Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2578~2587
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.3173

Feature selection for improving Indian spoken language identification in utterance duration mismatch condition

Aarti Bakshi[1], Sunil Kumar Kopparapu[2]
[1]Department of Electronics and Communication Engineering, UMIT, SNDT University, Mumbai, India.
[2]TCS Research, TATA Consultancy Services, Yantra Park, Thane, India

Article Info

Article history:

Received Mar 2, 2021
Revised Jul 2, 2021
Accepted Aug 17, 2021

Keywords:

Artificial neural network
DN FS
openSMILE
Random forest
Support vector machine

ABSTRACT

In spoken language identification (SLID) systems, the test data may be of a sufficiently shorter duration than training data, known as duration mismatch condition. Duration normalized features are used to identify a spoken language for nine Indian languages in duration mismatch conditions. Random forest-based importance vectors of 1582 OpenSMILE features are calculated for each utterance in different duration datasets. The feature importance vectors are normalized across each dataset and later across different duration datasets. The optimal number of duration normalized features is selected to maximize SLID system accuracy. Three classifiers, artificial neural network (ANN), support vector machine (SVM), and random forest (RF), and their fusion, weights optimized using logistic regression, are used. The speech material comprised utterances, each of 30 sec, extracted from the All India Radio dataset with nine Indian languages. Seven new datasets of smaller utterance durations were generated by carefully splitting each utterance. Experimental results showed that 150 most important duration normalized features were optimal with a relative increase in 18-80% accuracy for mismatch conditions. The accuracy decreased with increased duration mismatch.

Corresponding Author:

Aarti Bakshi
Department of Electronics and Communication Engineering
SNDT University
Santacruz (w), Mumbai, Maharshtra, India
Email: aarti.bakshi@kccemsr.edu.in

1. INTRODUCTION

Spoken language identification can be defined as automatically identifying the language in which the person spoke by analyzing, typically a short duration, of the user's speech utterance [1]. Spoken language identification plays a vital role in human-machine voice interactions [2]. A typical requirement is to quickly identify the speaker's language based on a very short utterance so that the user can be provided a personalized service in his or her own language. With the recent development of computer technology, Indian language identification has gained significance in applications such as vernacular call centers to assist customers, services to assist farmers in their regional language, etc. This is because of the need to provide service and communicate to the user in their own language. However, in a vernacular call center, a sufficient dataset of long-duration utterances may be available to train the system. However, equally long duration utterances may not be available; in some cases, the availability of test utterances may be significantly smaller in length (less than 3 sec). This problem is known as duration mismatch. Although there is a number of methods suggested in the literature to enhance the accuracy of short duration utterance, practically the spoken language identification (SLID) system fails to improve the performance for mismatched training and testing utterance durations, especially short and very short duration utterances. That is what training and testing utterance duration mismatch is a long-standing issue in spoken language identification. In the literature, most spoken language identification systems are designed using state-of-the-art i-vector modeling for fixed utterance duration using the total variability subspace technique. It provides an elegant framework for language identification and maps the number of frames to low dimensional vector space. However, the performance of the SLID system drastically degrades with short-duration utterances [3]. Different i-vector-based techniques such as modified prior estimation [4] and exemplar-based representation [5] were used to address the short-duration language identification problem.

Even though these methods reduce utterance duration mismatching in the i-vector space, improvement in the SLID system's performance is not significant. Recently long short-term memory (LSTM) recurrent neural network (RNN) with limited computational resources was used to develop a SLID in [6] RNN outperformed over the i-vector framework [7] for 0.1 to 2.5 sec utterances. The accuracy of 70% is achieved with 2 sec duration, but the accuracy is reduced to 50% for 0.5 sec duration utterance in matched conditions. MFCC and Gammatone frequency cepstral coefficients (GFCC) feature extraction techniques for very short duration utterance (0.8 sec) using bidirectional long short-term memory (BLSTM) neural networks were suggested in [8]. The system's performance has been evaluated using MFCC and GFCC features and a combination of both feature sets for samples of 0.27 sec to 1.5 sec. It has been shown that a 50% accuracy can be achieved for a 0.4 sec duration utterance in matched conditions. However, all the above features used in the literature are for short duration utterances; however, these in the duration mismatched condition reduce the SLID system's recognition accuracy drastically for the short utterance durations (below 3 sec) [7], [8].

To compensate duration mismatch, features such as shifted delta coefficient [9]-[10], eigenfeatures, gaussian mixture model (GMM) [11], [12], total variability i-vector transform [3]-[11], and probabilistic linear discriminant analysis (GPLDA) [11] were used previously. However, the system's performance for short-duration utterances is not significant and cannot address very short-duration mismatched conditions. Although these features carry some information about the speech sample, each feature may not be important for language identification. The selection of relevant or language discriminating features improves recognition accuracy and reduces the computational cost of the system [13], [14]. The main objective of feature selection is to select discriminating features to improve the SLID system's performance. Several feature selection (FS) techniques such as genetic algorithm [15], estimation of distribution algorithm (EDA), and greedy search [15]-[16] have been proposed in the literature. Chowdhury *et al.* [17] presented a grey wolf optimizer (GWO) feature selection algorithm for Indian language identification. In this case, speech samples are converted into spectrogram images, and then all three texture descriptors, namely local binary pattern (CLBP), local binary pattern histogram fourier (LBPHF), and discrete wavelet transform are used to extract the features from spectrogram images. A nature-inspired feature selection (FS) algorithm by combining binary bat algorithm (BBA) and late acceptance hill-climbing (LAHC) feature selection algorithm for Indian languages identification is proposed in the [18]. The MFCC [19]-[21], LPC, i-vector, x-vector, fusion of MFCC + DWT, and MFCC + GFCC feature extraction techniques are used to extract the features from an audio signal. Guha *et al.* [22] reported a hybrid feature selection (FS) algorithm using harmony search (HS) and naked mole-rat (NMR) algorithm labeled as HS-NMR for Indian language identification. Mel-Spectrogram features and relative spectral transform-perceptual linear prediction (RASTA-PLP) features were extracted from audio signals. However, all feature selection algorithms are proposed and employed for fixed duration representation of utterances [17]-[19] for train test conditions. So, it may not be possible to rely on the feature selection algorithm proposed in the literature for a duration mismatch condition as it could reduce the SLID system's language recognition accuracy.

Acoustic features from spoken utterances have often been used as input to a spoken language identification system. This paper uses openSMILE [23] features to represent a speech utterance's overall characteristic. Note that openSMILE features have been used effectively in audio speech emotion recognition. We have used the publicly available openSMILE toolkit [24]. The number of feature sets and their functionals is discussed in detail in [23]. In all our experiments, we use these three popular classifiers, namely artificial neural network with backpropagation algorithm (ANN), OvA multi-class support vector machine (SVM), and random forest (RF). For output score fusion, each class's fused score is calculated as a weighted combination of M classifiers' scores. First, we have built a dataset [25], which consists of a total of 9 languages, of which five belong to the Indo-Aryan family (Assamese (A_S), Bengali (B_N), Gujarati (G_J), Hindi (H_N), Marathi (M_R)), and 4 belong to the Dravidian family (Kannada (K_N), Malayalam (M_L), Tamil (T_M), and Telugu (T_L)). It is to be noted that languages with the same root languages are more likely to be confused. It is a studio-quality news speech recording in 9 Indian languages scraped from the All India Radio portal. The original news recording has been manually segmented into 30 sec duration. A total of 100 speech utterances per language (total of 900) sampled at 16 kHz forms the dataset. Additionally, the initial 30 sec speech utterances have been manually segmented into a smaller duration of 15, 10, 5, 3, 1, 0.5, and 0.2 seconds to form a varying duration dataset. In all, there are eight datasets with different utterance durations.

All speech utterances listen carefully, and any segment with music, silence or unwanted voice has been filtered out. This speech corpus has utterances by newsreaders, both male, and female (equal in number), on varying sets of topics.

In the baseline system, 1582 features using the openSMILE toolkit [24] were extracted from all the utterances across different duration datasets. A five-fold cross-validation method is used to evaluate these 1582 features across different utterance durations at each iteration. However, a classifier trained on one duration (say, 30 sec) when tested with utterances of different duration, namely, 0.2, 0.5, 1, 3, 5, 10, 15 sec, the system's performance degrades irrespective of the type of classifiers. In a nutshell, a mismatch in the duration of the utterance used to train, and the duration of the utterance used to test significantly deteriorate performance. This is seen across all the classifiers when there is a difference between the train and the test utterances. The same observation has been noted for mismatch train-test utterance duration using output score fusion. This issue can be mitigated by selecting proper relevant features with a machine learning framework which can adapt duration mismatched train-test condition. The feature selection approach helps speed up the classifier's training and often improves recognition accuracy because of a selection of discriminative features. For this reason, we introduced the proposed duration normalized feature selection (DNFS) algorithm to evaluate SLID system performance in duration-matched and mismatched conditions. The rest of this paper is organized as follows: The proposed DNFS method for spoken Indian language identification is discussed in section 2. The experimental setup of research work and different experimental results using ANN, SVM, RF, and score fusions are discussed in section 3. Finally, the conclusion is given in the last section.

2. PROPOSED DNFS

Figure 1 shows the proposed model for Indian language identification using normalized feature selection in the duration-matched and mismatched conditions in this study. As discussed earlier, complete 1582 openSMILE features degrade the SLID system accuracy in duration mismatched condition. To improve the SLID system's performance, we focused on the proposed normalized DNFS algorithm's outcome. As shown in Figure 1, five cross-validation techniques are used where 80% of spoken utterances from all datasets are used to train the classifier, while 20% of spoken utterances not used in training are used for testing purposes. Empty circles and triangles indicate the spoken utterances used for training while testing utterances are illustrated as filled circles. Figure 1 indicates the duration-matched condition and duration mismatched condition. Firstly, a complete set of acoustic features is extracted. The set contains both relevant and redundant features. The most discriminative features are selected using the proposed duration normalized feature selection to improve the SLID system's performance in duration mismatched conditions. These discriminative features are used to train the classifiers which predict the correct class in duration mismatched condition. The system's performance is analyzed on our own eight different duration dataset.

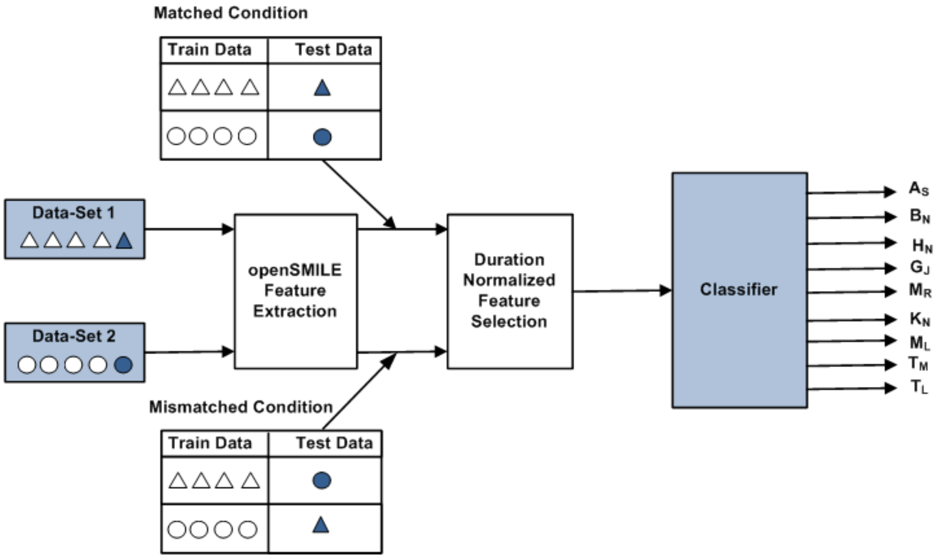

Figure 1. Proposed duration normalized feature selection method for language identification

The selection of most discriminative features helps to speed up classifier training and improve the system's robustness. In mismatch utterance duration, increasing the difference between the duration of the training and test utterances decreases the system's identification accuracy. We propose a duration normalized feature selection algorithm to improve identification accuracy under mismatched utterance duration conditions. Figure 2 shows a flow chart of the proposed duration normalized feature selection. Random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the identification accuracy while controlling the over-fitting.

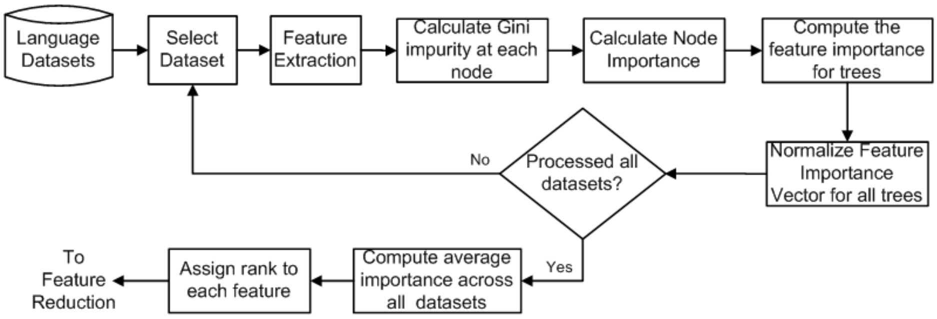
Language Datasets
Select Dataset
Feature Extraction
Calculate Gini impurity at each node
Calculate Node Importance
Compute the feature importance for trees
Normalize Feature Importance Vector for all trees
Processed all datasets?
No
Yes
Compute average importance across all datasets
Assign rank to each feature
To Feature Reduction

Figure 2. Flow chart of proposed duration normalized feature selection algorithm

Let $\mathrm{n_{tree}}$ represent the number of trees and let $\mathrm{n_f}$=1584 and n_c=9 be the number of input features and the number of output (languages) respectively. Let $\mathrm{X_t, Y_t}$ where t represents the duration of the datasets, $\mathrm{n_f}$ −dimension feature vectors and $\mathrm{n_c}$-dimension label vectors, respectively.

For every $\mathrm{X_t, Y_t}$ a separate random forest models each with ensemble of $\mathrm{n_{tree}}$ decision trees is trained. In each decision tree of the random forest model, a random set of features are selected from $\mathrm{n_f}$ features and best possible binary split at each node is performed based on the most important feature to achieve overall $\mathrm{n_c}$-class classification. At node $\mathrm{n_j}$ importance of set of randomly selected features to estimate best possible binary split (i.e. left and right) is calculated as:

$$I_{node}(j) = W_j G_j - W_j^{left} G_j^{left} - W_j^{right} G_j^{right} \tag{1}$$

where W_j is the weighted number of samples at node j its left and right split. G (.) is GINI impurity index calculated as [26]:

$$G = \sum_{i=1}^{n_c} f_i(1 - f_i) \tag{2}$$

where fi is frequency of ith label and n_c $c$lasses
For each decision tree, an importance of feature k is calculated as the ratio of number of nodes with k as most important feature to all nodes in the tree, namely:

$$\mathrm{I_{tree}(K)} = \frac{\sum \mathrm{n_j}:_{\mathrm{j\in=1....n_f}} \overset{\mathrm{argmax}}{} \mathrm{I_{node}}\,(\mathrm{j})=\mathrm{k}}{\sum_{\mathrm{j=1}}^{\mathrm{n_f}} \mathrm{n_j}} \tag{3}$$

Repeat steps for n_{tree} decision trees in random forest model to get feature importance of all n_f features.
Based on random forest model trained using t-sec duration dataset(X_t, Y_t), the feature importance of all n_f features is normalized as:

$$\hat{I_t} = \frac{\hat{I}_{tree}(k)}{\sum_{j=1}^{n_{tree}} \hat{I}_{tree}(k)} \qquad 1 \le k \le n_f \tag{4}$$

Repeate the process to calculate normalized feature importance vector for each of the different segment-length data-sets. Importance of all n_f features is averaged over all different duration datasets as:

$$I_{avg}(k) = \frac{\hat{I}_t(k)}{\sum_{\forall=1} \hat{I}_t(k)} \qquad 1 \le k \le n_f \tag{5}$$

Assign rank to all n_f features such that, the most important feature has rank 1 and the least important feature has rank n_f.

3. RESULTS AND DISCUSSION

3.1. SLID system performance using DNFS

The DNFS method is used to improve the SLID system's performance. In order to verify relevant features, logarithmic power of mel frequency band (logMelBand), spectral pair frequency (IspFreq), mel frequency cepstral coefficient (MFCC), PCM-loudness, shimmerLocal have been used. These features are represented mean, a linear approximation of the contour (linregc), outlier robust signal range max-min (pctlrange), percentile, quartile, standard deviation (stddev), and skewness. This comprises a feature vector of 1582 dimensions for each speech signal. The goal of this phase is to select the most important features using DNFS. It is to be noted that the top 25 and 50 feature sets are related to logMelFreqBand-sma (low-level descriptors smoothed by a moving average filter) and their functional; the top 75 and 100 features include additionally logMelFreqBand-sma-de (1st order delta coefficient of the smoothed low-level descriptor), IspFreq-sma, IspFreq-sma-de and mfcc-sma and their functional. The top 125 features additionally include mfcc-sma -de and their functional, and the top 150, 175, and 200 features contain mfcc-sma-de, pcm-loudness-sma, pcm-loudness-sma-de and shimmerLocal-sma and their functional. The performance of the different feature sets was evaluated using ANN, SVM, and RF classifiers and output score fusion of ANN+SVM and ANN+SVM+RF. Figure 3 shows the performance of the SLID system for 30sec training dataset and 15 and 0.2 sec testing dataset by varying the number of features from 25 to 200 in the step of 25. Figure 3 indicates that the proposed method selects the better feature subset and achieves the highest accuracy over the 15 sec testing dataset.

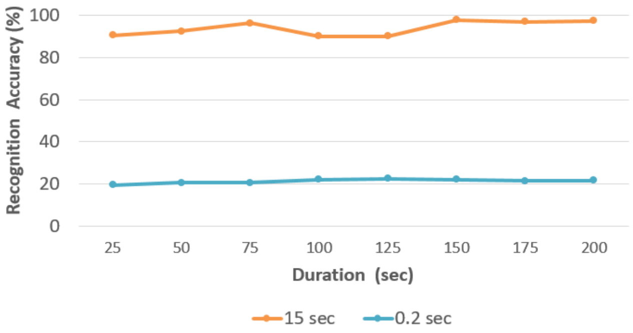

Figure 3. The performance of SLID system in duration mismatched condition for 30 sec-training dataset and 15 and 0.2 sec–testing dataset

The complete set of accuracy for duration mismatch condition using ANN are shown in Table 1. The 30 sec is used to train the classifiers, and the remaining datasets are used for testing. All classifiers performed better for all reduced feature sets. An incremental trend was observed for all classifiers are trained by 30 sec utterance durations. However, for 175 and 200, feature set recognition accuracy started reducing, so 150 feature set is taken as an optimum feature set. The performance SLID system for duration mismatched condition was evaluated by all reduced feature sets (25, 50, 75, 100, 125, 150, 175, and 200), and the best results obtained by 150 optimum feature set is presented in the paper.

Table 1. Accuracy (%) of ANN based SLID system in mismatched condition trained using 30sec datasets with varying number of duration normalized features

Test Dataset (sec)	25	50	75	100	125	150	175	200
30	96.1	99.5	98.6	97.3	94.6	97.7	97.9	97.3
15	90.6	92.4	96.4	90.1	90.1	97.8	97.0	97.4
10	87.3	96.1	96.4	88.4	88.4	97.3	96.4	97.2
5	83.8	94.2	93.5	87.2	87.0	94.2	92.6	93.6
3	70.2	82.8	83.5	77.1	75.6	86.0	85.6	85.8
1	58.4	59.8	61.0	62.0	62.0	63.3	62.7	62.8
0.5	41.7	41.3	43.1	44.6	44.6	44.8	44.0	44.1
0.2	19.5	20.5	20.6	22.1	22.5	22.0	21.3	21.5

As described in section 2, the DNFS is used to alleviate the short utterance duration and the mismatched condition issues in the baseline system. Comparative analysis for varying features according to important values showed that the first 150 most important features are optimum for SLID system under mismatched conditions. Tables 2 to 6 compare the effect of optimum duration normalized features with an entire set of features using three individual and two fusion classifiers for varying utterance duration datasets. The diagonal values depict the matched conditions, while off-diagonal values illustrate mismatched conditions. The comparative analysis indicates an increase in performance for all mismatched conditions with a possible slight decrease in performance for some matched conditions.

Table 2. Comparative performance original baseline (B) and proposed (P) SLID systems in mismatched condition using ANN classifiesr (%)

Train (sec)	Test (sec)															
	30		15		10		5		3		1		0.5		0.2	
	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
30	98.1	97.7	31.5	97.8	31.2	97.3	30.7	94.2	29.2	86.0	20.8	63.3	10.7	44.8	10.4	22.0
15	20.1	90.7	98.6	90.7	33.9	99.7	34.6	96.5	33.1	90.7	31.2	70.1	21.0	49.8	12.4	24.3
10	28.4	98.9	29.0	99.9	98.4	99.3	37.8	98.9	35.3	92.2	36.1	73.6	25.4	51.7	12.6	24.8
5	38.2	98.7	44.2	99.8	44.4	99.9	97.7	99.2	40.4	94.4	40.2	80.2	27.8	56.5	13.1	27.3
3	43.1	98.6	47.6	99.3	46.7	99.5	45.0	99.1	98.3	98.6	47.8	86.1	32.2	63.7	20.2	30.0
1	49.9	97.8	53.1	99.5	51.3	99.4	47.2	99.5	46.8	97.4	96.1	95.5	46.3	81.3	25.3	36.0
0.5	52.3	94.4	58.3	94.9	56.1	94.8	51.4	94.6	48.9	92.5	49.1	94.7	91.5	88.9	36.0	51.3
0.2	36.2	47.8	37.1	45.8	37.2	45.8	38.8	46.4	39.3	45.6	46.1	53.1	61.2	62.6	76.5	75.1

Table 3. Comparative performance original baseline (B) and proposed (P) SLID systems in mismatched condition using RF classifier (%)

Train (sec)	Test (sec)															
	30		15		10		5		3		1		0.5		0.2	
	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
30	99.1	97.2	61.9	96.7	62.2	95.9	58.3	92.3	54.1	85.0	43.7	60.5	32.6	41.6	15.4	24.2
15	57.2	88.2	98.8	88.2	65.8	99.4	62.3	96.5	56.3	88.2	48.9	65.9	36.3	44.9	16.2	24.4
10	63.1	98.9	63.8	99.8	99.2	98.7	64.4	97.6	57.2	89.7	52.5	69.1	36.3	47.3	18.9	25.0
5	68.2	98.9	69.1	99.7	67.7	99.6	97.8	98.0	60.2	92.7	56.1	76.0	40.4	53.0	23.1	26.5
3	69.4	98.7	71.4	98.7	71.3	98.5	66.0	97.7	96.1	96.0	58.0	79.9	42.1	58.3	25.2	28.8
1	73.1	96.2	73.7	97.3	73.9	97.2	67.9	96.9	63.1	92.9	88.8	89.5	43.7	77.0	30.7	37.8
0.5	77.0	89.3	76.3	90.0	76.4	90.0	70.4	90.3	63.8	84.1	60.4	89.0	91.9	81.1	40.2	53.8
0.2	36.8	59.5	36.9	58.7	39.4	58.9	39.6	59.1	39.9	59.1	42.2	62.4	44.7	66.1	69.1	68.0

Table 4. Comparative performance original baseline (B) and proposed (P) SLID systems in mismatched condition using SVM classifier (%)

Train (sec)	Test (sec)															
	30		15		10		5		3		1		0.5		0.2	
	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
30	99.1	97.1	38.9	97.5	36.0	97.1	32.2	95.3	28.1	89.5	24.6	66.9	11.1	41.0	11.1	13.2
15	55.0	92.0	53.0	92.0	50.1	99.2	48.2	98.2	49.2	92.0	35.6	70.9	67.0	42.5	11.2	12.3
10	44.2	98.7	42.9	99.6	98.7	99.0	38.8	98.7	30.4	93.1	28.9	74.2	14.3	45.9	11.5	13.2
5	47.3	98.6	45.8	99.5	38.9	99.5	97.8	98.6	34.5	94.7	31.3	81.2	16.2	52.3	11.7	17.0
3	50.1	98.1	46.3	98.3	43.1	97.8	42.3	97.6	98.6	96.6	33.8	85.3	20.1	59.4	13.2	22.7
1	53.2	78.0	47.9	79.6	48.3	79.5	46.7	78.7	36.9	76.5	87.5	77.8	23.4	64.5	14.3	28.7
0.5	55.0	57.0	53.0	57.9	50.1	57.6	48.2	59.1	49.2	55.5	35.6	59.7	67.0	51.9	21.4	32.2
0.2	18.2	18.6	22.8	23.9	23.9	24.1	23.2	23.9	23.1	25.0	22.2	25.0	20.2	25.4	37.1	31.3

Table 5. Comparative performance original baseline (B) and proposed (P) SLID systems in mismatched condition using ANN+SVM classifier (%)

Train (sec)	Test (sec)															
	30		15		10		5		3		1		0.5		0.2	
	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
30	99.4	98.6	62.5	98.3	61.7	97.8	58.8	95.9	55.0	90.0	44.0	67.6	31.1	45.6	15.7	25.0
15	58.1	92.7	99.0	92.6	65.2	99.8	63.0	98.3	56.9	92.7	45.5	71.5	35.5	49.9	16.5	24.9
10	63.8	99.3	64.3	99.9	99.2	99.3	64.9	99.4	57.9	93.0	52.1	74.9	37.1	52.1	19.2	25.5
5	68.6	99.3	69.8	99.9	67.3	99.3	98.0	99.4	60.6	93.0	56.8	74.9	40.0	52.1	23.4	25.5
3	69.9	99.4	71.8	99.5	70.7	99.8	66.7	99.5	98.8	99.2	58.7	87.1	41.5	64.6	25.5	30.8
1	73.6	98.4	74.1	99.6	73.5	99.6	68.4	99.8	63.7	98.4	96.5	96.1	43.5	81.5	32.0	38.5
0.5	77.6	95.4	76.5	94.9	75.8	95.6	70.8	95.6	64.5	92.9	60.8	95.7	91.9	89.0	40.8	54.4
0.2	37.2	60.0	36.9	59.2	38.2	59.2	39.8	59.6	40.5	59.9	43.0	62.7	44.5	62.2	76.2	76.1

Table 6. Comparative performance original baseline (B) and proposed (P) SLID systems in mismatched condition using ANN+SVM+RF classifier (%)

Test (sec)	Train (sec)															
	30		15		10		5		3		1		0.5		0.2	
	B	P	B	P	B	P	B	P	B	P	B	P	B	P	B	P
30	99.4	99.3	61.3	99.0	62.2	98.7	59.2	96.8	55.4	90.7	44.6	68.8	32.6	46.3	16.4	25.9
15	58.5	93.6	99.0	93.4	65.8	100	63.5	99.0	57.3	93.8	46.1	72.6	36.3	50.4	16.8	25.6
10	64.5	99.8	64.7	99.9	99.2	100	65.3	99.8	57.3	93.2	46.1	75.4	36.3	52.4	16.8	25.9
5	69.3	99.7	70.2	100	67.7	100	98.0	100	61.2	95.8	57.3	82.8	40.4	57.5	23.9	28.3
3	70.3	99.8	72.4	100	71.3	100	68.1	99.9	98.8	99.8	59.2	88.0	42.1	65.1	25.8	31.6
1	74.1	99.0	74.6	99.9	73.9	100	68.9	100	64.2	98.9	96.5	96.8	43.7	82.3	32.5	39.1
0.5	78.0	96.1	76.7	95.4	76.4	96.7	71.2	93.7	64.7	93.7	61.4	96.4	91.9	89.9	41.7	55.1
0.2	37.8	60.4	37.5	59.9	39.4	59.8	40.1	60.2	40.8	60.5	43.6	63.3	44.7	67.4	76.2	77.0

It is noticeable that despite discarding 90% of the initial features, the performance of optimum feature set is comparable to using all features. Incremental trends are observed in recognition accuracy for different utterance durations. The results indicate that recognition accuracy greatly improved with the proposed feature selection algorithm, especially for mismatched train-test conditions while, reducing feature dimensionality. It is to be noted that there is a significant saving in terms of computational time required, as shown in Figure 4. For 30 and 0.2 sec utterances, the computational time required to extract optimum feature set is 2.14 sec and 15.67 msec.

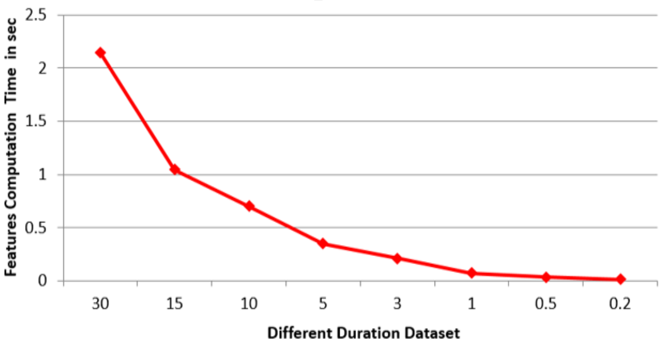

Figure 4. Computation time required for feature extraction

3.2. SLID system performance using mirture of variable-duration utternces

To explore the generalization competency of ML algorithms, we developed a SLID system with a mixture of variable-duration utterances for training and tested using different duration utterances. We used an equal number of training utterances from all eight duration datasets and different languages for training each model. The results of the experiment are shown in Table 7. It can be observed that the system is biased towards higher-duration utterances and provides less than a random chance for very small-duration utterances.

Table 7. Comparative performance SLID system for mix duration train and test dataset (%)

Classifiers	Durations (sec)							
	30	15	10	5	3	1	0.5	0.2
SVM	85.8	85.9	86.0	83.3	77.7	66.7	53.7	30.0
RF	82.8	83.4	83.9	79.5	72.4	57.0	47.8	35.8
ANN	86.6	87.1	87.6	84.3	78.2	66.3	52.6	35.8
ANN+SVM	87.4	87.6	87.8	84.6	78.7	67.1	53.9	35.9
ANN+SVM+RF	87.7	87.8	87.9	84.9	78.9	67.3	53.9	35.6

3.3. SLID system performance comparison with a state-of-the-art system

Das *et al.* [18] selected features using a state-of-the-art relief algorithm to improve the performance of of SLID system. Table 8 shows the accuracy of the SLID system for Indian languages using openSMILE features and relief feature selection algorithm. The number of features were optimised by forward selection method. The best performance, as shown in Table 8, was found for 180 features. Comparative analysis of the proposed and state-of-the-art feature selection algorithm, shows that the proposed algorithm selects more relevant features for improving accuracy at all mismatch conditions.

Table 8. Comparative performance SLID system with sate-of-the-art system (%)

Train (sec)	Test (sec)							
	30	15	10	5	3	1	0.5	0.2
30	97.0	92.7	88.6	85.8	71.7	58.0	50.1	26.7
15	93.4	95.5	83.0	93.6	76.3	69.3	51.6	26.6
10	91.8	91.8	95.4	93.8	79.4	71.7	51.3	27.3
5	91.6	90.6	94.8	93.1	80.8	72.9	55.2	27.2
3	89.1	90.4	94.9	92.6	87.6	78.2	60.0	31.4
1	88.6	88.4	89.3	90.2	85.7	89.1	78.4	39.2
0.5	87.4	87.1	86.1	89.0	71.3	88.6	81.7	53.0
0.2	53.6	52.6	51.9	51.2	51.3	52.4	57.6	69.6

Chowdhury *et al.* [17], used a Grey wolf optimization (GWO) feature selection algorithm, reported 96.6% accuracy using ANN classifier. In comparison, Das *et al.* [18] showed 92.3% and 100 % using the BBA-LAHC feature selection algorithm for the Indic TTS dataset of IIIT Madras and Speech and Vision Laboratory (SVL) IIIT-Hyderabad, respectively for 5 sec dataset. The proposed work yields 100% accuracy using a duration normalized feature selection algorithm for 5 sec dataset in duration-matched condition. A comprehensive study by Sarith Fernando *et al.* [27] used i-vector+ BLSTM to compensate for mismatched duration conditions and reported 66.8% accuracy for the 1 sec dataset. In comparison, the proposed duration normalized feature selection algorithm yielded 68.8% accuracy on 30 sec training dataset and tested with 1 sec dataset using ANN+ SVM + RF output score classifier.

4. CONCLUSION

Indian language identification is crucial for vernacular call centers for automatically routing incoming customer calls to respective language experts. Paper proposed a novel DNFS for spoken language identification using Indian languages for different utterance durations. Each utterance was represented using 1582 features extracted using the openSMILE toolkit. Random forest-based models are developed using reduced features to calculate importance vectors for each feature. The optimum 150 duration normalized features were calculated by averaging over different duration utterance datasets. These features improved SLID system accuracy under training and test duration mismatched conditions, but the system's accuracy reduced with decreasing utterance duration. All experiments were evaluated using the All India Radio dataset developed by us. The dataset was carefully processed to generate eight small-duration databases. Results showed that a combination of duration normalized features improved accuracy for short-duration utterances and mismatched conditions. The drastic improvement exhibits in recognition accuracy from 61.3% to 99.0% accuracy for utterance duration 15 sec and 44.6% to 68.8 % for a very short utterance duration of 1 sec when the classifier is trained with a 30 sec dataset using ANN+ SVM+ RF classifier. Simultaneously, a minor improvement in the recognition accuracy, 16.4% to 25.9% for 0.2 sec duration utterances, was observed. In future work, emphasis will be given to improve the recognition accuracy for very short-duration utterances in mismatched conditions.

REFERENCES

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin and V. Sethu, “Language Identification: A Tutorial,” *in IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82-108, Secondquarter 2011, doi: 10.1109/MCAS.2011.941081.

[2] B. Aarti and S. K. Kopparapu, “Spoken Indian language identification: a review of features and databases,” *Sādhanā*, vol. 43, no. 5, 2018, doi: 10.1007/s12046-018-0841-y.

[3] N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, “Language Recognition via i-vectors and Dimensionality Reduction,” *INTERSPEECH*, pp. 857-860, 2011.

[4] R. Travadi, M. Van Segbroeck, and S. S. Narayanan, “Modified prior i-vector estimation for language identification of short duration utterances,” *INTERSPEECH 2014 15th Annual Conference of the International Speech Communication Association*, pp. 3037-3041, 2014.

[5] M. Wang, Y. Song, B. Jiang, L. Dai and I. McLoughlin, “Exemplar based language recognition method for short-duration speech segments,” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7354-7358, doi: 10.1109/ICASSP.2013.6639091.

[6] R. Zazo, A. LDiez, J.G. Dominguez, D.T. Toledano, and G.J Rodriguez, “Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks,” *PLOS ONE*, vol. 11, no. 1, pp. 1-17, 2016, doi: 10.1371/journal.pone.0146917.

[7] F. Adeeba and S. Hussain, “Native Language Identification in Very Short Utterances Using Bidirectional Long Short-Term Memory Network,” in *IEEE Access*, vol. 7, pp. 17098-17110, 2019, doi: 10.1109/ACCESS.2019.2896453.

[8] F. Adeeba, and S. Hussain, “Acoustic Feature Analysis and Discriminative Modeling for Language Identification of Closely Related South-Asian Languages,” *Circuits, Systems, and Signal Processing*, vol. 37, no. 8, pp: 3589-3604, 2018, doi: 10.1007/s00034-017-0724-1.

[9] S. G. Koolagudi, D. Rastogi, and K. S. Rao, “Spoken Language Identification Using Spectral Features,” *Contemporary Computing. Communications in Computer and Information Science*, pp 496-497, vol. 360, 2012, doi: 10.1007/978-3-642-32129-0_52.

[10] R. K. Vuddagiri, H. K. Vydana and A. K. Vuppala, “Improved Language Identification Using Stacked SDC Features and Residual Neural Network,” *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 210-214, 2018, doi: 10.21437/SLTU.2018-43.

[11] A. Poddar, M. Sahidullah and G. Saha, “Performance comparison of speaker recognition systems in presence of duration variability,” *2015 Annual IEEE India Conference (INDICON)*, 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443464.

[12] A. Bakshi and S. K. Kopparapu, "Spoken Indian Language Classification using GMM supervectors and Artificial Neural Networks," *2019 IEEE Bombay Section Signature Conference (IBSSC)*, 2019, pp. 1-6, doi: 10.1109/IBSSC47189.2019.8972979.

[13] S. M. Kasongo and Y. Sun, "A Deep Learning Method With Filter Based Feature Engineering for Wireless Intrusion Detection System," in *IEEE Access*, vol. 7, pp. 38597-38607, 2019, doi: 10.1109/ACCESS.2019.2905633.

[14] A. Arruti, I. Cearreta, A. Álvarez, E. Lazkano, and B. Sierra, “Feature selection for speech emotion recognition in spanish and basque: On the use of machine learning to improve human-computer interaction,” *PLOS ONE*, vol. 9, no. 10, pp. 1–23, 10 2014, doi: 10.1371/journal.pone.0108975.

[15] B. Wutzl, K. Leibnitz, F. Rattay, M. Kronbichler, M. Murata, and S. M. Golaszewski, “Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness,” *PLOS ONE*, vol.14, no. 7, pp. 1-16, 2019, doi: 10.1371/journal.pone.0219683.

[16] B. Venkatesh and J. Anuradha, “A Review of Feature Selection and Its Methods,” *Cybernetics and Information Technologies*, vol.19, no. 1, pp. 3-26, 2019, doi: 10.2478/cait-2019-0001.

[17] A. A. Chowdhury, V. S. Borkar, and G. K. Birajdar, “Indian language identification using time-frequency image textural descriptors and GWO-based feature selection,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 32, no. 1, pp. 111-132, 2020, doi: 10.1080/0952813X.2019.1631392.

[18] A. Das, S. Guha, P. K. Singh, A. Ahmadian, N. Senu and R. Sarkar, “A Hybrid Meta-Heuristic Feature Selection Method for Identification of Indian Spoken Languages From Audio Signals,” in *IEEE Access*, vol. 8, pp. 181432-181449, 2020, doi: 10.1109/ACCESS.2020.3028241.

[19] D. Sengupta and G. Saha, “Automatic recognition of major language families in India,” *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, 2012, pp. 1-4, doi: 10.1109/IHCI.2012.6481844.

[20] C. C. Bhanja, M. A. Laskar, and R. H. Laskar, “A Pre-classification-Based Language Identification for Northeast Indian Languages Using Prosody and Spectral Features,” *Circuits, Systems, and Signal Processing*, vol. 38, no. 5, pp. 2266-2296, 2019, doi: 10.1007/s00034-018-0962-x.

[21] S. Jothilakshmi, V. Ramalingam and S. Palanivel, “A hierarchical language identification system for Indian languages,” *Digital Signal Processing*, vol. 22, no. 3, pp. 1051-2004, 2012, doi: 10.1016/j.dsp.2011.11.008.

[22] S. Guha, A. Das, P. K. Singh, A. Ahmadian, N. Senu and R. Sarkar, “Hybrid Feature Selection Method Based on Harmony Search and Naked Mole-Rat Algorithms for Spoken Language Identification From Audio Signals,” in *IEEE Access*, vol. 8, pp. 182868-182887, 2020, doi: 10.1109/ACCESS.2020.3028121.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 2794- 2797.

[24] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor," *MM '10: Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp.1459-1462, doi: 10.1145/1873951.1874246, 2010.

[25] A. Bakshi and S. K. Kopparapu, "Spoken Indian Language Identification," *IEEE Dataport*, 2020, doi: 10.21227/xm4q-s210.

[26] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, no. 213, pp. 1-6 2009, doi: 10.1186/1471-2105-10-213.

[27] S. Fernando, V. Sethu, E. Ambikairajah, J. Epps, "Bidirectional Modelling for Short Duration Language Identification," *INTERSPEECH 2017*, 2017, pp. 2809-2813, doi: 10.21437/Interspeech.2017-286.

BIOGRAPHIES OF AUTHORS

Aarti Bakshi is currently pursuing a Ph. D from UMIT, SNDT University, Mumbai. She has completed her BE (Electronics Engineering) from Pune University and ME (Electronics and Telecommunication) from the University of Mumbai. Her current area of interest includes speech processing, language recognition, image processing, and machine learning. She has several papers in international conferences, journals. She is a Life member of ISTE and IETE.

Sunil K. Kopparpu is a Principal Scientist TCS Research, TATA Consultancy Services, Mumbai. He received his Ph. D degree from IIT Bombay. He has several conferences, journals, publications, and patents. He is co-author of the book at Springer brief. His area of interest is the image, speech, and natural language processing. His is a member of IEEE.