Automatic Native Language Identification Using Novel Acoustic and Prosodic Feature Selection Strategies

Chiranjeevi Yarra
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
chiranjeeviy@iisc.ac.in

Achuth Rao MV
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
achuthr@iisc.ac.in

Prasanta Kumar Ghosh
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
prasantg@iisc.ac.in

***Abstract*—We consider the problem of automatic identification of native language (L1) of non-native English (L2) speakers from eleven L1 backgrounds. Analyzing the influence of each L1 pronunciation variabilities on L2 pronunciation, different sets of linguistic units are chosen to compute supra-segmental features by considering the acoustic and prosodic variations within and across these sets. Using these features, we implement a multi-class classifier comprising 55 binary (one versus another) support vector machine (SVM) classifiers. We select optimal set of features for each binary classifier using two feature selection strategies (FSSs) based on Fisher discriminant ratio (FDR). The first strategy considers the features that maximizes the each binary classifier performance. However, the second strategy selects the features by maximizing a multi-class classifier performance for which an algorithm is proposed. Experiments are performed on the ETS corpus of non-native spoken English, comprising 4099 files. When the proposed features along with FSSs are used, the unweighted average recall (UAR) on the test set for each selection strategy is found to be 1.3% and 2.1% (absolute) higher compared to using all features; as well as 3.0% and 3.8% higher than the baseline technique respectively.**

***Index Terms*—Native language Identification, acoustic and prosodic features, feature selection strategies.**

I. Introduction

Speech carries linguistic information as well as para-linguistic information including emotional states and speaker idiosyncrasies [1]. The para-linguistic information could be useful in predicting the person's identity/social class, for example nativity [1] [2]. Most of the existing works on nativity address the problem of language identification when a speaker is speaking his/her own native language (L1) [3]. However, identifying speaker's nativity (L1) remains a challenge when the speaker speaks English (L2) [2]. This problem is similar in many ways to the problem of accent/dialect identification [2]. An automated way of identifying native language/dialect/accent could help in accent morphing and in improving the automatic speech recognition (ASR) accuracy [2]. Arslan et al. proposed an isolated word and phoneme based algorithm to identify three different foreign accents [4]. Piat et al. showed the benefit of prosodic patterns in L1 identification of four different nativities [5]. Biadsy examined frame-based acoustic and phonetic features along with prosodic features for identifying three different foreign accents [6]. Lope et al. used acoustic features from Gaussian mixture model (GMM) super vectors and prosodic features for identifying the nativeness in the TED talks, where speakers have high proficiency in speaking English [7]. Unlike having speakers with high proficiency in speaking English (L2), we consider nativeness (L1) identification from spontaneous speech with speakers having wide variety in their spoken English proficiency.

Most of the existing works in nativeness identification have been addressed based on acoustic and prosodic features computed within different segments recognized using ASR system. However in the case of spontaneous speech with varying accents, obtaining reliable segment boundaries could be challenging [8]. On the other hand, computing the features using erroneous segment boundaries can reduce the performance of nativeness identification [8]. The performance could degrade further for spontaneous speech with different accents when the accent specific pronunciation lexicon is unavailable for each L1 [8] [9]. Obtaining features robust to these variabilities is challenging [10]. We, in this work, select a subset of segments instead of all ASR decoded segments to compute the acoustic and prosodic features. These selected segments contain the words that are chosen based on their frequency of occurrence in the training data. This is because, we observe that frequently occurring words are decoded correctly by the ASR across all L1. We also observe that the ASR decoded text has common language structures within the nativities; to account these structural properties, features based on topic models are proposed [11]. In addition to all the above features, we compute features based on the speech acoustics over the entire sentence instead of using ASR decoded output. Proposed acoustic and prosodic features are combined with the openSMILE features [2] to identify the nativeness.

While different types of features can provide several useful cues for nativity identification, Zhang et al. described that the accuracy could degrade with increasing dimension of the feature vector [10]. So it could be useful to select a subset of features that maximizes the classifier performance. However, a brute forced method for feature selection could be time consuming when the feature vector dimension is large. In order to circumvent this problem, different feature selection strategies (FSSs) have been proposed in the literature [12]. However most of these methods are designed for binary

classifier (BC) and then generalized to multi-class classifier (MCC), because a MCC could be implemented with multiple BCs [13]. In all of these methods, a common subset of features is selected for all BCs; however, different subset of features for different BC could improve the classifier performance further. Based on this hypothesis, we develop a strategy by following the work by Huang et al. [14] using Fisher discriminant ratio (FDR) to select a subset of features for each BC separately. Further, we propose a method to select the best combination of those features at each BC for achieving maximal MCC performance.

We implement the MCC with multiple support vector machine (SVM) BCs, each of which classifies one versus another class. For each BC, features are analyzed separately and ranked based on the FDR. We select top-ranked features for each BC separately from the rank ordered list, that maximizes the overall MCC performance. The features are computed based on acoustic and prosodic variations within the selected subset of ASR decoded segments, along with the features based on the speech acoustics over the entire sentence without considering ASR decoded output. In addition, features based on topic models are proposed. Experiments are performed on the ETS non-native spoken English data. The proposed method performs better than the baseline technique [2] using selected features with FSSs as well as using all features without FSS. When FSS is used, the unweighted average recall (UAR) is found to be highest and it is more than the baseline by 5.2% and 3.8% on the development and the test sets respectively.

II. Database and experimental setup

A. Database

We use the ETS corpus[1] of non-native spoken English [2]. This corpus includes more than 50 hours of speech from 4099 non-native English speakers, with 11 different L1 backgrounds (Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR)). Every language has 358-385 speakers with varying degree of proficiency in English speaking. Each audio recording is about 45 seconds long and has been obtained in the context of the TOEFL iBT assessment. This test is designed to measure a non-native speaker's ability to use and understand English at the university level by asking a question from a predefined set of seven questions. Each recording from the data has been labeled with the speaker's nativity along with prompted question.

B. Experimental setup

We randomly divide the entire data into three sets – 1) 2,459 instances (60%, approximately 30.7 hours) as the training set; 2) 821 instances (19%, approximately 10.3 hours) as the development set; 3) 819 instances (17%, approximately 10.2 hours) as the test set. We use CMUSphinx ASR engine [15] for decoding audio into text and obtain time aligned word boundaries. ASR engine acoustic models are learned from the Wall Street Journal-based CSR corpus [16]. The ASR uses HUB4-trigram language model [17]. Using the time aligned word boundaries, we obtain the prosodic markings by using AuToBI [18] with Boston directions corpus based prosodic models. The obtained prosodic information consists of 8 different accents (NOACCENT, H*, L+H*, !H*, H+!H*, L*, L*+H and L+!H*) along with their time of occurrence, intonation boundaries and sentence boundaries [19]. Using the decoded text, we build the topic models by using Stanford topic modeling toolbox with latent dirchlet allocation [20]. For nativeness classification, we use SVM classifiers using LIBSVM toolkit [13].

[1]Source: Derived from data provided by ETS. Copyright © 2016 by ETS (www.ets.org). Disclaimer: The opinions set forth in this publication are those of the author(s) and not ETS.

III. Proposed approach

Figure 1 shows the block diagram summarizing the steps involved in the proposed method, consisting of three main stages. The first stage involves the computation of three different types of features – 1) acoustic and segmental features, which consider the acoustic variations within the selected set of ASR decoded words as well as within the entire sentence. 2) prosodic features, which consider the variation in the prosodic markings including stress, tone and accent 3) topic model features, which consider the common language structures within each nativity based on ASR decoded text. The second stage implements the MCC comprising multiple BCs and determines the final decision strategy from each BC output. The third stage uses the proposed plus OpenSMILE features and selects the best subset of features for every BC using FDR based rank ordering.

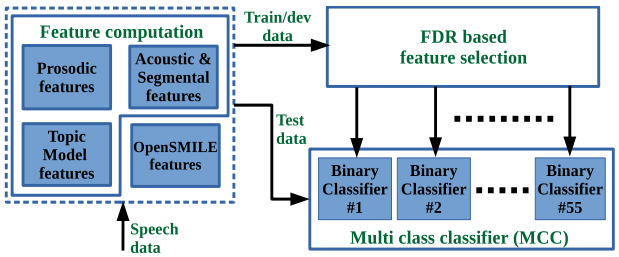

Fig. 1. Block diagram describing the proposed approach

A. Feature computation

1) Acoustic and segmental features (ASFs): The acoustic properties within a segment could vary across the nativity. For example Robert has shown that the vowel formants change across the dialects of US English [21]. Similarly Hönig et al. has used short-time energy variations within segments for identifying the degree of nativity [22]. In this work, instead of using all the segments, we select two sub-sets of segments for computing the features. Within these segments, we hypothesize that the acoustic features would be consistent and less noisy thereby reflects the nativeness specific cues effectively. We believe that these segments are as follows – 1) the words from the ASR decoded text which appear one or more times in every audio file of the training set, called at

least once words (ATOWs); the total count of such words is found to be 19. 2) top L mono-syllabic words (MSWs) from the rank ordered list of all MSWs in the training data; the ranking is done according to their frequency of occurrence. The L is empirically chosen to be 500. Within ATOWs and MSWs, we obtain the features as follows:

- Average duration of each word across a sentence, results in a total of 519 features.
- Statistics of the averaged short-time energy and temporal correlation & selected sub-band correlation [23] (TCSSBC) of each word. The statistics are computed across a sentence and are median, mean, geometric mean, mode, range, skewness, kurtosis, standard deviation, maximum and minimum. This result in a total 1038 features.

In addition to these features, we compute a total of 84 features considering the segments containing silences, filled pauses, cough and noise following above two steps. These together results in a total of 1621, referred to as ASFs#1.

Nativity could also depend on phoneme specific variations. For example, phoneme /w/ is typically replaced by /v/ in Indian English [24]. This, in turn, could cause different nativity specific patterns in the occurrence of various phonetic units. In the absence of accent specific pronunciations in the lexicon, we aim to capture this pattern using 46 phone models [15] and a frame-wise likelihood computation. For this, we consider two sets of models – 1) 138 GMMs (46 phone models ×3 states per phone), 2) 42 GMMs by combining state GMMs with equal weights for all phones excluding breath, cough, smack and noise. Using each set of models, we propose two sets of features: 1) fraction of the frames in an audio file with the highest likelihood in each frame from every model, 2) average likelihood across all frames in an audio file using every model by ignoring likelihoods below a threshold, empirically set to 0.001. This results in 360 ((138+42)×2) features, referred to as ASFs#2.

Speaking rate (number of syllables per second) is another factor for discriminating the nativity. We observe that the Japanese have lower speaking rate than Telugu nativity. For each audio file, we compute the speaking rate contour (analysis window of 2 seconds with a shift of 0.125 seconds) in two methods – 1) counting the number of syllabic nuclei locations obtained directly from the acoustics as proposed by wang et al. [23] 2) counting the number of syllabic segments from ASR output whose centers lie within the analysis window. Similar to the speaking rate contour, a silence rate contour is computed using ASR decoded silence segments with the same analysis interval by following the second method. Considering these three contours, we obtain 300 features from the statistics of the statistics computed within every 2 seconds segment of each contour, referred to as ASFs#3.

2) Prosody based features (PFs): Every L1 has its own prosodic patterns, for example Hindi and Telugu are syllable timed language whereas English is a stress timed language [24]. The voice of the L1 speakers speaking L2 is also influenced by their own native prosodic patterns [4]. To consider this, we compute three sets of features considering prosodic markings obtained from AuToBI [18]. Further, we group the markings into four categories such as syllables with – 1) accent 2) accent and high pitch tone 3) no accent 4) primary stress. The first set of features are obtained based on the duration variations in the four categories computing statistics on 1) segment intervals between two consecutive syllables and 2) syllable duration in each category. This results in a total of 80 ((4×10)×2) features, referred to as PFs#1. The second set of features are obtained from the statistics across a sentence computed on the ten statistics of TCSSBC and short-time energy separately within all the syllables as well as within the syllables of MSWs belonging to all the four categories. This results in a total of 1600 ((((10×10)×2)×2)×4) features, referred to as PFs#2. The third set of features obtained from the statistics of the statistics computed separately within every 2 seconds segment of four speaking rate contours, which are computed as mentioned in section III-A1 considering the syllables belonging to each category. This results in a total of 400 features, referred to as PFs#3.

3) Topic model based features (TMFs): We observe that the words spoken by the L1 speakers for a question, which is asked during the test, have a common pattern within one nativity and different patterns across different nativities. To explore these properties we built topic models specific to each L1 and question (overall 11x7=77 cases). In the absence of training data for two cases, a total of 75 topic models are built using the ASR hypotheses from the training data. From these topic models we obtain 75 features comprising likelihoods from each model for an unknown spoken text.

Combining acoustic, prosodic and topic model based features, a total of 4436 proposed features are combined with the baseline (openSMILE) features [2], resulting in a 10809-dimensional feature vector which is used for classification. Schuller et al. have used the openSMILE features for native language identification task [2].

B. Multi-class classifier

We use SVM classifier with linear kernel for identifying 11 nativities. The linear kernel has been found to perform better than other choices of kernel [13], when the feature vector dimension (10809) is greater than the number of training instances (2459). In general SVM classifiers are binary and discriminative. However, an effective multi-class SVM classifier can be designed based on pair-wise coupling strategy [25]. In this work, we use the pair-wise coupling strategy; based on this, the 11 class classification problem is broken into 55 ($^{11}C_2$) BC problems. In this strategy, each BC classifies/votes every feature vector to one of the nativities along with a confidence score. The final decision on the nativity is made based on a majority voting scheme. However, in the case of a tie the nativity with maximum total confidence score is selected [25]. By implementing this pair-wise strategy, we adapt each BC separately for achieving the best MCC performance.

C. Fisher Discriminant Ratio based feature selection

When the dimension of the feature vector is larger than the number of training instances, the SVM classifier tends to overfit the data [26]. Hence, the classifier can perform poorly on unseen data [10]. However eliminating features without any specific strategy could also degrade the classifier performance because the features which discriminate well between a pair of nativities might not work well with other pair of the nativities. So a feature selection strategy (FSS) is required. By following the work by Huang et al. [14], we propose FSSs based on Fisher discriminant ratio (FDR). The FDR between class i and j for each feature is computed by following the work proposed by wang et al. [27]; and this value is high for the features which discriminate well between two classes. So most discriminative features could be selected for a BC by ranking the features based on the FDR values.

Algorithm 1 FSS-2 algorithm – input=$\mathscr{S}_I^k$: Optimal set of features selected in FSS-1 for k-th BC, output=$\mathscr{S}_{II}^k$: Optimal set of features selected for k-th BC; $\forall k : 1 \leq k \leq 55$.

$\triangleright$ N : Incrementing steps.
$\triangleright$ $\mathscr{R}_{\{1:m\}}^k$: Set of top m ranked features from $\mathscr{R}^k$.
$\triangleright$ MCC($\mathscr{X}$): Function that returns the overall MCC performance for the given set ($\mathscr{X}$) of features for each BC.
$\mathscr{S}_{II}^k \leftarrow \mathscr{S}_I^k$,$\forall k : 1 \leq k \leq 55$;
for each $k \in \{1 : 55\}$ **do**
  for each $i \in \{0 : N : 10809\}$ **do**
    $\mathscr{X} \leftarrow \{\mathscr{S}_{II}^1, .., \mathscr{S}_{II}^{k-1}, \mathscr{R}_{\{1:i\}}^k, \mathscr{S}_{II}^{k+1}, .., \mathscr{S}_{II}^{55}\}$
    A_i = MCC($\mathscr{X}$);
  end for
  i^* = $\arg\max_i A_i$; $\mathscr{S}_{II}^k \leftarrow \mathscr{R}_{\{1:i^*\}}^k$;
end for

Considering this, we select the features for each BC using two different strategies. In the first strategy (FSS-1), each BC is trained separately, where we find the top-ranked features from a ranked order features set ($\mathscr{R}^k$ $\forall k : 1 \leq k \leq 55$) by maximizing the BC performance on the development set for every increment of N features. In FSS-1, we hypothesize that the selected features which maximizes individual BC performance could result in better MCC performance. In contrast to this, in the second strategy (FSS-2), top-ranked features are selected for each BC that maximizes the overall MCC performance instead of individual BC performance as in FSS-1. The maximal performance is evaluated on the development set. The detailed steps involved in FSS-2 are provided in Algorithm 1. Using these optimal choices of features, the 11 nativities classification is performed.

IV. Experimental results

A. Hyper parameter optimization

We consider classification accuracy and unweighted average recall (UAR) as objective measures for SVM based BC and MCC respectively. We normalize every feature vector by the mean and the variance of the training set before training as well as performing classification on the test and development sets. The complexity parameter (C) of each BC is learnt on the development set using the model learnt on the training data by searching over the values 1E-5, 1E-4, 1E-3, 1E-2, 1E-1, 1 and 10. The optimized range of C for all 55 BCs is found to be from 1E-3 to 1E-1. The optimal feature set for each BC is separately learnt on the development set with N=1000. Figure 2 shows the selected number of top-ranked features by FSS-1 and FSS-2 for each BC. From the figure it is observed that the optimum feature set sizes for all BCs found within FSS-1 and FSS-2 are not identical. This observation is consistent across the feature set sizes found in FSS-1 and those in FSS-2. For example, in FSS-1 the optimal feature set sizes are found to be 1000 and 3000 for FRE vs ARA and SPA vs ARA BCs respectively, while, in FSS-2 they are 6000 and 5000 respectively. This indicates that the optimum feature set changes from FSS-1 to FSS-2 as well as across BCs within a selection strategy.

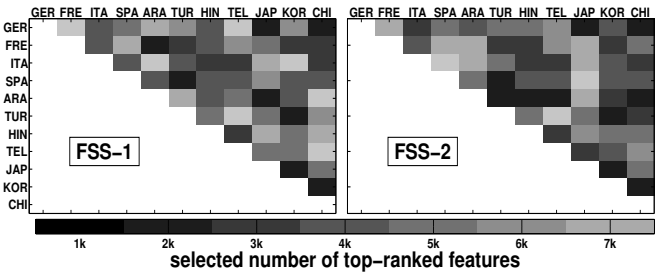

Fig. 2. Plot describing the selected number of top-ranked features for each BC by using FSS-1 and FSS-2

B. Results and discussion

Table I shows the UAR on the development (devel) and the test set for the baseline technique and the proposed method with three different conditions – without using FSS (Wo-FSS); with FSS-1 and with FSS-2. From the table, it is observed that the proposed method performs better than the baseline for all cases on devel and test sets. On the development and test sets, the proposed method without FSS (Wo-FSS) is higher than the baseline by 0.9% and 1.7%(absolute) respectively. This indicates that the proposed features are complementary to the baseline features. Also when the proposed FSS-2 is used, the accuracy increases by 4.3% & 2.5% and 2.1% & 0.8% (absolute) compared to Wo-FSS & FSS-1 for development and test sets respectively. This indicates that FSS improves the classification performance.

TABLE I
UAR comparison between the baseline and the proposed methods on the development and test sets

%	Baseline	Wo-FSS	FSS-1	FSS-2
Devel	42.4	43.3	45.1	47.6
test	41.3	43.0	44.3	45.1

Table II shows confusion matrix among all 11 native languages on the test set. From the table, it is observed that the accuracies for FRE, ITA, TUR, HIN, JAP, KOR and CHI (marked in blue) improve over the baseline; and the highest improvement is found to be 13.5% (absolute) for JAP. This could be because JAP has lower speaking rate compared to

TABLE II

CONFUSION MATRIX COMPUTED ON THE TEST SET USING THE PROPOSED METHOD FSS-2 (DIFFERENCES BETWEEN CONFUSION MATRICES OF THE FSS-2 AND BASELINE ARE SHOWN IN BRACKETS).

%	GER	FRE	ITA	SPA	ARA	TUR	HIN	TEL	JAP	KOR	CHI
GER	**52.0(-2.7)**	5.3(1.3)	4.0(0.0)	4.0(-6.7)	8.0(0.0)	9.3(2.7)	1.3(0.0)	0.0(-2.7)	4.0(2.7)	8.0(1.3)	4.0(4.0)
FRE	8.2(-2.7)	**35.6(2.7)**	12.3(1.4)	8.2(-1.4)	9.6(1.4)	4.1(-2.7)	4.1(0.0)	0.0(-1.4)	9.6(4.1)	4.1(-1.4)	4.1(0.0)
ITA	6.7(-2.7)	17.3(4.0)	**49.3(4.0)**	10.7(-1.3)	9.3(5.3)	1.3(-5.3)	0.0(0.0)	1.3(1.3)	1.3(1.3)	0.0(-4.0)	2.7(-2.7)
SPA	5.3(-5.3)	9.2(0.0)	11.8(3.9)	**25.0(-1.3)**	7.9(2.6)	6.6(-2.6)	3.9(2.6)	5.3(-1.3)	7.9(2.6)	10.5(-1.3)	6.6(0.0)
ARA	4.1(-4.1)	10.8(0.0)	6.8(-6.8)	1.4(0.0)	**42.5(-1.4)**	8.1(0.0)	5.4(1.4)	5.4(-1.4)	5.4(-2.7)	5.4(0.0)	1.4(0.0)
TUR	6.7(4.0)	2.7(-5.3)	4.0(-1.3)	1.3(-5.3)	14.7(-2.7)	**49.3(12.0)**	1.3(0.0)	0.0(-6.7)	5.3(1.3)	8.0(2.7)	6.7(1.3)
HIN	1.4(-1.4)	2.7(1.4)	4.1(1.4)	2.7(0.0)	4.1(-2.7)	2.7(-1.4)	**47.3(2.7)**	31.1(0.0)	0.0(0.0)	0.0(-1.4)	4.1(1.4)
TEL	1.4(-1.4)	0.0(0.0)	1.4(-1.4)	2.7(-2.7)	5.5(2.7)	1.4(-1.4)	27.4(6.8)	**54.8(-1.4)**	1.4(0.0)	1.4(0.0)	2.7(-1.4)
JAP	1.4(-1.4)	2.7(-2.7)	1.4(0.0)	5.5(-8.2)	8.2(2.7)	2.7(0.0)	4.1(2.7)	0.0(0.0)	**45.9(13.5)**	16.4(4.1)	15.1(4.1)
KOR	1.4(-1.4)	2.7(0.0)	2.7(1.4)	10.8(-2.7)	2.7(-4.1)	2.7(1.4)	2.7(-1.4)	2.7(0.0)	12.2(-1.4)	**39.2(1.4)**	20.3(6.8)
CHI	6.5(-2.6)	2.6(-2.6)	0.0(-1.3)	6.5(-1.3)	5.2(3.9)	2.6(-6.5)	3.9(-1.3)	1.3(-1.3)	7.8(-3.9)	9.1(5.2)	**54.5(11.7)**

TABLE III

COMPARISON OF UARS OBTAINED USING PROPOSED THREE CATEGORIES OF THE FEATURES AND THEIR COMBINATIONS ALONG WITH THE BASELINE FEATURES ON TEST SET UNDER THREE CONDITIONS WO-FSS, FSS-1 AND FSS-2.

	Feature Type	UAR (%)	Selected Features (%)		UAR-drop (%)	
			FSS-1	FSS-2	FSS-1	FSS-2
1	ASFs#1	29.5	25.8(18.0)	22.2(17.0)	4.0	6.3
2	ASFs#2	30.3	25.1(12.4)	23.2(12.1)	2.9	4.4
3	ASFs#3	17.1	26.8(17.2)	24.4(17.8)	2.4	3.1
4	PFs#1	16.7	31.1(20.6)	26.6(22.0)	1.9	3.5
5	PFs#2	20.1	32.6(19.5)	28.5(19.2)	3.9	5.1
6	PFs#3	15.4	23.0(15.9)	20.1(16.3)	2.9	4.2
7	ASFs	33.6	25.9(16.4)	22.6(15.6)	3.6	5.9
8	PFs	21.0	29.9(18.1)	26.1(18.1)	2.8	6.2
9	TMFs	7.7	8.8(13.1)	7.0(12.1)	1.8	3.2
10	7+8	33.0	27.5(16.8)	24.0(16.4)	6.9	8.6
11	7+9	33.1	25.4(16.2)	22.2(15.4)	4.6	5.9
12	8+9	20.6	29.0(17.8)	25.3(17.8)	4.7	5.2
13	7+8+9	28.0	27.2(16.7)	23.7(16.3)	7.7	8.2
14	OpenSMILE	41.3	39.2(18.6)	35.7(17.6)	8.4	10.2
15	7+14	44.6	35.2(17.6)	31.8(16.6)	9.9	10.1
16	8+14	38.9	37.1(18.2)	33.5(17.2)	11.7	12.3
17	9+14	41.6	38.8(18.5)	35.4(17.5)	9.0	10.4
18	7+8+14	43.3	34.3(17.5)	30.8(16.6)	28.3	30.4
19	7+9+14	44.2	35.0(17.6)	31.6(16.5)	10.0	10.1
20	8+9+14	38.6	36.8(18.1)	33.3(17.1)	12.1	13.1

other languages so the proposed features based on speaking rate helps in better discrimination of JAP from other classes. The accuracies of GER, SPA, ARA and TEL (marked in red) reduce from the baseline; and the highest reduction is found to be 2.7% for GER. In the table, GER, which is a stressed timed language, is highly confused with its geographically close syllable timed languages, namely, FRE, ITA and SPA. So we hypothesize that the prosody based features could discriminate GER from these languages. However, a drop in UAR for these languages could be due to noisy prosody marking because of poor ASR and/or AuToBI. From the table, it is also observed that most of the confusions are among geographically close languages – HIN-TEL; JAP-KOR-CHI; GER-FRE-ITA-SPA, expect ARA-TUR. Therefore the accuracies within these groups could be improved by adding most discriminative features among these groups.

We, further, investigate the performance of the classifier on the test data using different sub-set of features from the total set of features. The sub-sets are considered from three sub-categories of ASFs and PFs and 14 sets from all possible combination of ASFs, PFs, TMFs and OpenSMILE, excluding the combination containing all 10809 features. For the comparison, we consider three performance measures – 1) UAR under Wo-FSS 2) average percentage (standard deviation in brackets) of selected features across all BCs and 3) UAR-drop. The percentage is the ratio of selected number of features to total number of features in each sub-set. The UAR-drop is computed separately for FSS-1 and FSS-2. It is the difference between the UAR in Table I and a UAR computed with MCC using selected features in each sub-sets of features separately. Considering this measure, we compare the contribution of each sub-set in the overall UAR; the sub-set with the highest UAR-drop indicates the most significant sub-set of features among all.

Table III shows the three measures obtained from each sub-set of features. From the table, it is observed that the UAR obtained from each sub-set is not always higher than the UARs obtained from its sub-partsall. For example, UAR obtained with ASFs is found to be 33.6%, while that, with ASFs plus PFs is 33.0%. However, it is also observed that the percentage of selected features is more than 20% in all sub-sets except TMFs with FSS-1 and FSS-2. This indicates that the selected features in the sub-sets are complementary to each other. From Table I and III, it is observed that FSS-2 achieves highest UAR with lesser selected features in all sub-sets compared to FSS-1. This indicates FSS-2 is better than FSS-1. It is also interesting to observe that the standard deviations in the percentage of selected features (show in brackets in Table III) are high in all sub-sets. This indicates that the selected features are largely vary among all BCs. This observation is consistent with the observations made from Figure 2.

Among TMFs and sub-categories of ASFs & PFs, least UAR and UAR-drops are achieved using TMFs. This could be because of the errors in the ASR due to out of vocabulary (OOV) words. This in turn causes inconsistent topic models and provides less discrimination across the nativities. This less discrimination is evident from the lower contribution of

TMFs in the selected features by FSS-1 and FSS-2. Highest UAR and UAR-drops are obtained using ASFs#2 and ASFs#1 respectively. This indicates that the features deduced from phoneme specific variations and based on ATOWs & MSWs carry maximal cues for nativities. However, it is observed that the UAR is lower and the UAR-drops are higher in ASFs#1 compare to ASFs#2. This could be due to elimination of noisy features from ASFs#1 with FSSs. Among all sub-sets, it is observed the highest UAR-drops (28.3% and 30.4%) are obtained from 18-th sub-set (ASFs+PFs+OpenSMILE) under both FSS-1 and FSS-2. However, UAR-drops (using FSS-1 & FSS-2) obtained using ASFs+PFs and OpenSMILE are found to be 6.9% & 8.6% and 8.4% & 10.2%. This indicates that ASFs+PFs features are more complementary to OpenSMILE features and hence together achieve the highest UAR-drop.

Further, it is observed that the UAR-drop obtained from PFs is lower than that from PFs#2 and PFs#3 under FSS-1. However, under FSS-2, the UAR-drop from PFs is higher than that from those two. This indicates that FSS-2 selects the features better than the FSS-1. Incontrast to this, UAR-drops obtained from ASFs are lower than the ASFs#1 under both FSS-1 and FSS-2. This indicates that both the FSSs fail to select the optimal set of features. This could be because of higher incremental step size ($N = 1000$). However, we observe that the low value of N increases the computation complexity. Hence, it is required to choose N by balancing computational complexity as well as UAR.

V. Conclusions

We propose an FDR based acoustic and prosodic feature selection strategy for MCC to identify the native language of the speaker. We implement 11-class classifier by combining 55 SVM BCs using pairwise strategy. We propose a method to learn the parameters and select the features for each BC separately. We use FDR for selecting a subset of the features, which are proposed based on acoustic, prosodic and linguistic properties. Experiments with ETS corpus of non-native spoken English reveal that the UAR of the proposed method improves over the baseline technique. This improvement is highest when the proposed FSS is used. Further investigations are required to reduce confusion among the geographically close languages by adding robust features. Future works also include the selection of features jointly in FSS unlike ranking them individually as done in the present work.

References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language-state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, Sincerity & Native language," *Proc. of INTERSPEECH, San Francisco, USA*, pp. 2001–2005.

[3] A. F. Martin, C. S. Greenberg, J. M. Howard, D. Bansé, G. R. Doddington, J. Hernández-Cordero, and L. P. Mason, "NIST language recognition evaluation plans for 2015," *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 3046–3050, 2015.

[4] L. M. Arslan and J. H. Hansen, "Language accent classification in american English," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[5] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," *Proc. of INTERSPEECH, Brisbane, Australia*, pp. 759–762, 2008.

[6] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011.

[7] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for TED talks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5672–5675, 2011.

[8] M. Hasegawa-Johnson, J. Cole, P. Jyothi, and L. R. Varshney, "Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications," *Laboratory Phonology*, vol. 6, no. 3-4, pp. 381–431, 2015.

[9] P. Jyothi and M. Hasegawa-Johnson, "Transcribing continuous speech using mismatched crowdsourcing," *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 2774–2778, 2015.

[10] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognitioin," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5672–5675, 2011.

[11] S. Jarvis, M. Paquot *et al.*, *Native language identification*. The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 2015.

[12] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[14] P. X. Huang and R. B. Fisher, "Individual feature selection in each one-versus-one classifier improves multi-class SVM performance," *Available at http://homepages.inf.ed.ac.uk/s1064211/thesis/icpr14.pdf: last accessed on 11 November, 2016.*

[15] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," *Sun Microsystems Inc., Technical Report SML1 TR2004-0811*, 2004.

[16] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," *Proceedings of the workshop on Speech and Natural Language*, pp. 357–362, 1992.

[17] sourceforge2015, "https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/."

[18] A. Rosenberg, "AuToBI-a tool for automatic ToBI annotation." *Proc. of INTERSPEECH, Makuhari, Japan*, pp. 146–149, 2010.

[19] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody." *International Conference on Spoken Language Processing*, vol. 2, pp. 867–870, 1992.

[20] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 457–465, 2011.

[21] R. Hagiwara, "Dialect variation and formant frequency: The american English vowels revisited," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 655–658, 1997.

[22] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," *Proc. of ISADEPT–International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 21–30, 2012.

[23] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[24] P. Sailaja, *Indian English (Dialects of English)*. Edinburgh University Press, 2009.

[25] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[26] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[27] S. Wang, D. Li, Y. Wei, and H. Li, "A feature selection method based on Fisher discriminant ratio for text sentiment classification," *Web Information Systems and Mining*, pp. 88–97, 2009.