data is taken from the training data set. Since the goal of this work is not to compare the performance of the four systems but to investigate similarity among languages, it is not necessary that the GMM and SVM systems be trained with the same data.

For UBM preparation, the same 40 minutes data is used as that of SVM. Therefore, the UBM data is also a subset of the training data.

For computing SFCC features, data dependent SFCC filter bank is computed with the UBM data and is used universally for all languages.

Using each of the four systems mentioned above, language models are created for 16 languages and they are tested with utterances of length 3 seconds, 10 seconds, and 30 seconds.

6. Results and Discussion

The experiments have two parts. The first part deals with analysis of the verification systems. The second part uses the verification systems to find out similarity among languages.

6.1. Analysis of Language Verification Systems. The average EER values of all the language verification systems are given in Table 1. The table shows that mostly average EER values of the systems are highest when 3-second test utterances are used. Average EER decreases as test utterance length increases from 3 seconds to 10 seconds to 30 seconds. This is expected because the length of feature vectors produced by 3-second utterances are very small and so carry less language discriminative information. So the overlap between target and nontarget scores is high resulting in high EER. The only exception to this fact is shown by Konkani for the GMM systems and to a small extent by English. This can be considered as abrupt behavior of the system. For this reason four different systems are used for the experiments.

6.1.1. Analysis of EER Values

(i) Bengali, English, Kannada, Kashmiri, and Telugu systems show average EER values less than 20 for most of the techniques.

(ii) Assamese, Hindi, Konkani, and Sanskrit systems have average EER in the range 10 to 30 for most of the techniques.

(iii) Gujarati, Malayalam, Marathi, Odia, and Urdu mostly have average EER values in the range of 20 to 40. Gujarati and Urdu have EER values above 40 in certain cases. But these occurrences are very rare (two cases in Urdu and one case in Gujarati) and can be considered as outliers.

(iv) Punjabi and Tamil show wide variation in EER using GMM and SVM systems. For Punjabi average EER values mostly vary between 7 and 22 using GMM. But with SVM results vary between 26 and 37. For Tamil, average EER values using GMM are less than 20 whereas using SVM values are greater than 20. This implies that the features of these languages are better discriminated by GMM than SVM. But since system comparison is not the goal of this work, results of both GMM and SVM systems are considered.

DET plots of 3-second test utterances of Bengali, Assamese, Gujarati, and Punjabi for the five test data sets are shown in Figures 9 and 10 in order to show the nature of the graphs. The 10-second and 30-second plots are similar in nature.

6.1.2. Cause of High Variation of EER from Language to Language. All language models have been prepared using the same techniques and with the same amount of data. Also the data for all languages are collected from the same source; therefore there is very little scope for any one language to get more affected by noise or other external factors. In spite of this, the language verification systems show high variations in EER. This can only be due to the discriminative property of the languages. Since most of the languages have developed from a common source language and belong to two major language families, it is obvious that there will be overlapping features among them. The greater the overlap is, the lesser the discriminative capacity of the language would be, hence showing higher EER. This can be explained from Figure 11. In the figure, language A has the highest amount of overlap (shaded region represents overlap); hence a language verification system for language A would show the highest EER whereas a verification system for language C would show the lowest EER since its overlapped region is minimum.

6.2. Detection of Similarity among Languages Using Language Verification Systems. The concept of the experiment is to select a particular false acceptance percentage (X) and analyze the languages of the false accepted samples at that percentage as discussed in Section 5. Initially, false acceptance of 15% ($X = 15$) is chosen for the experiments since it lies nearly at the middle of the average EER values of most of the languages. Since there are 60 nontarget samples
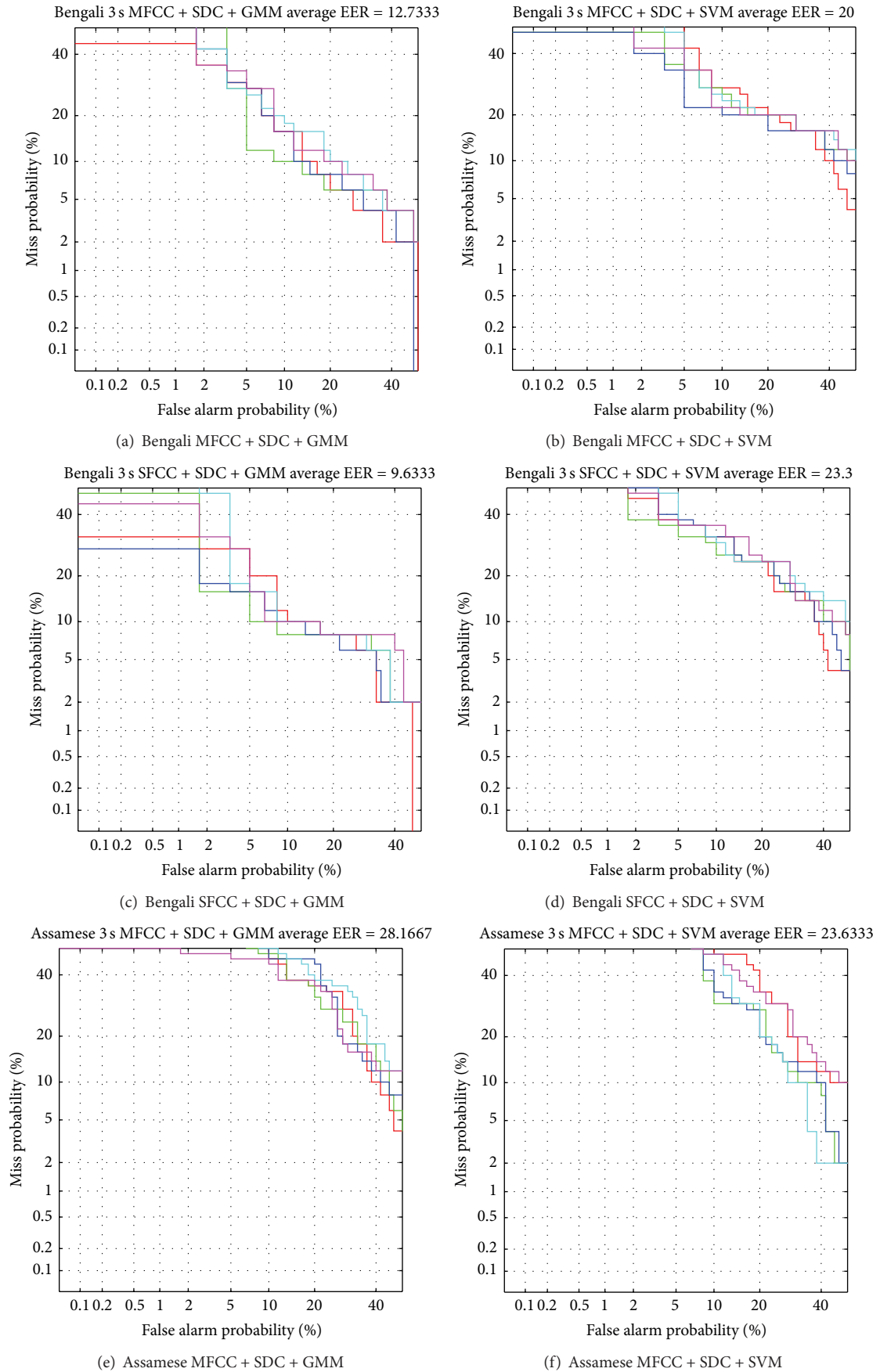
(a) Bengali MFCC + SDC + GMM

(b) Bengali MFCC + SDC + SVM

(c) Bengali SFCC + SDC + GMM

(d) Bengali SFCC + SDC + SVM

(e) Assamese MFCC + SDC + GMM

(f) Assamese MFCC + SDC + SVM

Figure 9: Continued.

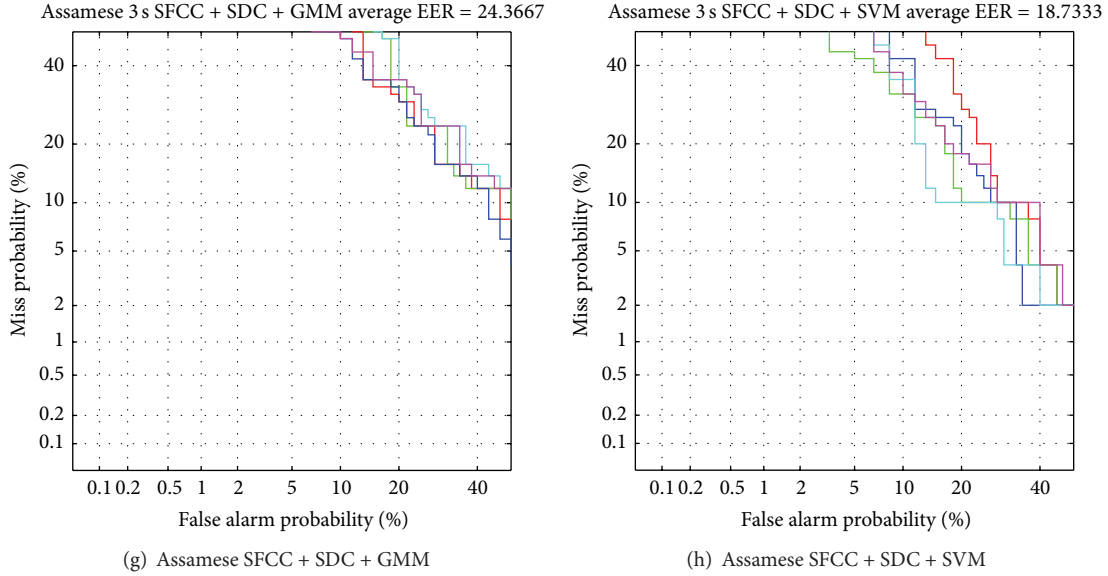(g) Assamese SFCC + SDC + GMM

(h) Assamese SFCC + SDC + SVM

FIGURE 9: DET plots of Bengali and Assamese for 3-second test utterance length (plots of five data sets are represented by five colors) (s = second).

TABLE 1: Average EER values of verification systems of 16 languages using the four systems (sec = second).

Language	MFCC + SDC + GMM			MFCC + SDC + SVM			SFCC + SDC + GMM			SFCC + SDC + SVM		
	3 sec	10 sec	30 sec	3 sec	10 sec	30 sec	3 sec	10 sec	30 sec	3 sec	10 sec	30 sec
Assamese	28.17	17.43	15.97	23.63	14.20	4.77	24.37	22.20	22.73	18.73	7.47	1.83
Bengali	12.73	7.43	3.30	20.00	10.73	6.57	9.63	4.03	0.00	23.30	7.07	2.93
English	15.60	8.53	9.63	28.70	17.07	6.37	11.10	4.03	5.83	32.93	14.90	6.73
Gujarati	34.37	29.63	29.63	40.73	30.73	28.53	34.90	28.53	29.63	35.97	26.70	20.00
Hindi	29.27	23.67	23.10	26.33	25.80	26.33	21.83	18.37	16.70	27.07	26.17	24.37
Kannada	16.53	11.83	10.00	17.07	5.83	4.93	19.83	14.37	14.90	18.53	10.00	7.80
Kashmiri	22.93	10.00	2.93	16.53	3.67	2.20	19.47	3.63	1.47	19.10	4.40	1.83
Konkani	24.00	29.27	36.33	21.83	14.37	10.00	19.80	23.63	31.10	20.00	11.10	4.93
Malayalam	25.07	25.07	22.73	30.37	23.83	14.53	25.80	23.63	21.80	29.63	25.80	15.43
Marathi	28.17	28.53	27.80	28.00	21.10	17.07	25.27	28.90	23.83	30.37	18.53	18.53
Odia	30.73	26.17	24.73	26.33	13.67	8.90	38.17	27.80	26.53	28.53	22.93	12.73
Punjabi	18.53	13.80	7.43	32.90	28.90	26.73	21.83	16.17	12.20	36.90	29.27	30.00
Sanskrit	26.70	24.53	22.53	27.80	17.27	11.47	22.37	18.53	14.90	24.20	19.27	11.83
Tamil	19.10	10.73	7.80	33.63	29.83	21.63	18.73	10.37	8.17	30.00	26.17	21.10
Telugu	15.63	14.53	8.53	24.00	16.90	7.10	15.27	12.57	6.37	28.17	21.47	13.10
Urdu	30.00	33.43	21.47	38.37	38.90	28.90	37.80	41.47	27.07	37.80	43.63	34.90

for each target language in a particular nontarget set, nine false accepted samples are taken for analysis ($Y = 9$).

6.2.1. Experiment with 15% False Acceptance ($X = 15$). The graphs in Figures 12, 13, and 14 show the false acceptance of 3-second, 10-second, and 30-second utterances for all target and nontarget language combinations at $X = 15$. The rows represent target languages and the columns represent the languages of the false accepted samples. The bars represent percentage of false acceptance of a particular nontarget language (Y_A%). The higher the bars, the greater the false

acceptance and hence the greater the similarity between the two languages represented by the row and column.

(1) Analysis and Observation. The 3-second graph has 8 blank plots, the 10-second graph has 15 blank plots, and the 30-second graph has 43 blank plots. This indicates that, in the 3-second graph, the nine false accepted samples ($Y = 9$) for each target language are more or less evenly distributed among all nontarget languages with lesser number of peaks. This further indicates that all nontarget languages give scores close to each other for 3-second test samples. Therefore
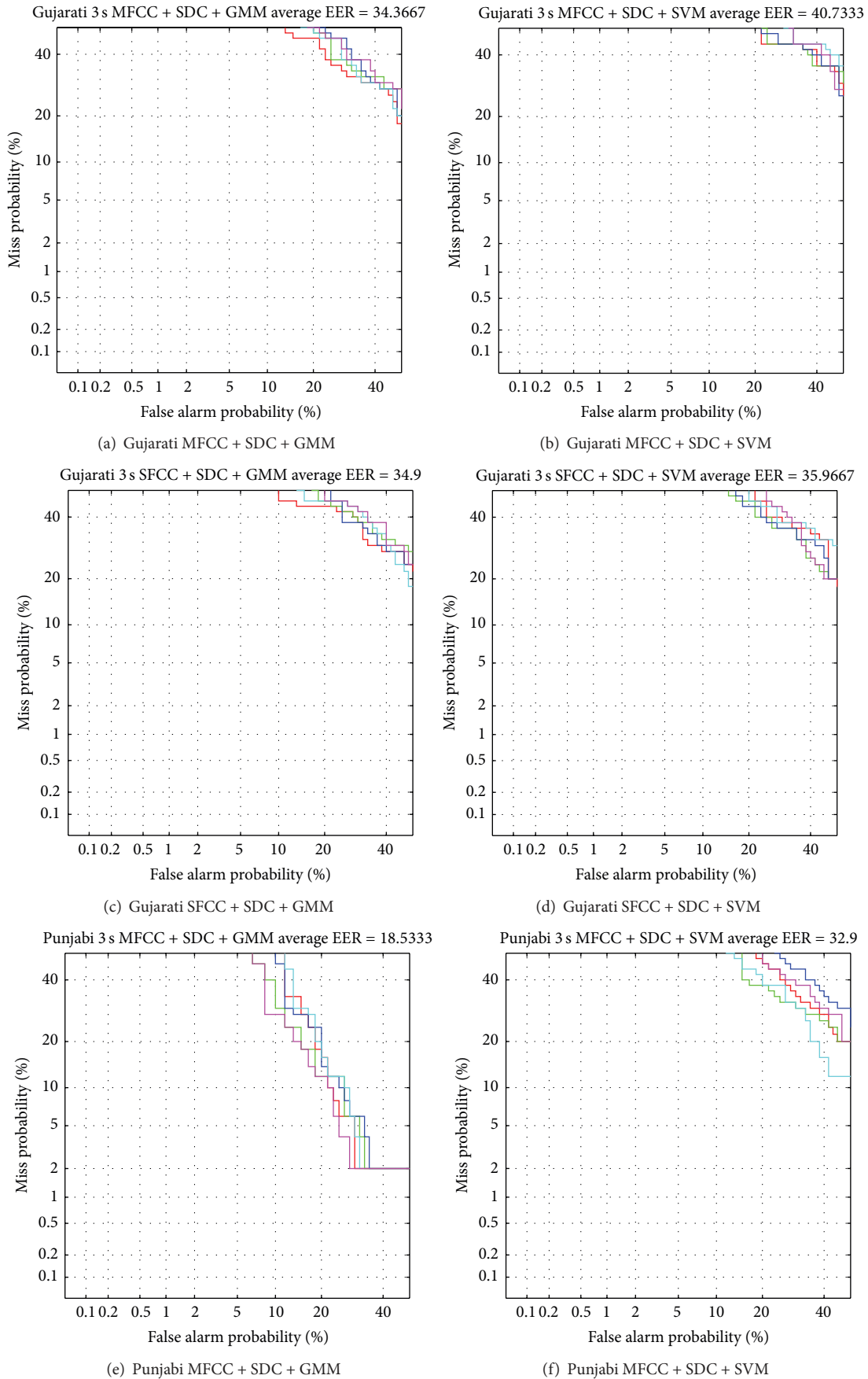
(a) Gujarati MFCC + SDC + GMM

(b) Gujarati MFCC + SDC + SVM

(c) Gujarati SFCC + SDC + GMM

(d) Gujarati SFCC + SDC + SVM

(e) Punjabi MFCC + SDC + GMM

(f) Punjabi MFCC + SDC + SVM

FIGURE 10: Continued.

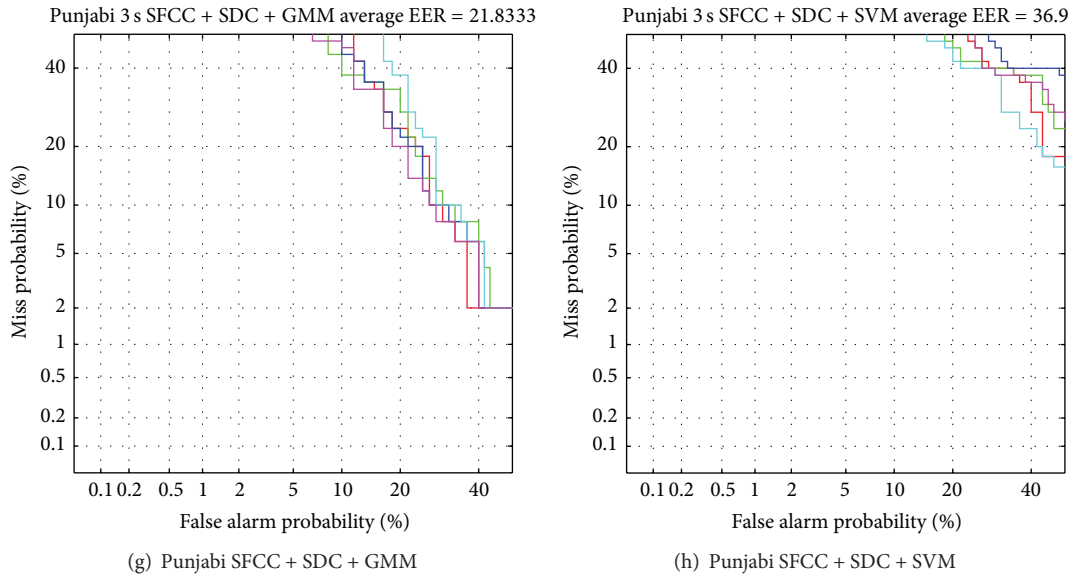(g) Punjabi SFCC + SDC + GMM

(h) Punjabi SFCC + SDC + SVM

FIGURE 10: DET plots of Gujarati and Punjabi for 3-second test utterance length (plots of five data sets are represented by five colors) (s = second).
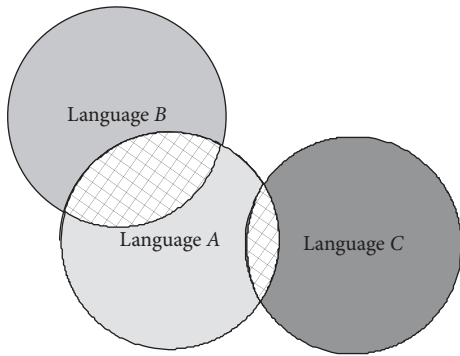

FIGURE 11: Venn diagram showing overlapping features of languages.

language discriminating information is available in the least amount in 3-second graph. The reason for this is that 3-second samples are too small in length to give a reasonable score. Information content increases with test utterance duration.

Ideally the upper triangular and lower triangular plots of the graphs should be equal because the (i, j)th plot represents false acceptance of jth language (Y_j%) in ith target language and (j, i)th plot represents false acceptance of ith language (Y_i%) in jth target language. Therefore, if ith language is false accepted as jth language, ideally the reverse should also happen. In the figures, the upper triangular and lower triangular plots, though not exactly equal, are mostly similar.

Since the 30-second graph contains maximum language discriminative information, further analysis is done with the 30-second graph only. If nontarget language A shows Y_A% of 25% or more with target language T in at least three out of the four verification systems, then languages A and T are considered to be similar. Based on this assumption, the following observations can be made from the graph.

(i) Malayalam shows Y_{Tamil}% in the range 75%–90% (range signifies Y_A% of the three systems) while Tamil shows $Y_{\text{Malayalam}}$% in the range 55%–75% indicating high degree of similarity between the two languages.

(ii) English shows Y_{Kashmiri}% in the range 65%–90% and Kashmiri shows Y_{English}% in the range 60%–75% which indicates that the two languages have high similarity.

(iii) Gujarati shows Y_{Hindi}% in the range 25%–50% while Hindi shows Y_{Gujarati}% in the range 30%–55%. This indicates similarity between Hindi and Gujarati.

(iv) Gujarati also shows Y_{Marathi}% in the range 25%–75% while Marathi shows Y_{Gujarati}% in the range 50%–70%. This indicates that Gujarati and Marathi are similar.

(v) Hindi shows Y_{Marathi}% in the range 30%–75% and Marathi shows Y_{Hindi}% in the range 30%–50% indicating similarity between these two languages.

The observations obtained above are valid and can be justified from a geopolitical point of view.

(i) Tamil and Malayalam belong to Dravidian language family and are neighboring languages spoken in Tamil Nadu and Kerala, respectively (Figure 15). This justifies similarity between Tamil and Malayalam.

(ii) It is a well known fact that Gujarati, Marathi, and Hindi languages sound similar. Also locationwise they are neighboring languages since Gujarati is spoken in Gujarat, Marathi in Maharashtra, and Hindi in Madhya Pradesh which are neighboring states (Figure 15). Therefore, similarity among the languages is obvious.

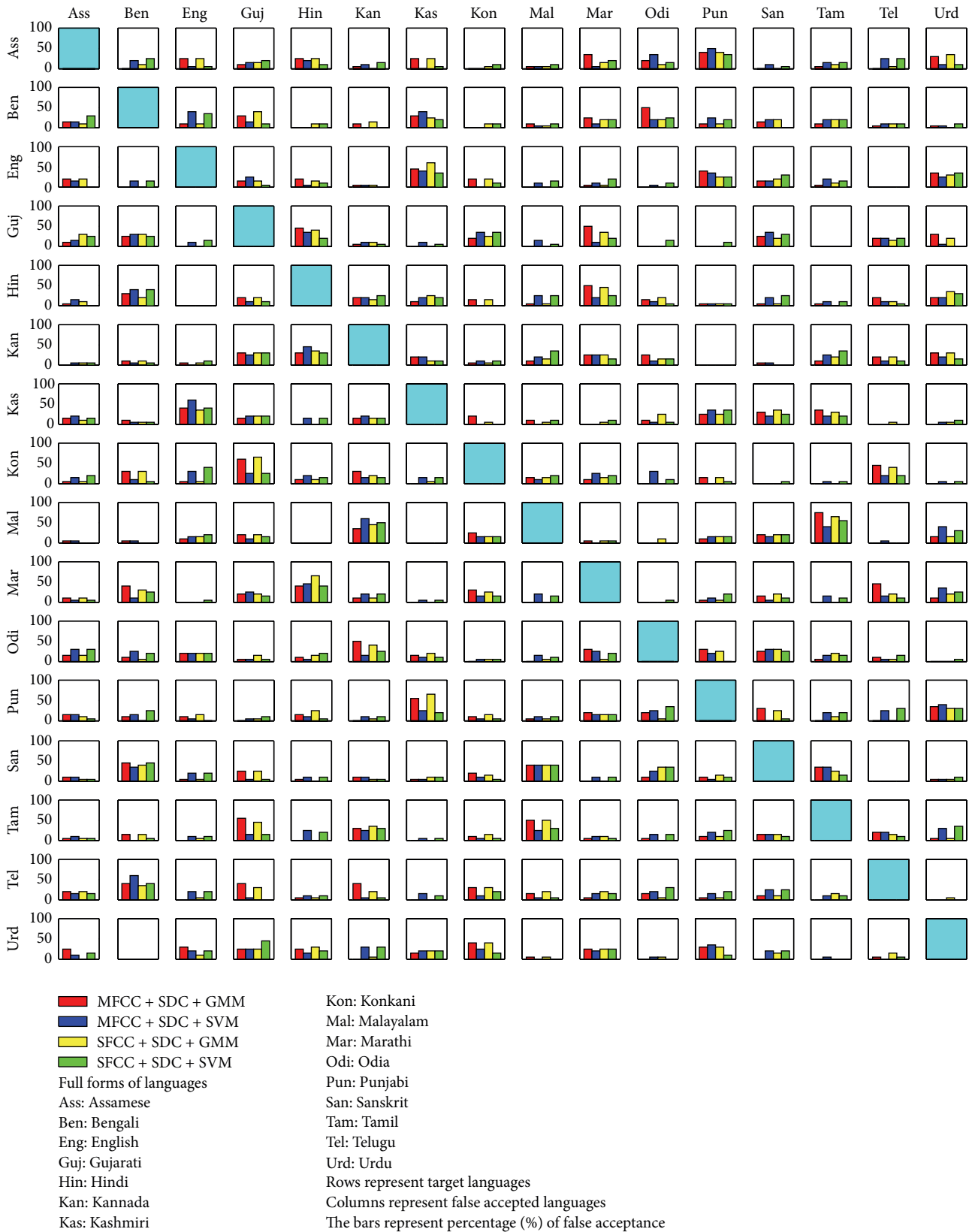(iii) The similarity between English and Kashmiri can be attributed to the fact that Kashmir is a tourist

FIGURE 12: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 3-second test samples using the four systems at $X = 15$.
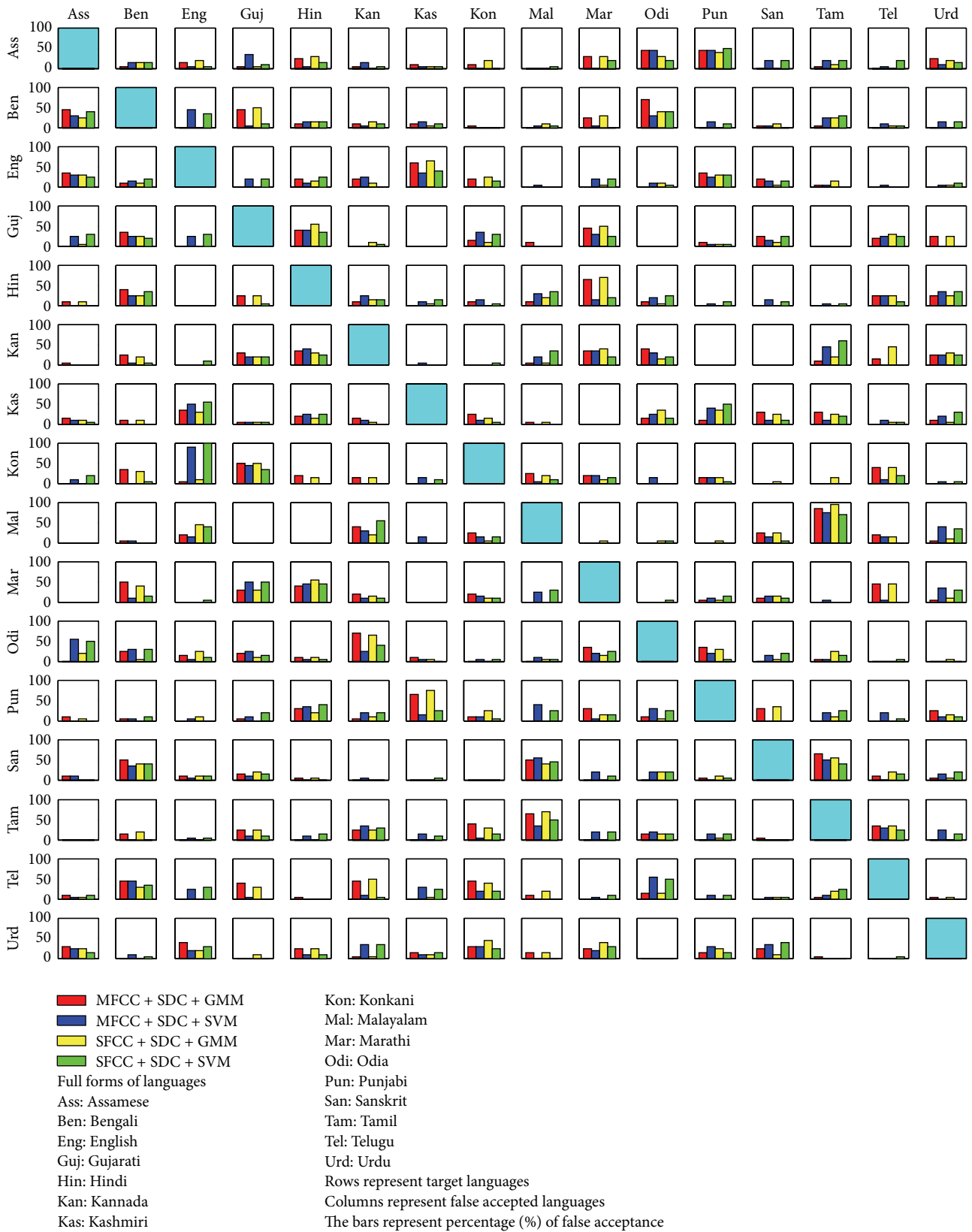
FIGURE 13: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 10-second test samples using the four systems at $X = 15$.
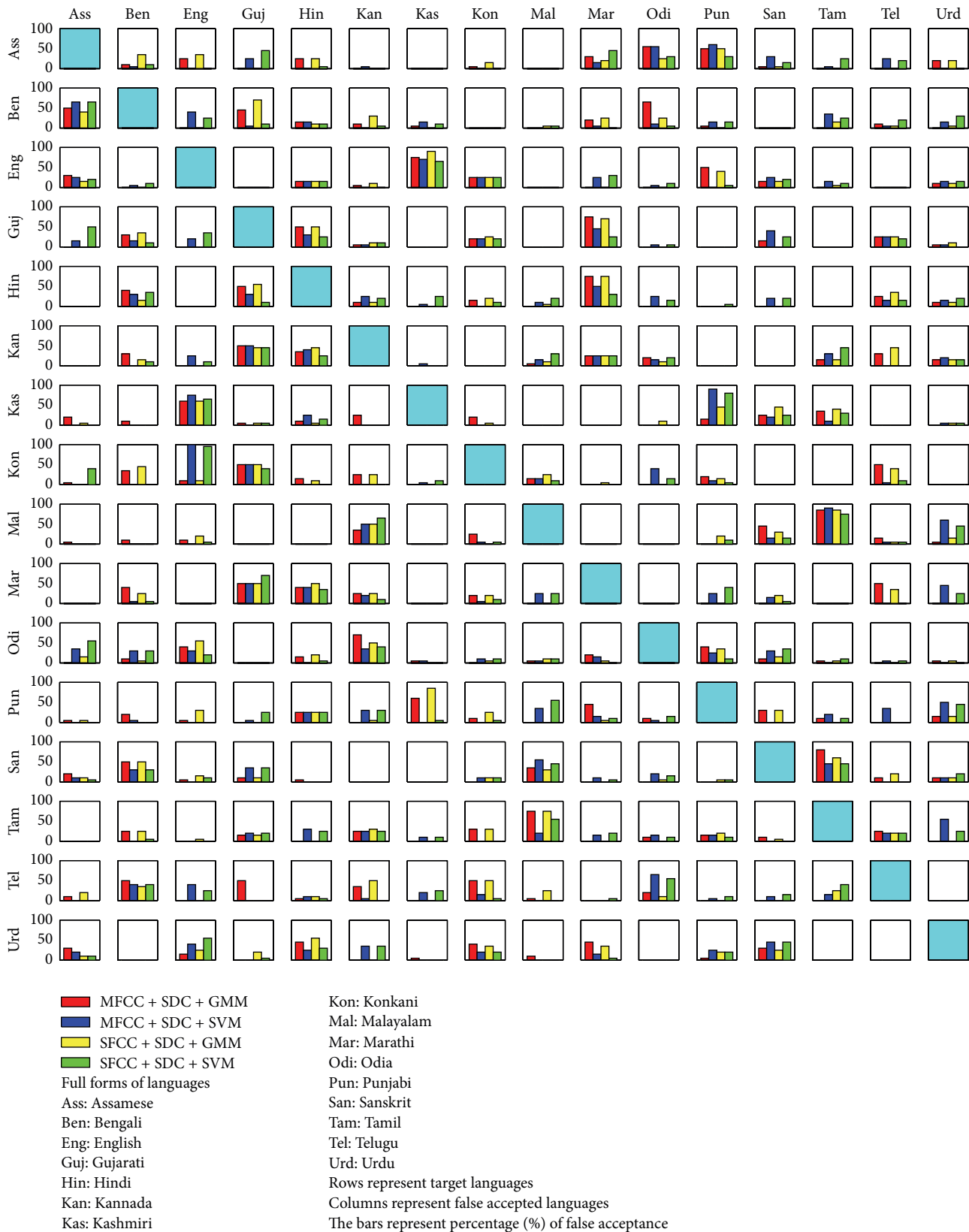
Figure 14: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 30-second test samples using the four systems at $X = 15$.
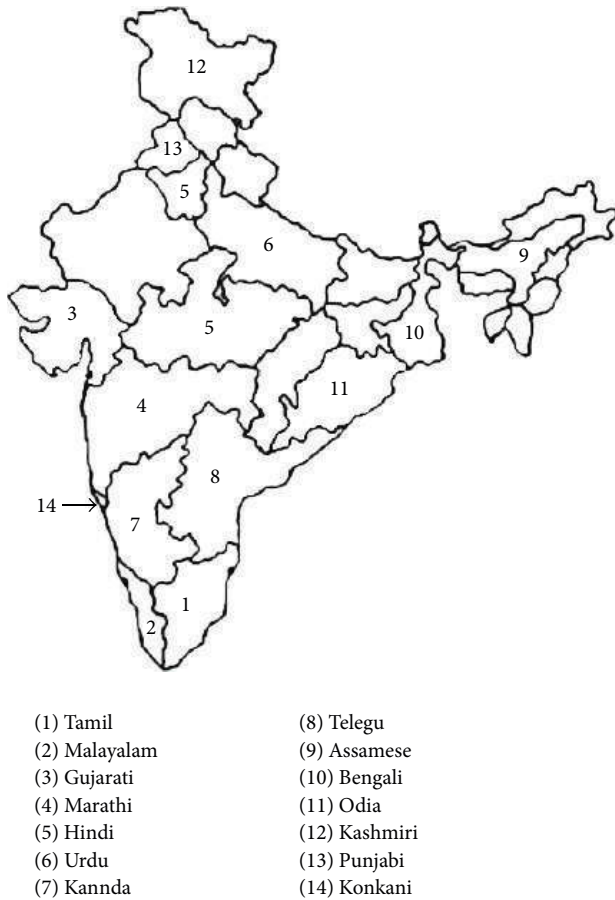
(1) Tamil
(2) Malayalam
(3) Gujarati
(4) Marathi
(5) Hindi
(6) Urdu
(7) Kannda
(8) Telegu
(9) Assamese
(10) Bengali
(11) Odia
(12) Kashmiri
(13) Punjabi
(14) Konkani

FIGURE 15: Map indicating languages and states where they are spoken.

place which people from all over the country visit and since English is universally used as the language of communication in the country, intermingling of the two languages has taken place. Kashmiri people also use English frequently as a second language (ETHNOLOGUE: Languages of the World, http://www.ethnologue.com/, accessed: March 2014) which might be another cause of intermingling of the two languages.

Though the results obtained in the above analysis are valid and logical, there are some plots in the graph which are incoherent. For example, Konkani shows Y_{Gujarati}% in the range 40%–50%, Urdu shows Y_{English}% in the range 25%–55%, Sanskrit shows Y_{Tamil}% in the range 45%–80%, Assamese shows Y_{Punjabi}% in the range 30%–60%, and Urdu shows Y_{Hindi}% in the range 25%–55%, whereas the opposite false acceptances are less than 25% in each case; that is, Y_{Konkani}% shown by Gujarati, Y_{Urdu}% shown by English, and so forth are <25%. These can either be due to abrupt behavior of the test data or can be due to the fact that the total number of false accepted samples (Y) is restricted to nine, which means if language A is similar to language B and language C, it would give Y_B% > 25% if it is more similar to language B but Y_C% would be <25%. Since Gujarati has

shown high false acceptance with Hindi and Marathi, it might not have enough false accepted samples left to show similarity with Konkani. Similar to this might be the case with the other languages showing incoherence. So if Y is increased, then these similarities might be detected. But increasing Y to a high value is not a good idea because that would permit unwanted test samples to enter into the graph. So Y is increased by a small amount (from 9 to 12) and the experiment is done again. Increasing Y from 9 to 12 implies increase of X% from 15% to 20%.

6.2.2. Experiment with 20% False Acceptance ($X = 20$). The graphs in Figures 16, 17, and 18 show the false accepted samples of 3-second, 10-second, and 30-second utterances for all target and nontarget language combinations at $X = 20$. The rows, columns and the bars represent the same things as discussed in Section 6.2.1.

(1) Analysis and Observation. The 3-second graph has 2 blank plots, the 10-second graph has 8 blank plots, and the 30-second graph has 27 blank plots. The reason behind this increasing number of blank plots as test utterance duration increases is already described in Section 6.2.1(1). Number of blank plots is less compared to the earlier case because here Y is greater. As in previous case, only the 30-second graph is used for analysis. If nontarget language A shows Y_A% of 25% or more with target language T in at least three out of the four verification systems, then languages A and T are considered to be similar as in previous case. Based on these assumptions, the observations are listed below.

(i) Malayalam shows Y_{Tamil}% in the range 85%–100% while Tamil shows $Y_{\text{Malayalam}}$% in the range 45%–75% indicating high degree of similarity between the two languages.

(ii) English shows Y_{Kashmiri}% in the range 75%–95% and Kashmiri shows Y_{English}% in the range 70%–90% which indicates that the two languages have high similarity.

(iii) Gujarati shows Y_{Hindi}% in the range 35%–50% while Hindi shows Y_{Gujarati}% in the range 35%–65%. This indicates similarity between Hindi and Gujarati.

(iv) Gujarati also shows Y_{Marathi}% in the range 25%–75% while Marathi shows Y_{Gujarati}% in the range 50%–75%. This indicates that Gujarati and Marathi are similar.

(v) Hindi shows Y_{Marathi}% in the range 45%–85% and Marathi shows Y_{Hindi}% in the range 50%–75% indicating similarity between these two languages.

(vi) Gujarati shows Y_{Konkani}% in the range 25%–45% while Konkani shows Y_{Gujarati}% in the range 45%–55% indicating similarity between the two languages.

(vii) Assamese shows Y_{Odia}% in the range 35%–65% while Odia shows Y_{Assamese}% in the range 30%–70% which indicates similarity between Assamese and Odia.

(viii) English shows Y_{Konkani}% in the range 25%–30% and Konkani shows Y_{English}% in the range 25%–100% which indicates that English and Konkani are similar.
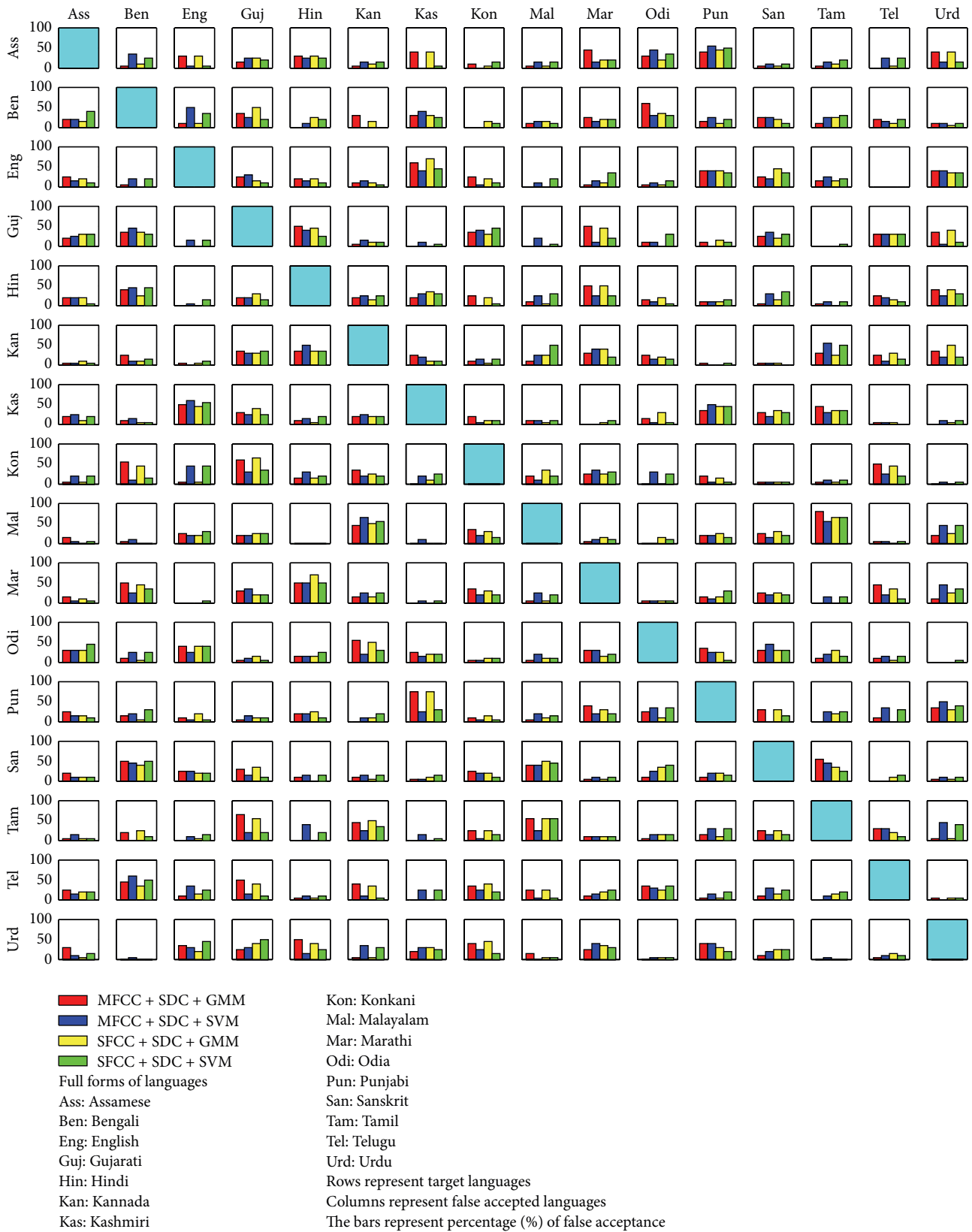
FIGURE 16: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 3-second test samples using the four systems at $X = 20$.
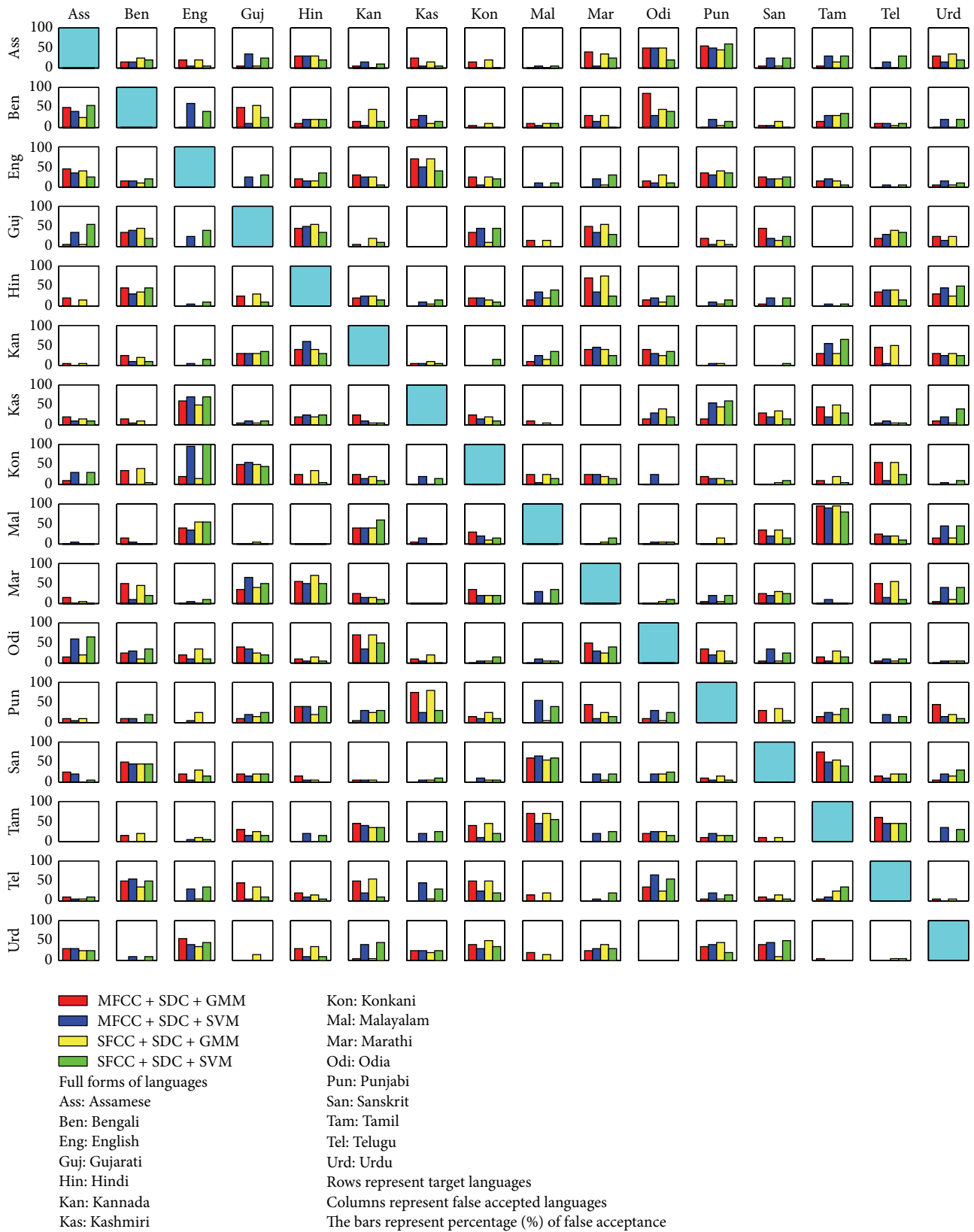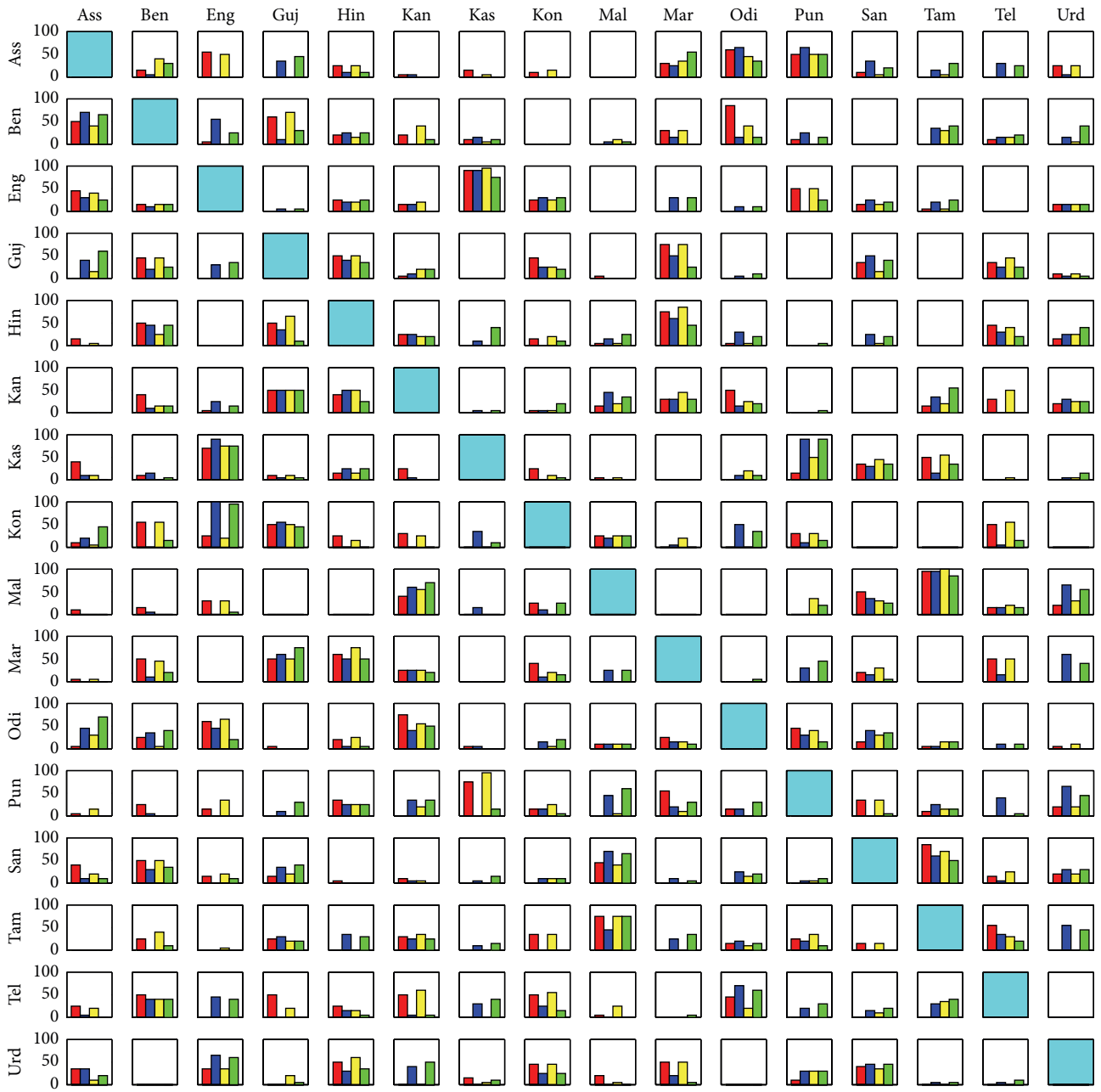
FIGURE 17: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 10-second test samples using the four systems at $X = 20$.

FIGURE 18: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 30-second test samples using the four systems at $X = 20$.
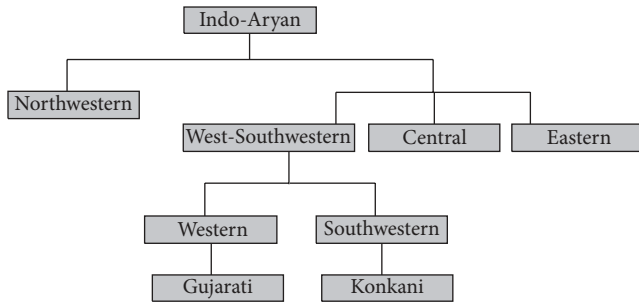
FIGURE 19: Similarity between Konkani and Gujarati.

(ix) Hindi and Urdu show similarity. Hindi shows Y_{Urdu}% in the range 25%–40% while Urdu shows Y_{Hindi}% in the range 30%–60%.

(x) Kannada shows Y_{Marathi}% in the range 30%–45% and Marathi shows Y_{Kannada}% = 25% (consistent 25% using MFCC + SDC + GMM, MFCC + SDC + SVM, and SFCC + SDC + GMM) which again indicates similarity between the two languages.

(xi) Malayalam shows Y_{Sanskrit}% in the range 25%–50% while Sanskrit shows $Y_{\text{Malayalam}}$% in the range 40%–70% indicating similarity between Malayalam and Sanskrit.

(xii) Tamil shows Y_{Telugu}% in the range 30%–55% and Telugu shows Y_{Tamil}% in the range 30%–40%. This indicates similarity between the two languages.

(xiii) Bengali shows Y_{Gujarati}% in the range 30%–70% and Gujarati shows Y_{Bengali}% in the range 25%–45% indicating similarity between the two languages.

It is worth noting that a small increase in X% (from 15% to 20%) reveals a huge amount of information. The results obtained here are valid and most of them can be supported either from a geopolitical point of view or from records of linguists and historians.

The first five points are the same as observed with $X = 15$ (with higher values of Y_J%).

(i) The similarity between Konkani and Gujarati is justified in [18] where it is said that Konkani can be assigned to the South Western group of Indo-Aryan languages with Gujarati as its nearest kin. This is depicted in Figure 19.

(ii) Similarity between Assamese and Odia is a well known fact. Mention about this similarity is found in [19].

(iii) Similarity between English and Konkani can be attributed to the fact that Konkani is mainly spoken in Goa which has been in contact with foreign land from very early period of time either due to colonization or for trade purpose [20, 21]. Besides, Goa is also a tourist place like Kashmir. Since English is used as a universal language by tourists, it can be a cause of intermingling of Konkani and English.

(iv) Similarity between Hindi and Urdu is also a well known fact. Moreover they are mostly spoken in neighboring states. Urdu is mostly spoken in Uttar Pradesh and Andhra Pradesh. Uttar Pradesh and Madhya Pradesh (speaking Hindi) have a long common boundary which is a possible cause of intermingling of the languages. Hindi is also spoken in Haryana (Hindi and Punjabi are the official languages of Haryana) which is another neighboring state of Uttar Pradesh (Figure 15).

(v) Kannada and Marathi show similarity. This can also be justified from the fact that the two are neighboring languages (Figure 15).

(vi) Similarity between Malayalam and Sanskrit has been mentioned a number of times in literature. Eighty percent of Malayalam vocabulary is constituted of Sanskrit (Kerala Sahitya Akademi, http://www.keralasahityaakademi.org/, accessed: June. 2014).

(vii) Similarity between Tamil and Telugu is also justified by the fact that they are neighboring languages spoken in Tamil Nadu and Andhra Pradesh, respectively (Figure 15).

(viii) Similarity between Bengali and Gujarati cannot be justified from existing data on linguistics. But considering the fact that rest of the results are valid and logical, it can be said that this is a new piece of information which linguists are yet to find out.

It is observed that, with $X = 20$, most of the peaks in the graph can be justified. Incoherence in result regarding Konkani and Gujarati, Hindi and Urdu noted with $X = 15$ are resolved. But a few incoherences still exist. For example, Urdu shows Y_{English}% in the range 35%–65%, Sanskrit shows Y_{Tamil}% in the range 50%–85%, and Assamese shows Y_{Punjabi}% in the range 50%–65% whereas the opposite false acceptances are less than 25% in each case; that is, Y_{Urdu}% shown by English, Y_{Sanskrit}% shown by Tamil, and so forth are <25%. It is to be noted that these were present with $X = 15$ also (with lower values of Y_J%). Therefore, it can be concluded that these are nothing but abrupt behavior of the test data which give high score with a particular target language.

6.2.3. Experiment with 25% False Acceptance ($X = 25$). Though increasing Y to a high value is not a good idea, a third experiment is performed with $Y = 15$, that is, with X% = 25% in order to get an idea about how increasing X affects the results.

The graphs in Figures 20, 21, and 22 show the false accepted samples of 3-second, 10-second, and 30-second utterances for all target and nontarget language combination at $X = 25$. The rows, columns, and the bars represent the same things as discussed in Section 6.2.1.

(1) Analysis and Observation. The 3-second graph, in this case, has 1 blank plot, the 10-second graph has 5 blank plots, and the 30-second graph has 17 blank plots. The reason behind the lesser number of blank plots as X is increased is

FIGURE 20: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 3-second test samples using the four systems at $X = 25$.

FIGURE 21: Percentage of false acceptance for all target and nontarget combination (Y_j%) of 10-second test samples using the four systems at $X = 25$.
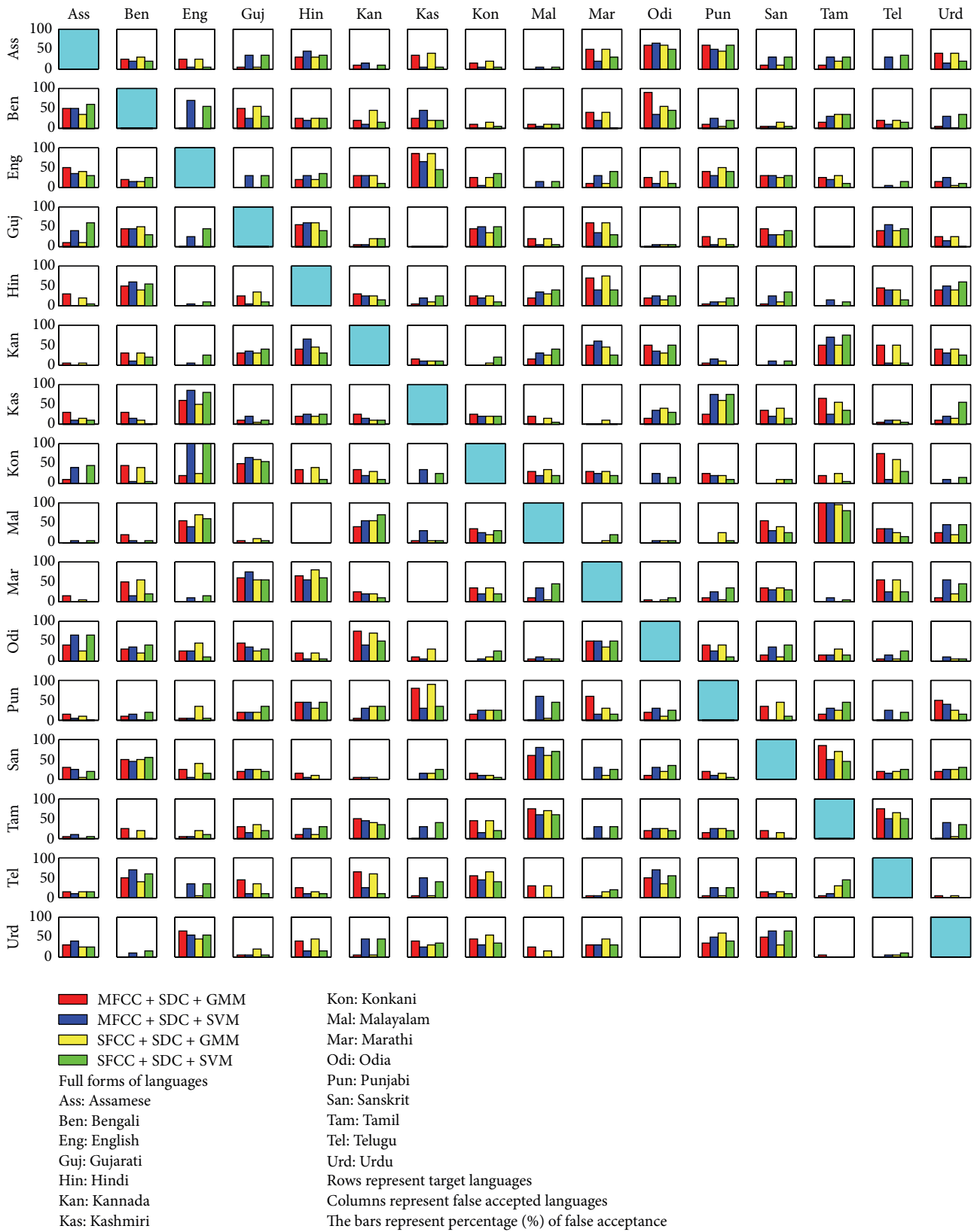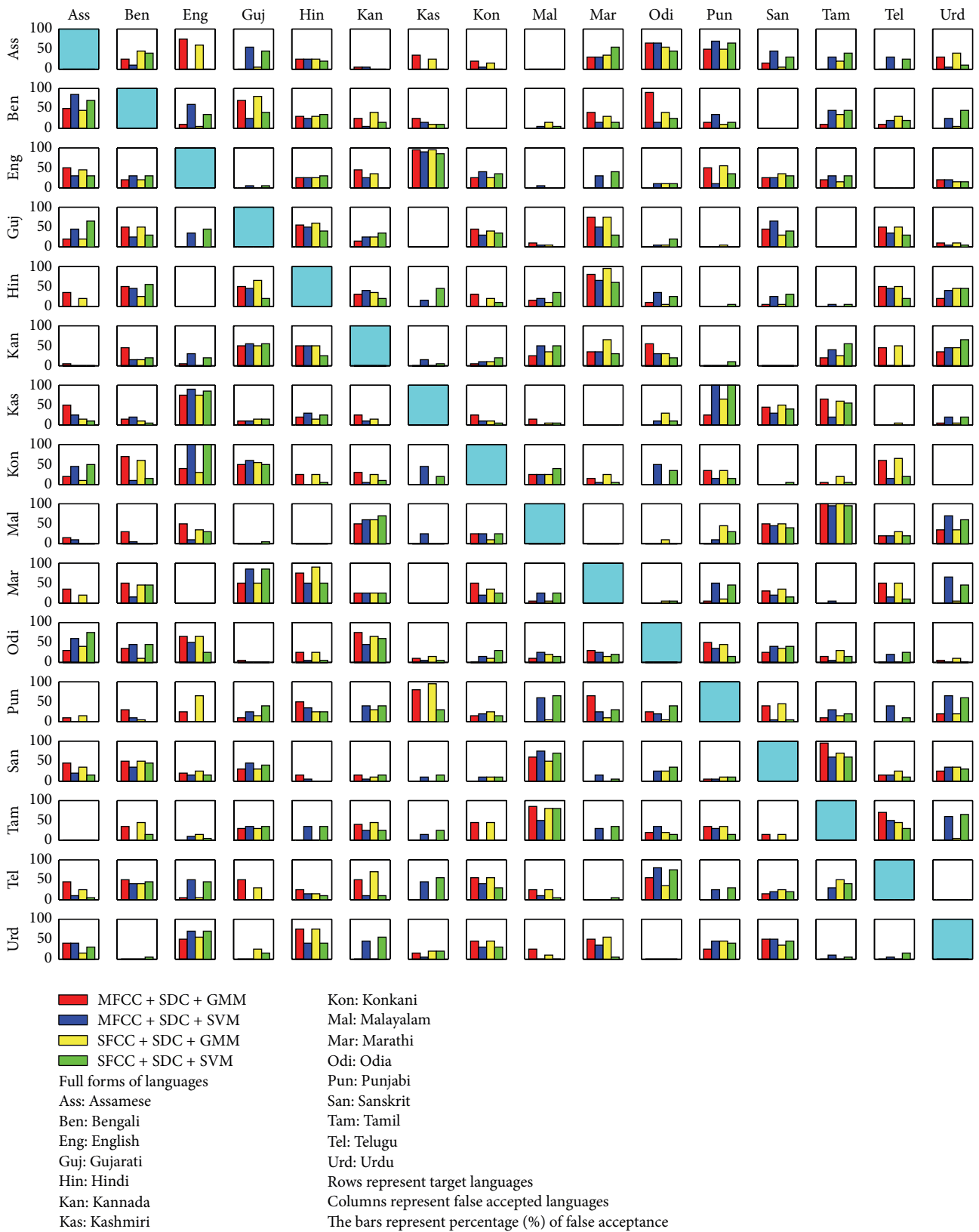
FIGURE 22: Percentage of false acceptance for all target and nontarget combination ($Y_j\%$) of 30-second test samples using the four systems at $X = 25$.

already explained earlier. As in previous cases, only the 30-second graph is used for analysis. The criteria for similarity measurement also remain the same.

The first thirteen points observed with $X = 25$ are the same as those observed with $X = 20$ (with higher values of $Y_J\%$). Apart from these, the other similarities which emerge from $X = 25$ are similarities between Assamese and Bengali, Bengali and Hindi, Bengali and Odia, Gujarati and Kannada, Hindi and Kannada, Kannada and Malayalam, Kannada and Odia, Kannada and Tamil, Kashmiri and Punjabi, Konkani and Malayalam, Odia and Sanskrit, and Sanskrit and Urdu.

Similarity between Assamese and Bengali and that between Bengali and Odia are justified from the fact that they are neighboring languages (Figure 15). Besides, these three languages are called sister languages and similarity among them is a well known fact. Mention about this similarity is found in [19]. Similarity between Kannada and Malayalam and between Kannada and Tamil can be attributed to the fact that the three languages belong to the Dravidian language family. Also, the three are neighboring languages. Kashmiri and Punjabi are also neighboring languages. So similarity between them is expected. Similarity between Odia and Kannada might be due to the fact that Kannada is a Dravidian language and ancient Orissa comprised a large Dravidian speaking region (Government of Odisha, http://orissa.gov.in/). Though political boundaries of Orissa have shrunk in modern times, intermingling of population has resulted in the influence in their language.

Justification of similarity between the other pairs of languages could not be found. The incoherent results seen with $X = 20$ have persisted for $X = 25$ (e.g., Urdu shows $Y_{\text{English}}\%$ in the range 50%–70%, Sanskrit shows $Y_{\text{Tamil}}\%$ in the range 60%–95%, and Assamese shows $Y_{\text{Punjabi}}\%$ in the range 50%–70% whereas the opposite false acceptances are less than 25% in each case).

It is found that, with lower values of false acceptance percentage ($X = 15$), less information is available and a lot of incoherences are observed. As false acceptance percentage is increased by a small amount ($X = 20$), more information becomes available and most of the incoherences are resolved. Increasing false acceptance percentage further ($X = 25$) gives little more information. Incoherences also persist. So $X = 20$ can be considered an optimum choice of false acceptance percentage.

7. Computational Effort

The code was run on an IBM server machine with Intel(R) Xeon(R) 2.40 GHz dual processor and 128 GB RAM. For a single test utterance of 30 seconds, MFCC + SDC + GMM system takes 0.3227 seconds to generate verification result. MFCC + SDC + SVM system takes 0.3522 seconds, SFCC + SDC + GMM takes 0.3381 seconds, and SFCC + SDC + SVM system takes 0.3519 seconds. Similarity between two languages is not measured by a single test utterance because result shown by a single utterance cannot be relied upon. Hence, 20 utterances of a particular nontarget language are used for this purpose. The total time taken by 20 utterances to verify and show similarity is 27.2988 seconds.

For a 10-second test utterance, MFCC + SDC + GMM takes 0.1601 seconds to generate verification result, MFCC + SDC + SVM takes 0.2345 seconds, SFCC + SDC + GMM takes 0.1573 seconds, and SFCC + SDC + SVM takes 0.2231 seconds. The total time taken by 10-second utterances for verification and similarity measurement is 15.5 seconds. For a 3-second utterance, MFCC + SDC + GMM takes 0.0750 seconds for verification, MFCC + SDC + SVM takes 0.1936 seconds, SFCC + SDC + GMM takes 0.0761 seconds, and SFCC + SDC + SVM takes 0.1896 seconds. And the total time taken by 3-second utterances for verification and similarity measurement is 10.6868 seconds.

8. Conclusion

The work consists of two major tasks: (i) development of language verification systems for the major Indian languages and (ii) finding out similarity among the languages using the verification systems.

The experiments show that EER values of the verification systems vary widely from language to language which is attributed to the language discriminating capacity of the languages in comparison to other languages present in the database. As far as similarity among languages is concerned, it is found that, fixing the false acceptance percentage at different levels, similarity among languages can be explored better. The results show that mostly the neighboring languages show similarity. Cases where similarity is seen between nonneighboring languages have been justified with historical and linguistic findings. The method developed in this paper is an easy and efficient way to detect similarity among languages. Similarity not yet known to linguists can be detected and thus it can serve as a useful tool in the study of languages.
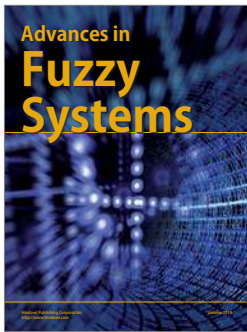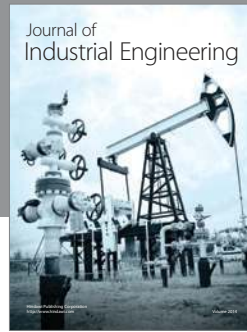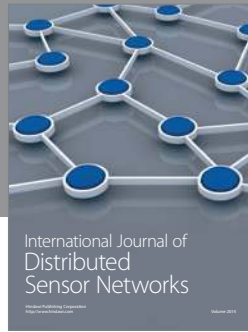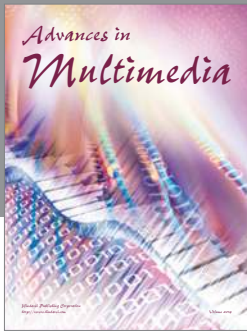
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] K. Kettunen, M. Sadeniemi, T. Lindh-Knuutila, and T. Honkela, "Analysis of EU languages through text compression," in *Advances in Natural Language Processing*, vol. 4139 of *Lecture Notes in Computer Science*, pp. 99–109, Springer, Berlin, Germany, 2006.

[2] S. S. Downey, B. Hallmark, M. P. Cox, P. Norquest, and J. S. Lansing, "Computational feature-sensitive reconstruction of language relationships: developing the aline distance for comparative historical linguistic reconstruction," *Journal of Quantitative Linguistics*, vol. 15, no. 4, pp. 340–369, 2008.

[3] G. Kondrak, "A new algorithm for the alignment of phonetic sequences," in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pp. 288–295, Association for Computational Linguistics, San Francisco, Calif, USA, 2000.

[4] N. Oco, L. R. Syliongka, R. E. Roxas, and J. Ilao, "Dice's coefficient on trigram profiles as metric for language similarity," in *Proceedings of the International Conference Oriental COCOSDA Held Jointly with the Conference on Asian Spoken Language*

Research and Evaluation (O-COCOSDA/CASLRE '13), pp. 1–4, IEEE, November 2013.

[5] A. Bradlow, C. Clopper, R. Smiljanic, and M. A. Walter, "A perceptual phonetic similarity space for languages: evidence from five native language listener groups," *Speech Communication*, vol. 52, no. 11-12, pp. 930–942, 2010.

[6] S. Chakroborty and G. Saha, "Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification," *Speech Communication*, vol. 52, no. 9, pp. 693–709, 2010.

[7] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[8] K. Paliwal, B. Shannon, J. Lyons, and K. Wójcicki, "Speech-signal-based frequency warping," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 319–322, 2009.

[9] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[10] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proceedings of the INTERSPEECH*, Denver, Colo, USA, September 2002.

[11] M. A. Kohler and M. Kennedy, "Language identification using shifted delta cepstra," in *Proceedings of the 45th Midwest Symposium on Circuits and Systems*, vol. 3, pp. 69–72, August 2002.

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

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

[14] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

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

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," Technical Report, DTIC Document, 1997.

[17] N. Dehak, *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*, Ecole de Technologie Superieure, Montreal, Canada, 2009.

[18] S. M. Katre, *The Formation of Konkani Language*, Orient Book Distributors, 1966.

[19] D. Deb, "On case marking in assamese bengali and oriya," *International Journal of Applied Linguistics & English Literature*, vol. 1, no. 2, 2012.

[20] B. W. Diffie and G. D. Winius, *Foundations of the Portuguese Empire, 1415–1850*, University of Minnesota Press, 1977.

[21] B. S. Shastry, *Goa-Kanara Portuguese Relations, 1498–1763*, XCHR Studies Series, no. 8, Concept Publishing Company, 2000.

Advances in
Multimedia

The Scientific World Journal

International Journal of
Distributed Sensor Networks

Journal of
Industrial Engineering

Applied Computational Intelligence and Soft Computing

Advances in
Fuzzy Systems

Modelling & Simulation in Engineering

Journal of
Computer Networks and Communications

Advances in
Artificial Intelligence

Advances in
Computer Engineering

International Journal of
Computer Games Technology

International Journal of
Biomedical Imaging

Advances in
Artificial Neural Systems

Advances in
Software Engineering

Journal of
Robotics

Advances in
Human-Computer Interaction

Computational Intelligence and Neuroscience

International Journal of
Reconfigurable Computing

Journal of
Electrical and Computer Engineering

Hindawi

Submit your manuscripts at
http://www.hindawi.com