FEATURE SELECTION FOR ARABIC MISPRONUNCIATION DETECTION BASED ON SEQUENTIAL FLOATING FORWARD SELECTION AND DATA MINING CLASSIFIERS

M. Maqsood, H. A. Habib[*] and T. Nawaz

Software Engineering Department, University of Engineering and Technology Taxila, Pakistan
[*]Computer Science Department, University of Engineering and Technology Taxila, Pakistan
Corresponding author's email: muazzam.maqsood@uettaxila.edu.pk

ABSTRACT: Feature selection process is used to reduce the feature vector length and identify the discriminative features. Many acoustic-phonetic features including Mel-Frequency Cepstral Coefficient (MFCC), Energy, Pitch, Zero-crossing, spectrum were tested individually for Arabic mispronunciation detection using three classifiers; Random Forest, Bayesian classifier, and Bagged Support Vector Machine (SVM). The results for Bagged SVM were better than the other two classifiers. Top three individual features with highest accuracies were identified for each isolated Arabic consonant. To validate the results, a modified form of Sequential Floating Forward Selection (SFFS) process was used. Results showed that MFCC along with its first and second derivatives, energy, spectrum, and zero-crossing were the most suitable acoustic features for Arabic mispronunciation detection system. The proposed approach provided an average accuracy of 94.9% which was better than the previous best 92.95% for Arabic consonants.

Keywords: Feature selection, Acoustic-phonetic features, Sequential Floating Forward Selection and Mispronunciation detection.

(Received 25-05-2016 Accepted 09-12-2016).

INTRODUCTION

Pronunciation training and automatic mispronunciation detection have received a lot of attention in last decade due to advancement in artificial intelligence and machine learning. Mostly pronunciation scoring and mispronunciation detection are considered as a single task. In fact, both these tasks are different from each other and serve different purposes. Pronunciation scoring only rates one's proficiency in the language while mispronunciation detection points out the specific mistakes in pronunciation. Therefore, mispronunciation detection is more useful than the pronunciation scoring (Troung *et al.*, 2006 and Witt *et al.*, 2000).

Mispronunciation detection systems can be divided into two categories; Confidence Measure (CM) and Acoustic-Phonetic Feature. Most of the mispronunciation detection systems are developed using CM because of well-defined mathematical models and Automatic Speech Recognition (ASR). Speech can be represented in the form of a signal with many acoustic features. These features include MFCC, Pitch, Fundamental frequency, formants, zero-crossing and spectrum etc. Any change in a pronunciation should ideally be representable from the acoustic-phonetic features rather than CM scores. Therefore, mispronunciation detection problem can be designed more comprehensively using Acoustic Phonetic Features (APF). It is highly desirable to find the most suitable acoustic features so that APF based mispronunciation detection can be designed (Wei *et al.*, 2008).

The objective of this study is to identify the most discriminative acoustic-phonetic features for each specific Arabic consonant, using a modified form of Sequential Floating Forward Selection (SFFS). To cross-validate the suitability of the identified features, three machine learning classifiers i.e. Random Forest, Bayesian, and Bagged SVM are used for mispronunciation detection. The selected features show promising results for mispronunciation detection.

MATERIALS AND METHODS

A general framework for mispronunciation detection included the isolated Arabic consonants as input and then feature selection process was performed manually by using SFFS (Fig-1). Then correlation between these two methods was calculated and discriminative acoustic features were identified for each isolated Arabic consonant.

Dataset: In this study, a large Arabic dataset from 200 Pakistani speakers was collected for 28 Arabic consonants. These speakers included both, highly proficient speakers as well as new learners. The recording was carried out in an office environment, these recording conditions were far from ideal. Five Arabic language experts were asked to label this dataset and assign them native or non-native classes. A class of native or non-native pronunciation for a particular phoneme was assigned if at least three of the judges agreed on the same label. The details of the number of speakers, the number of correct and incorrect phonemes are presented in Table-1.

Feature Extraction: A relatively large acoustic feature vector comprised of Pitch, Entropy, and zero-crossing, 14 MFCC coefficients with its first and second delta. Energy, spectral features, and statistical features were extracted. First and second derivative for some of these features were calculated (Table-2).

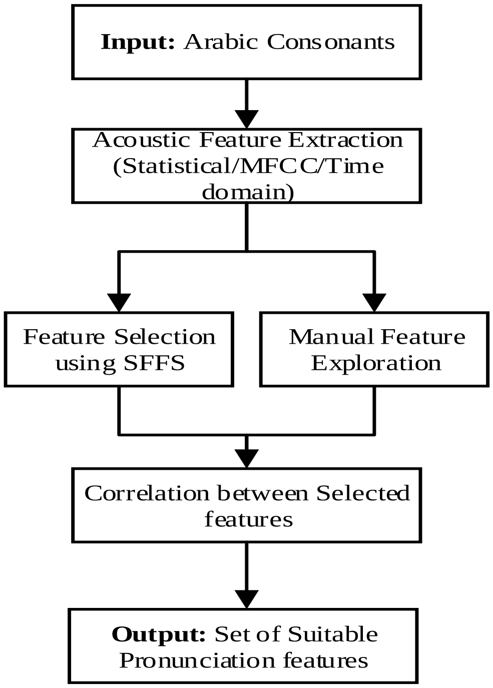
Input: Arabic Consonants
Acoustic Feature Extraction (Statistical/MFCC/Time domain)
Feature Selection using SFFS
Manual Feature Exploration
Correlation between Selected features
Output: Set of Suitable Pronunciation features

Fig. 1: Feature Selection Framework for Mispronunciation Detection

Table 1: Dataset details show the number of speakers and number of labeled phonemes.

		No. of Speakers		
	Adult Male	**Adult Female**	**Children**	**Total**
No. of Speakers	100	50	50	200
		No. of Labelled Phonemes		
	Adult Male	**Adult Female**	**Children**	**Total**
Native	780	275	240	1295
Non-Native	220	225	260	705
Total	1000	500	500	2000

Table 2: Explored Acoustic Features for Mispronunciation Detection.

#	Features	Dimensions	Feature Vector
[1]	Entropy of Spectrum (EOS)	6	EOS (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[2]	Lower Energy (LE)	6	LE (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[3]	MFCC & its derivatives (MFCC)	252	14 coefficients of MFCC, delta-MFCC, double delta-MFCC (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[4]	Pitch	6	Pitch (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[5]	RMS Energy (RMSE)	6	RMSE (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[6]	Spectrum	6	Spectrum (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[7]	Statistical Features* (Statistical)	6	Statistical (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)
[8]	Zero-Crossing Rate (ZCR)	6	ZCR (Mean, Std, Slope, periodic amplitude, periodic freq, periodic entropy)

Zero-Crossing Rate: The zero-crossing rate, a time domain feature, was calculated following the standard pattern reportedly (Zahid *et al*., 2015). Zero-crossing was calculated as follows:

$$ZCR = \frac{1}{2(M-1)} \sum_{n=1}^{M-1} |sgn[x(n+1)] - sgn[x(n)]| \quad (1)$$

In Equation (1), sgn[…] represents sign function while x(n) represents the values of the discrete signal from n=1,……, M.

Mel-Frequency Cepstral Coefficient (MFCC): Mel-frequency Cepstral Coefficient (MFCC) was calculated first, by Fourier transform of each small audio frame, then mapping the power spectrum on to Mel Scale obtained from the last step. The logarithm of power for each Mel frequency and Discrete Cosine Transform of Mel log power was calculated. The resulting amplitudes of the spectrum were MFCC coefficient (Dong *et al.*, 2006).

MFCC were calculated as:

$$\sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos\left[\frac{n(k-0.5)\pi}{K}\right] \quad (2)$$

$$where n = 1,2,3 \ldots . L$$

In equation (2), K represents the number of band pass filters and L represents the number of MFCC.

Spectral Features: Spectrum flux, the changeable power spectrum an audio signal, was extracted by calculating the Euclidian distance between two consecutive audio frames (Hacker *et al.*, 2007).

It was calculated as follows:

$$SF = \frac{1}{(M-1)(N-1)} \times \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} [\log(X(m,n) + \Delta) - \log(X(m-1,n) + \Delta)]^2 \quad (3)$$

In equation (3), k represented the input discrete audio signal and window function (k) represented length L. N, and M represented the order of DFT and total number of frames respectively.

Pitch: The rate at which vocal folds vibrates, when pressurized air coming from lungs pass through vocal folds. It was used for emotion recognition through speech. Mispronunciation detection systems also use Pitch due to its discriminative power to differentiate different sounds (Hermansky *et al.*, 1990).

Short-Time Energy: Shirt-Time energy, a feature that represents the energy of an audio signal was calculated using following standard procedure (Zhang *et al.*, 2014). It was calculated as:

$$E_m = \sum_{n=-\infty}^{\infty} [x(n)\omega(m-n)]^2 \quad (4)$$

Equation (4), $x(n)$represented a signal, *m* and $\omega(n)$ represented the frames and window size respectively.

Feature Selection using SFFS: Sequential Floating Forward Selection (SFFS) was used for automatic feature selection. SFFS being a greedy feature selection process, often find itself in local maxima by not covering all features. To cater this issue, in this study, SFFS was allowed to run for a complete set of features (Hassan *et al.*, 2009 and Ververidis *et al.* 2008). The algorithm of SFFS has been presented below:

Algorithm 1: A Modified SFFS Algorithm

```
Input: Set of all features, F= $f_1, f_2, \ldots, f_K$
Output: A subset of features $X^d$, where d<K
    $X^0 \leftarrow \{\ \}$ ; OC ← 0 ; NC ← 1 ;
    for (i=1:F)
    if (Converged) Break
    else if (NC >OC) then
    $X^{d-1} := X^d - f^-$
    Function 2 ()
    else
    function 1 ()
    function 2 ()
    end
    end else if
    end if
    end for
    function 1 ()
    $f^+ :=$ argmax $J(X^d + f_i)$ ;
    $X_{d+1} = X_d + f^+$
    OC ← $J(X^{d+1})$;
    end Function
    function 2 ()
    $f^- :=$ argmax $J(X^d - f_i)$ ;
    NC ← $J(X^d - f_i)$;
    end Function
```

Manual Feature Exploration: To cross validate the results obtained from SFFS, three data mining algorithms were used. These data mining algorithms included Bayesian (Domingos *et al.*, 1997), Random Forest (Breiman 2001) and SVM (Weston *et al.*, 1999). An ensemble method was used to improve the results for SVM algorithm.

Statistical Analysis: In this study, Accuracy and MAE were used for the statistical analysis. Accuracy was calculated as below:

$$\text{Accuracy} = \frac{Correctlyclassified}{\text{total no.of samples}} \times 100\% \quad (5)$$

It was highly desirable to keep this error to a minimum. It was calculated as:

$$MAE = \frac{1}{k} \sum_{i=1}^{k} |P_i - A_i| \quad (6)$$

Here k is the total number of samples, P_i represents predicted labeled phonemes and A_i represent actual labels of the phonemes. The dataset was divided by 80-20 rule for training and testing. All the calculations were made by using 10-fold cross-validation.

RESULTS AND DISCUSSION

A large set of acoustic-phonetic features was extracted for mispronunciation detection. These features were individually evaluated for each Arabic consonant and their impact on pronunciation. All the accuracy values were rounded off for simplification (Table 3, 4 and 5). Each feature was explored individually against all three classifiers for all Arabic consonants. The best performing features for each Arabic consonant are highlighted in corresponding tables.

Table 3: Feature Exploration using Bayesian.

Features				/		/					/							/	/		/	/			/	/		/	**Avg.**
	/aː/	/b/	/t/	θ/	/g/	ħ/	/ x/	/ d/	/ð/	/r/	z/	/ s/	/ š/	/ş/	/ď/	/ţ/	/đ/	ʕ/	ɣ/	/f/	q/	k/	/l/	/m/	n/	h/	/w/	y/	
EOS	81	89	**95**	60	89	95	94	89	81	95	94	84	93	92	92	86	89	78	89	59	91	82	86	72	95	49	47	74	83
LE	56	81	54	77	43	84	50	84	92	84	57	51	54	51	65	65	59	51	51	50	82	71	73	52	70	49	44	50	63
MFCC	**84**	**100**	92	91	**95**	**97**	**100**	97	97	**97**	**95**	**89**	92	**100**	**97**	**100**	**97**	**95**	**95**	**97**	**97**	**94**	**97**	**97**	**97**	**97**	**94**	**94**	96
Pitch	53	78	89	83	86	73	69	86	81	65	78	81	73	76	73	86	56	84	89	62	79	77	23	89	73	73	67	59	74
RMSE	78	84	95	86	87	92	81	89	86	84	76	76	**97**	68	81	95	86	73	76	74	88	88	70	83	86	49	67	44	80
Spectral	75	84	92	**94**	89	81	**100**	92	94	81	70	70	54	46	92	81	78	62	68	59	91	76	73	75	81	51	58	68	76
Stat	75	**100**	92	89	89	84	83	**100**	**100**	86	86	86	89	78	86	76	72	86	86	76	94	**94**	93	78	81	59	83	73	85
ZCR	63	78	78	80	86	86	67	81	89	81	78	65	81	70	62	78	62	62	46	76	56	47	86	56	51	46	72	47	69

Table 4: Feature Exploration using Random Forest.

Features	/aː/	/b/	/t/	/ θ/	/g/	/ ħ/	/ x/	/ d/	/ð/	/r/	/ z/	/ s/	/ š/	/ş/	/ď/	/ţ/	/đ/	/ ʕ/	/ ɣ/	/f/	/ q/	/ k/	/l/	/m/	/ n/	/ h/	/w/	/ y/	Avg.
EOS	72	92	**95**	60	89	**95**	97	89	83	**95**	**92**	84	84	95	**98**	86	92	84	95	65	94	82	86	89	95	65	91	74	86
LE	59	73	41	77	68	70	47	78	83	78	68	43	51	49	57	65	68	62	49	65	74	76	76	58	62	68	58	74	64
MFCC	**81**	**100**	92	91	92	100	94	**100**	**100**	95	**92**	**92**	**100**	**97**	97	100	**97**	**92**	**97**	**97**	**97**	**94**	**97**	**97**	**97**	**100**	**97**	**91**	96
Pitch	78	78	84	74	**95**	78	81	92	89	78	89	86	86	76	29	73	70	81	89	74	85	79	78	92	78	59	72	74	78
RMSE	**81**	81	**95**	89	78	**95**	94	92	89	84	86	89	95	84	73	97	78	84	84	94	83	88	83	89	86	62	81	74	85
Spectral	**81**	89	**95**	**94**	92	89	**100**	86	92	86	81	78	86	81	81	84	84	73	78	74	88	88	68	75	81	51	77	68	82
Stat	**81**	95	89	86	81	84	89	97	94	81	86	**92**	92	78	89	95	81	**92**	92	79	94	97	78	78	81	68	86	65	86
ZCR	69	76	78	83	84	86	78	89	78	84	86	86	76	83	68	81	78	84	62	65	65	71	81	69	68	70	72	68	76

Table 5: Feature Exploration using Bagged SVM.

Features	/aː/	/b/	/t/	/ θ/	/g/	/ ħ/	/ x/	/ d/	/ð/	/r/	/ z/	/ s/	/ š/	/ş/	/ď/	/ţ/	/đ/	/ ʕ/	/ ɣ/	/f/	/ q/	/ k/	/l/	/m/	/ n/	/ h/	/w/	/ y/	Avg.
EOS	**88**	**84**	**92**	77	84	86	81	84	81	**89**	**95**	**92**	89	81	**92**	84	**92**	**89**	**95**	68	94	79	**89**	**89**	**95**	65	**83**	74	85
LE	56	84	70	86	54	84	69	86	**94**	89	73	49	68	62	70	70	73	59	43	65	79	85	81	53	62	59	58	**79**	70
MFCC	**98**	**100**	**92**	**94**	**97**	**92**	**94**	**100**	**100**	77	**97**	**95**	**95**	**95**	89	**97**	**95**	**95**	**97**	**97**	**97**	**94**	**97**	**97**	**97**	**100**	**97**	**94**	95
Pitch	72	73	92	83	**86**	78	83	78	86	86	**89**	**92**	78	78	81	78	70	84	**92**	79	**96**	76	78	**89**	**86**	54	78	**76**	81
RMSE	**81**	78	89	**91**	76	**97**	**94**	**89**	89	**92**	84	84	**95**	**86**	89	**97**	81	84	81	**88**	91	**88**	81	86	86	**70**	81	68	86
Spectral	69	78	**97**	**94**	**86**	**92**	**92**	**95**	89	**89**	86	78	86	51	**95**	89	**92**	68	81	77	86	85	73	75	86	57	78	59	82
Stat	75	**92**	86	86	78	89	89	82	**94**	86	86	81	**95**	73	**92**	**92**	81	**89**	89	**82**	**94**	**97**	**81**	72	86	**65**	**89**	62	84
ZCR	72	84	86	77	86	89	61	78	75	78	86	78	65	**89**	84	78	76	78	59	65	53	74	78	72	70	59	69	71	75

The results showed that for most of the phonemes MFCC along with its derivatives performed exceptionally well. There were some other acoustic features which also showed promising results other than MFCC along with their first and second derivative. These features included Entropy of Spectrum (EOS), Statistical features, and Root Mean Square Energy (RMSE). These features showed exceptionally good results and in some cases performed even better than MFCC. Pitch and ZCR were considered as good discriminative features for many speech classification applications (Su *et al.* 2014). The average class-wise accuracies for pitch and ZCR were not bad, but MFCC, EOS, RMSE and statistical features outperformed these features. Two energy features; LE and RMSE were used in this system. RMSE consistently performed better than low energy. EOS and spectral features both contributed significantly to mispronunciation detection system, but EOS perform slightly better than spectral features. Statistical features were the only other features besides MFCC, which performed consistently exceptional. Although, MFCC even outperformed statistical features by a large margin, but statistical features were better than other features. The reason behind the outstanding results produced by statistical features was the overall impact of the signal. Pitch, Low energy, and the Zero-Crossing Rate (ZCR) did not contribute much to mispronunciation detection system. For some consonants, even these features showed promising results but the overall performance of these features was not satisfactory (Fig-2).

Random Forest provided 94.28% of average class-wise accuracy and 0.1603 average MAE. Bayesian produced 94.7% average class-wise accuracy and 0.0481 average MAE. Bagged SVM provided 94.9% average class-wise accuracy and 0.067 average MAE. These results suggested that Bagged SVM performed slightly better than other two classifiers.

Therefore, Bagged SVM was selected for the identification of suitable features. For all Arabic consonants, the number of occurrences of each feature in top three best forming features was counted. The results showed that MFCC were selected as the best feature with 27 occurrences, the second best feature was Entropy of Spectrum with 14 occurrences. RMSE and Statistical features had the same number of occurrences in top three best performing features with 12 occurrences (Fig-3). MFCC performed consistently outstanding for all consonants. The other three features also performed very well and at times even better than MFCC for some consonant.

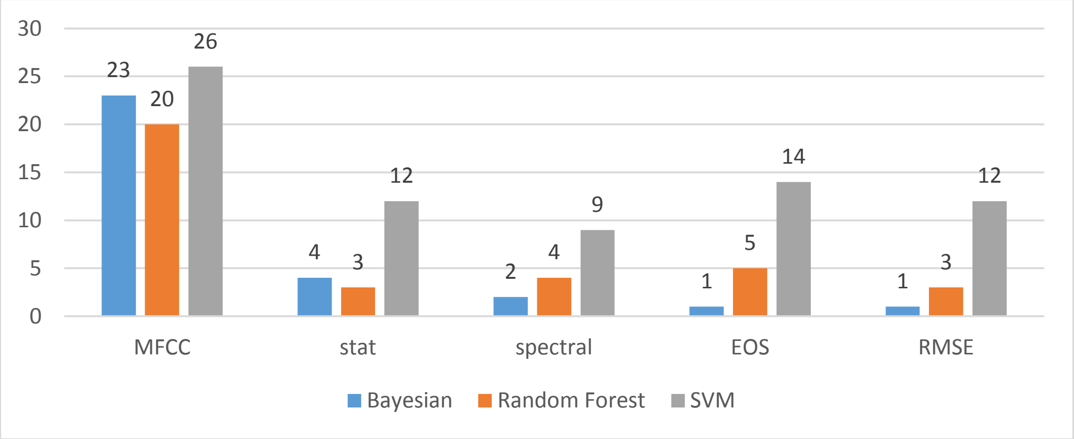

Fig. 2: A Comparison of Top three best performing features Selected by Classifiers

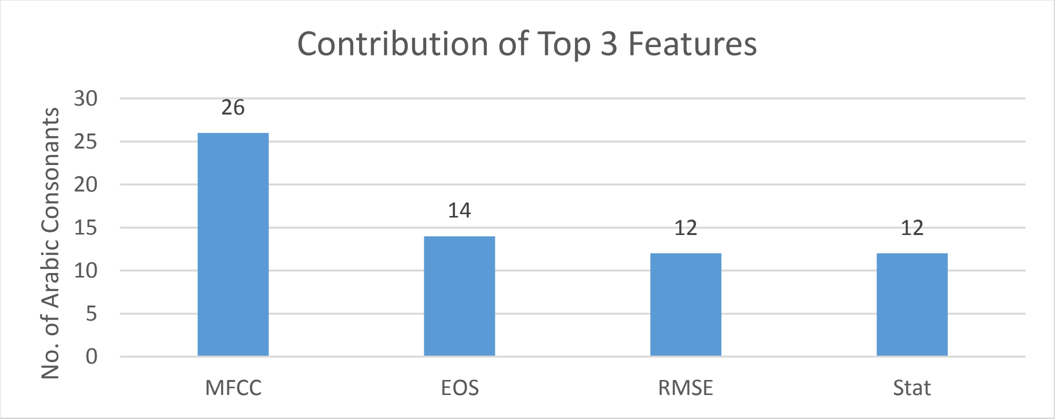

Fig. 3: Contribution of Top 3 best performing Features using Bagged SVM.

The SFFS process was used with default settings and it did not provide with any starting feature to start with. By keeping the drawback of greedy algorithms in mind, it was allowed to run for all features. Therefore, it covered all features to find out the best combination of acoustic features. Without this change, SFFS process terminated as soon it encountered the accuracy dropped for the first time. As Bagged SVM was used for the manual feature exploration, therefore the same classifier was used with SFFS. Detailed results of feature selection for each consonant and corresponding accuracies are presented in Table-6.

Discriminative features selected by SFFS were compared with top three handpicked features for each phoneme. This comparison was carried out to verify the feature selection results. The results confirmed that the same features were selected for both SFFS and manual feature selection. SFFS mostly selected more than three features for each phoneme. Top three features selected by manual testing were part of the feature vector selected by SFFS. Sometimes, feature selection process also selected some irrelevant features. The accuracy of selected features from both methods confirmed that classification results were better for handpicked features than SFFS. There were a group of acoustic features which repeatedly selected for each phoneme. MFCC along with its first and second derivatives were not the only suitable features for the pronunciation training systems. Therefore, it was highly recommended to not to use MFCC alone. MFCC were generic in nature while other acoustic-phonetic features were specific to the pronunciation mistakes (Franco *et al.*, 1999; Franco *et al.* 2000 and Ito *et al.*, 2007).

Correlation between handpicked features and SFFS was calculated. The results showed that handpicked features were highly correlated with the features selected from SFFS method. The correlation coefficient values were very good for most of the consonant. The overall average for correlation coefficient was 0.82.

Table 6: A Comparison for Features Selected through SFFS and top rated handpicked features using Bagged SVM.

Arabic Consonants	Features Selected through SFFS	Accuracy with Bagged SVM	Best Performing Features	Accuracy with Bagged SVM
/aː/	[1],[3],[4],[6],[7]	90.63	[1],[3],[5]	97.87
/b/	[1],[3],[5]	97.29	[2],[3],[7]	100
/t/	[1],[3],[4],[5],[7],[8]	91.89	[1],[3],[6]	97.29
/ θ/	[3],[4],[6],[7],[8]	91.43	[3],[5],[6]	94.28
/g/	[1],[3],[7],[8]	94.59	[3],[4],[6]	97.29
/ ħ/	[1],[3],[6],[7]	97.29	[3],[5],[6]	91.89
/ x/	[1],[3],[6],[7]	91.67	[3],[5],[6]	94.44
/ d/	[1],[3],[5],[7]	94.59	[3],[5],[6]	100
/ð/	[2],[3],[4],[7],[8]	98.87	[2],[3],[7]	100
/r/	[1],[3],[5],[6],[7],[8]	94.59	[1],[5],[6]	91.89
/ z/	[1],[3],[4],[7],[8]	99.00	[1],[3],[4]	97.29
/ s/	[1],[3],[4],[5],[7]	91.89	[1],[3],[4]	94.59
/ š/	[1],[3],[5]	89.19	[1],[3],[7]	94.59
/ş/	[1],[3],[4],[7]	97.29	[3],[5],[8]	94.59
/ď/	[1],[3],[6],[7]	86.49	[1],[6],[7]	94.89
/ţ/	[1],[3],[5],[7]	97.29	[3],[5],[7]	97.29
/đ/	[1],[3],[6],[7]	97.29	[1],[3],[6]	94.59
/ ʕ/	[1],[3],[5]	94.59	[1],[3],[7]	94.59
/ ɣ/	[1],[3],[4]	94.59	[1],[3],[7]	97.29
/f/	[3],[7],[8]	91.18	[1],[3],[4]	97.06
/ q/	[1],[3],[6],[7]	91.18	[3],[4],[7]	97.05
/ k/	[1],[3],[6],[7]	88.24	[1],[3],[4]	97.05
/l/	[3],[6]	97.29	[1],[3],[7]	97.29
/m/	[3],[4]	97.22	[1],[3],[4]	97.22
/ n/	[2],[3],[7]	97.29	[1],[3],[4]	97.29
/ h/	[3]	99.50	[3],[5],[7]	100
/w/	[3],[4],[7],[8]	94.44	[1],[3],[7]	97.22
/ y/	[3]	91.18	[2],[3],[4]	94.12

Many existing techniques used log-likelihood ratios as features for mispronunciation detection systems. In a study Abdou *et al.*, (2006) developed a system for Quranic Recitation pronunciation training and used Hidden Markov Model (HMM) for mispronunciation detection. The system produced 52.2% accuracy which was very low for a CALL system. In comparison, the present system produced an excellent accuracy of 94.9%. Mispronunciation detection system for only five Arabic consonants using GOP scores for mispronunciation detection has already been developed by (Al-Hindi *et al.*, 2014). By using confidence measure scores average accuracy of 92.95% was observed. The present system even outperformed the previous system by using discriminative and effective acoustic-phonetic features. The other notable systems developed by (Cucchiarini *et al.*, 2000; Muazzam *et al.*, 2015; Strik *et al.*, 2007 and Witt *et al.*, 2000) produced 86%, 92.15%, 80-92%, 81-

88% accuracy respectively (Table 7). The results showed that proposed method outperformed the existing systems and selected the best possible features for each Arabic Consonant.

Table 7: A comparative analysis between existing techniques and proposed system.

Mispronunciation Detection Systems for Arabic							
Systems	**Proposed Technique**	**Abdou *et al.* System**	**Strik *et al.* System**	**Alhindi *et al.* system**	**Witt. Technique**	**Cucchiarini *et al.* System**	**Muazzam et al. System**
Avg. Accuracy	94.9%	52.2%	81-88%	92.95%	80-92%	86%	92.15%

Conclusion: In this research, a feature extraction based study was conducted to identify the most discriminative acoustic pronunciation features for Arabic consonants. All features were manually explored and evaluated using Bayesian, Random Forest, and Bagged SVM. The bagged SVM produced significantly better results and top three best performing features were identified. A modified form of Sequential Floating Forward Selection (SFFS) process was used to validate the identified features and results showed perfect correlation.

REFERENCES

Abdou, S., S. E. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, and W. Nazih. (2006). Computer aided pronunciation learning system using speech recognition techniques. Interspeech 2006.

Al Hindi, A., M. Alsulaiman, G. Muhammad, and S. Al-Kahtani. (2014). Automatic pronunciation error detection of nonnative Arabic Speech. AICCSA, 190-197.

Breiman, L. (2001). Random forests. Machine learning, 45(1): 5-32.

Cucchiarini, C., H. Strik, and L. Boves. (2000) Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology. JASA 107 (2): 989-999.

Domingos, P., and M. Pazzani. (1997). On the optimality of the simple Bayesian classifier under zero-one loss." Machine learning 29, (2-3): 103-130.

Dong, B., Q. W. Zhao, and Y. H. Yan. (2006). Automatic scoring of flat tongue and raised tongue in computer-assisted Mandarin learning. In Proc. ISCSLP: 580-591.

Franco, H., L. Neumeyer, M. Ramos, and H. Bratt. (1999) Automatic detection of phone-level mispronunciation for language learning. In Eurospeech 1999.

Franco, H., L. Neumeyer, V. Digalakis, and O. Ronen. (2000) Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30: 121-130.

Hacker, C., T. Cincarek, A. Maier, A. Hebler, and E. Noth. (2007) Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children. ICASSP: IV-197.

Hassan, A., and R. I. Damper. (2009) Emotion recognition from speech using extended feature selection and a simple classifier. Interspeech: 2043-2046.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. JASA: 1738-1752.

Ito, A., Y. L. M. Suzuki, and S. Makino. (2007). Pronunciation error detection method based on error rule clustering using a decision tree. Acoust. Sci. & Tech: 131-133.

Maqsood, M., H. A. Habib, and T. Nawaz. (2015) Selection of Discriminative Features for Arabic Phoneme's Mispronunciation Detection. Pakistan J. Sci. 67 (4): 405-413.

Weston, J., and C. Watkins. (1999) Support vector machines for multi-class pattern recognition. In ESANN, 99: 219-224.

Strik, H., K. P. Truong, F. D. Wet, and C. Cucchiarini. (2007) Comparing classifiers for pronunciation error detection. In Interspeech: 1837-1840.

Su, L., C. M. Yeh, J. Liu, J. Wang, and Y. Yang. (2014) A systematic evaluation of the bag-of-frames representation for music information retrieval. Multimedia, IEEE Transactions, 16 (5): 1188-1200.

Truong, K. (2006) Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach. Proceedings of InSTIL/ICALL2004.

Ververidis, D., and C. Kotropoulos. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Processing 88 (12): 2956-2970.

Wei, S., G. Hu, Y. Hu, and R. Wang. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. Speech Communication 51(10): 896-905.

Witt, S. M., and S. J. Young. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication 30 (2): 95-108.

Zahid, S., F. Hussain, M. Rashid, M. H. Yousaf, and H. A. Habib. (2015). Optimized audio classification and segmentation algorithm by using ensemble methods. MPE, 2015.

Zhang, Y., D. J. Lv, and H. S. Wang. (2014). The application of multiple classifier system for environmental audio classification. AMM: 225-229.