An Efficient Mispronunciation Detection System Using Discriminative Acoustic Phonetic Features for Arabic Consonants

Muazzam Maqsood[1], Adnan Habib[2], and Tabassam Nawaz[1]
[1]Department of Software Engineering, University of Engineering and Technology Taxila, Pakistan
[2]Department of Computer Science, University of Engineering and Technology Taxila, Pakistan

Abstract: *Mispronunciation detection is an important component of Computer-Assisted Language Learning (CALL) systems. It helps students to learn new languages and focus on their individual pronunciation problems. In this paper, a novel discriminative Acoustic Phonetic Feature (APF) based technique is proposed to detect mispronunciations using artificial neural network classifier. By using domain knowledge, Arabic consonants are categorized into two groups based on their acoustic similarities. The first group consists of consonants having similar ending sounds and the second group consists of consonants with completely different sounds. In our proposed technique, the discriminative acoustic features are required for classifier training. To extract these features, discriminative parts of the Arabic consonants are identified. As a test case, a dataset is collected from native/non-native, male/female and children of different ages. This dataset comprises of 5600 isolated Arabic consonants. The average accuracy of the system, when tested with simple acoustic features are found to be 73.57%.While the use of discriminative acoustic features has improved the average accuracy to 82.27%. Some consonant pairs that are acoustically very similar, produced poor results and termed as Bad Phonemes. A subjective analysis has also been carried out to verify the effectiveness of the proposed system.*

Keywords: *Computer assisted language learning systems, mispronunciation detection, acoustic-phonetic features, artificial neural network, confidence measures.*

Received April 20, 2016; accepted November 9, 2016

1. Introduction

Computer Assisted Language Learning (CALL) systems have gained a lot of attention in recent years because of the advancements in artificial intelligence & machine learning. Mispronunciation detection is probably the most important feature of CALL systems. It is sometimes very difficult for language learners to take time out of their busy schedule to learn new languages. In language learning classes, pronunciation training is not primarily addressed and it becomes practically very difficult for the trainer to solve student's individual problems. Automatic mispronunciation detection systems can provide the language learners a platform to focus on their individual needs [6, 8, 12, 14].

In pronunciation assessment literature, mispronunciation detection is often confused with pronunciation scoring. Both functions are different from each other, serve different purposes and provide different outcomes. Pronunciation scoring rates someone's proficiency in speech but does not tell anything about the specific problems in pronunciation. Mispronunciation detection can point out the specific problems in someone's speech [11, 12].

Mispronunciation detection algorithms can be classified into two types; Confidence Measure (CM) based and classifier based [14]. CM based mispronunciation detection use Automatic Speech Recognition (ASR) systems to calculatestatistical scores [11]. Many mispronunciation detection systems are developed for different languages using ASR based CMs, which are ideally designed for speech recognition. While classifier based mispronunciation detection explores different sets of Acoustic-Phonetic Features. (APF) can represent the acoustic variations for anypronunciation mistakes in a better way. Therefore, APF can be used to formulate mispronunciation detection problem more comprehensively. However, it faces a major drawback, that discriminative pronunciation acoustic features are still unknown. That's why it's still an open research area, to find the set of most discriminative APF for pronunciation, and develop a mispronunciation detection system without using a traditional ASR system [14].

Arabic is the 5th most widely used language in terms of native speakers, more than 362 million speakers speak Arabic as their first language [2]. It is the language of the Holy Book of Muslims. So there are almost, more than a billion Muslims, who want to learn Arabic language pronunciation. A very little emphasis has been given to developing pronunciation training systems for Arabic.

On the other hand, many mispronunciation detection systems have been developed for English, Mandarin, Dutch, and French.

In this work, mispronunciation detection problem for Arabic consonants is formulated as a binary classification problem. For any phone, correct pronunciations are categorized in class 1 and all the mispronunciations for that phone are categorized in class 2. In this way, more acoustic features can be added and tested for mispronunciation detection. When mispronunciation detection is treated as a classification problem, it is required to train a classifier for each pronunciation mistake which requires a lot of memory and training time. To overcome this issue, by using domain knowledge, Arabic consonants are classified into two groups based on their acoustic similarities. A group having consonants with similar ending sounds and a group having totally different consonants. The most discriminative parts of the consonants having similar ending sounds are identified for features extraction. While complete consonants are used for feature extraction for the second group. A separate Artificial Neural Network (ANN) classifier for each groupis trained for mispronunciation detection. To evaluate the effectiveness of our proposed system, a medium size speech corpus of isolated Arabic consonantsisrecorded from 200 Pakistani speakers. Results demonstrate that the proposed system produce very good results. Which are comparable to the accuracies of CM based mispronunciation detection systems. A subjective evaluation is also carried out to validate the objective results.

In summary, this research work has the following contributions:

1. Discriminative parts for Arabic consonants are identified.
2. Set of Acoustic Phonetic features are identified for mispronunciation detection.
3. An Acoustic Phonetic features based mispronunciation detection classifier is developed and evaluated for Arabic consonants.

The rest of the paper is organized as follows; section 2 covers the related work of CALL systems and section 3 presents the proposed methodology. In section 4, results of all extensive experiments are presented along with detailed discussion followed by aconclusion and future work.

2. Related Work

Existing Computer Assisted language learning systems can be classified into two categorize; CM based systems and classifier based systems (acoustic phonetic based systems) [14].

In the first category, Witt and Young [16] proposed a pronunciation scoring method for non-native English speakers. This method achieved relatively high Scoring Accuracy (SA) of 80-92%. Cucchiarini *et al*. [6] proposed apronunciation training system for Dutch speakers. The system used a relatively medium size dataset from 15 speakers and achieved 86% accuracy. Ito *et al*. [9] proposed two new threshold calculation methods for mispronunciation detection. A series of HMM states were designed for both correct and incorrect pronunciation models. In the first method, a class dependent threshold is used to decide about mispronunciation which gives better results than the phone dependent thresholds. In the second method, a remarkable improvement in mispronunciation detection is observed when thresholds are calculated using decision tree-based approach. The limitation of pronunciation based model is that they can only give good results if pronunciation models truly represent the actual pronunciation variations. The representation of pronunciation variations is highly sensitive to speaker variations and limited availability of labeled corpus. Metawalli *et al*. [10] have developed a system for Quranic recitation training. The system was trained on the correct Quranic recitation (Tajweed). This system provides feedback about the pronunciation mistakes made by the user. The system use Hidden Markov Model (HMM) model and speaker adaptation for pronunciation scoring and also consider other factors like speaker variations. The system cover a large number of recitation mistakes and only manage to produce 52% accuracy. Abdou *et al*. [1] proposed apronunciation training system for Arabic, namedHAFSS. Articulation features are used in this system to calculate the pronunciation scoring. The HAFSS system uses a speech recognizer to detect mispronunciations fromuser'srecitation. A relatively large dataset is used for training and testing. The results were based on log likelihood ratios which are almost an approximation to posterior probabilities. The final results show that this proposed system reduce the false alarm to less than 25%. Al-Hindi *et al*. [2] proposed an ASR-based pronunciation training system for five Arabic Phonemes using standard Goodness of Pronunciation (GoP) algorithm. The average accuracy of this system is 92.15%. The limitation of this work is that it has been developed for only 5 Arabic phonemes. Confidence measure based methods calculate confidence measure by using an ASR systems. These systems estimate a threshold from labeled corpus to decide whether a word is correctly pronounced or not. These systems use sophisticated mathematical models and ASR toolkits to calculate the confidence measures. These systems produce good results but cannot identify the pronunciation error type.

In the second category, a combination of acoustic-phonetic features with a classifier are used to detect mispronunciations. Troung *et al*. [13] proposed a system to detect mispronunciations by using APF. Decision trees and Linear Discriminant Analysis (LDA) are used to discriminate plosives and fricatives

by using formants and duration features. It outperformed thetraditional ASR based GOP method [15]. Strik *et al*. [12] carried out a comparison method between GOP based systems and APF based systems. For this purpose, four types of classifiers are used to detect mispronunciations of velar fricative /x/ and the velar plosive /k/. The results show that APF based classifier outperformed other classifiers. It proves that if discriminative features can be identified and used by the suitability of the problem, APF based mispronunciation detection systems can outperform the traditional CM based systems.

Table 1. Details of LLD and statistical functions.

Feature	Description
Pitch	Pitch (f_0) in Hertz
Low Energy	Low Energy per frame
Spectral	Spectral features
Zero-Cross	Number of Zero-cross
Entropy	Entropy features
Cepstrum	14 Mel-Frequency Coefficient with delta and double delta
RMSE	Root mean square (RMS) energy
Statistical	Mean, periodic entropy, standard deviation, slope, periodic frequency, periodic amplitude

So mispronunciation detection should be made on the basis of discriminative features and not on the basis of confidence measures which are designed for ASR systems.

3. Proposed Mispronunciation Detection Techniques

The proposed Mispronunciation detection technique is presented in Figure 1.

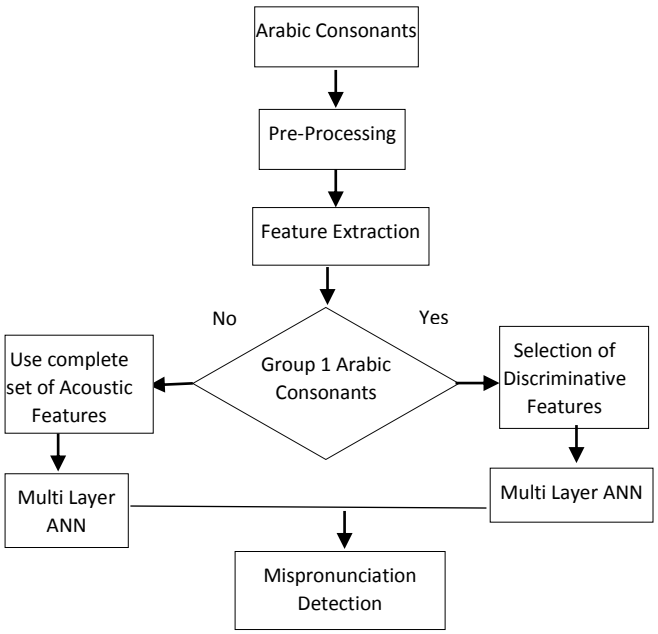

Figure 1. Proposed mispronunciation detection based on discriminative acoustic-phonetic features.

3.1. Pre-Processing

Noise is an important problem in this proposed work because data is recorded in an office environment. The recording files contain noise from different sources. The audio files also contain silence parts on both ends. The adobe audition software is used for noise removal and segmentation of Arabic consonants. The same software has been used for signal amplification to achieve a uniform *dB* level for all the signals.

3.2. Feature Extraction

Feature extraction and selection is a very important process in speech processing applications. It is a process to convert a signal into a series of features. Feature selection process can be used to choose the most discriminative features [17]. In mispronunciation detection systems, the discriminative acoustic features are still unknown. Therefore, researchers have used different sets of features to develop these systems. In this research, a large set of low-level descriptors were calculated which comprises of first 14 coefficients MFCCs along with its first and second delta, rms Energy, Pitch, Entropy, Spectral features, Cepstrum features, low energy and zero-cross. There are 6 statistical features that are also calculated. These statistical features include mean, standard deviation, periodic frequency, periodic amplitude, slope, and periodic entropy. These features are extracted using 25ms hamming window with 10ms shift. Each sample is divided into equal sizes using 44 KHz sampling rate. Details of these features are presented in Table 1.Some of these Acoustic features are explained here:

- *Zero-Cross Rate*: Zero-crossing is a time domain feature and tells that how many time a signal has changed its sign. Zero-cross is a frequency measure of the content of the signal. It describes the rate at which a signal crosses the zero value i.e. signal movement from a positive peak to negative peak. It is widely used in speech classification techniques [17].

Zero-crossing can be calculated as:

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

Here sgn[…] shows the sign function and the discrete signal and x(n)represents the values ranging from n=1,……,M.

- **Mel-Frequency Cepstral Coefficients (MFCCs)**:

Mel-frequency Cepstral Coefficients (MFCCs) is the most widely used feature in speech and music classification applications. Different sounds can be easily classified by using MFCCs because of its discriminative ability.This discriminative property has led its use in CALL systems. It can be calculated for frames as well as for speech segments [17]. Steps to calculate MFCCs can be explained as; first of all, an

audio signal is divided into frames to take Fourier transform.

Table 2. Details of arabic phonemes divided into two groups based on their similarity.

Group 1	ب	ت	ث	ح	خ	ر	ز	ط	ظ	ف	ه	ي						
Group 2	أ	ج	د	ذ	س	ش	ص	ض	ع	غ	ق	ك	ل	م	ن	و		

Then periodograms are estimated of the power spectrum for each frame, the logarithm of all energies are then taken followed by a Discrete Cosine Transform (DCT) of each Mel log power which gives MFCCs.

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

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

Here, *K* represents the number of band pass filters and *L* represents the number of MFCCs.

- *Spectral Features*: The spectral feature is a frequency domain feature. Formants are the most commonly used spectral feature, it is most widely used to disambiguate vowels. Mostly only first two formants are enough to disambiguate the vowels [17].
- *Pitch:* The rate at which vocal folds vibrates, when pressurized air coming from lungs are passed through vocal folds. The pitch is one of the important features used for speech recognition. It has also been used for emotion recognition through speech. The pitch is also used in mispronunciation detection systems by considering its discriminative power to differentiate different sounds [17].
- *Short-Time Energy*: This feature has been used by many researchers in speech classification applications [17]. It has also been used in mispronunciation detection systems. Short time energy can be defined as:

$$E_m = \sum_{n=-\infty}^{\infty}[\, x(n)\omega(m-n)]^2 \quad (3)$$

Here input signal is represented by$x(n)$, number of frames by m and window size by $\omega(n)$.

3.3. Discriminative Acoustic Phonetic Feature Selection

In any classification problem, machine learning classifiers heavily rely on input features. The quality of input features determines the performance of a classifier. In this research, the main aim is to identify and provide the discriminative acoustic-phonetic features as input. By using domain knowledge, it was observed that some Arabic phonemes have similar ending acoustic variations. There exist a group of phonemes which have similar ending sound (second half of the sound). While all the remaining phonemes have completely different sound from each other. Same language experts who labeled the training corpus were asked to group these Arabic consonants on the basis of these similarities. Language experts divided these phonemes into two groups named group 1 and group 2. Group 1 consonants include those consonants, which has similar ending sound (acoustic variations) while Group 2 consonants have totally different sounds from each other. The phonemes in each group are presented in Table 2.

Group 2 consonants are totally different from each other. Therefore, complete consonants were used for feature extraction. For Group 1, to extract the most discriminative acoustic features, the discriminative parts of the consonants have to be identified. Each Arabic consonant is divided into 10 equal segments. Acoustic features are extracted from these segments. An exhaustive processiscarried out to identify the discriminative parts of the Group 1 consonants. Different parts of the consonants are used to extract acoustic-phonetic features and tested to identify the most discriminative parts of the consonants.

3.4. Classification using Artificial Neural Network (ANN)

Artificial Neural Network is inspired by the human nervous system. It consists of different interconnected groups of multiple artificial neurons [3]. ANN is adaptive in nature, which means it can change its structure on the basis of information passing through it. It's a supervised learning algorithm and it is composed of simple elements called nodes. The input information is given to the nodes, these nodes calculate the output and the output is compared with already assigned target classes. If target class and output do not match, it is given back to the nodes and weights are readjusted to predict the new output. This process continues until the output error is minimum or zero. There are two types of ANN; single layer and multiple layers Neural Network [4]. Single layer neural network consists of a single layer of weights and nodes. This means input is directly connected to the output, so it can only handle linear problems. Multiple layer ANN consists of hidden layers other than input and output layer. As mispronunciation detection needs a supervised learning classifier.

In this work, a multiple layer ANN is used with multiple hidden layers and back propagation algorithms is used train the classifier. A separate ANN classifier is trained to detect mispronunciation for each group. Artificial neural network classifier is used to create nodes for each target label separately.

4. Experiments and Results

4.1. Dataset

The availability of standard corpus is very important for speech recognition related applications. The standard corpus should cover different acoustic and

speaker variations. There are no state-of-art corpus available specifically for Arabic mispronunciation detection, especially for Pakistani speakers. In this work, a dataset of Arabic consonants has been recorded for Pakistani speakers. These recordings have been carried out in 5 different sessions using a simple microphone in stereo using 44100 Hz sampling frequency in an open office environment. Total of 200 speakers including Males, females, and children of different ages were asked to record the data. The ages of these speakers ranged from 15-50 years with an average age of 25 years. These speakers include both types of speakers; speakers who are highly proficient in speaking Arabic (who have learned Tajweed) and speakers who had just started learning Arabic. Total of 130 speakers have excellent knowledge of Arabic and the rest of the speakers have just started learning Arabic.

Each speaker was asked to read all 28 Arabic consonants three times.As the recordings have been carried out in an open office environment. Therefore, many audio files were not suitable for experiments due to high noise. The repetitions per speaker enabled us to make sure that the best quality audio files are used for the experiments. A single, best quality recording, for each consonant, is selected per speaker for the experiment. Therefore, total dataset consist of 200 * 28 = 5600 consonants. The recorded Arabic consonants are available in separate audio files. List of 28 Arabic consonants along with their IPA is given in Table 3.

Five Arabic language experts from Pakistan having a large experience of teaching Tajweed in renowned institutions were asked to label the dataset. These Language experts labeled the Arabic phonemes separately. The labeled consonants are categorized into Native (correctly pronounced) and Non-Native (mispronounced) consonants. A consonant was assigned a native or non-native label, only if at least three of the language experts agree on the same label class. The Speaker distribution along with all the phonemes classified as native and non-native categories labeled by language experts are presented in Table 4. It also represents the data division for classifier training and testing.

Table 3. Details of all Arabic consonants.

Letter	IPA Symbol
أ	[ʔ]
ب	[b]
ت	[t̪]
ث	[θ]
ج	[dʒ]
ح	[ħ]
خ	[x]
د	[d̪]
ذ	[ð]
ر	[r]
ز	[z]
س	[s]
ش	[ʃ]
ص	[sˤ]
ض	[d̪ˤ]
ط	[t̪ˤ]
ظ	[ðˤ]
ع	[ʕ]
غ	[ɣ]
ف	[f]
ق	[q]
ك	[k]
ل	[l]
م	[m]
ن	[n]
ه	[h]
و	[W]
ي	[j]

Table 4. Details for dataset used for this experiment.

No. of Speakers				
	Adult Male	Adult Female	Children	total
No. of Speakers	100	50	50	200
No. of Labelled Phonemes for Training and Testing				
Training			Testing	
Native	Non-Native		Native	Non-Native
2900	1550		740	410

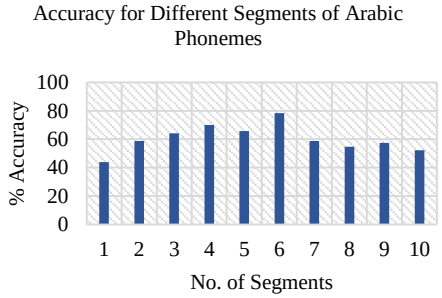

Figure 2. Mispronunciation detection accuracies for different parts of arabic consonants.

4.2. Metrics

Accuracy and Recall has been extensively used as a performance measure in Mispronunciation detection systems [2, 6, 7, 8, 12, 14]. This paper uses Accuracy

and Recall as the performance evaluation measures. Accuracy is a measure to present that how many instances have been correctly classified. It is desired to increase the accuracy and recall rate in order to achieve high performance. Accuracy can be calculated as:

$$\text{Accuracy} = \frac{Correctly\ classified}{\text{total no.of samples classified}} \times 100\% \quad (4)$$

Recall rate is used to measure that how many correctly classified mispronunciations are relative. It can be calculated as:

$$Recall = \frac{Number\ of\ True\ mispronunciations}{Total\ labelled\ mispronunciatoins} \times 100\% \quad (5)$$

4.3. Objective Results

To extract the most discriminative acoustic features, different segments (lengths) of Arabic consonants were tested. Each consonant is divided into ten equal segments. As group 2 consonants are completely different, therefore only group 1 consonants were tested for discriminative features. To start with, only firstsegments of all the consonants were used to extract the acoustic features. Then first two segments of all the consonants lengths were used and so on tillall ten segments (complete phoneme) of all the consonants were used for feature extraction. The acoustic features extracted for each part of all consonants were used to train aclassifier for mispronunciation detection and accuracies are presented in Figure 2. It shows that accuracy is not very good when first two segments of all the consonants are considered. The accuracy gradually rises after that and it reaches the highest value of 78.4% when first six segments of all the consonants are considered for feature extraction. After that, the accuracy again falls considerably for all the segments. This shows that the most discriminative acoustic features are extracted when we consider first six segments of all the Arabic consonants.

In our second experiment, two different methods were used for mispronunciation detection; Method 1 and Method 2. Method 1 use discriminative parts of the consonants for feature extraction of Group 1 consonants for both training and testing. The proposed method uses complete phonemes of group 2 phonemes for feature extraction.To verify the effectiveness of the method 1, another method is developed named method 2. The method 2 use complete phonemes of both groups for feature extraction. Two different testing conditions are used to evaluate both methods. First, when phonemes from group 1 are used for testing and second when phonemes from group 2 are used for testing.

When method-1 was tested with consonants from group 2, the accuracy of the system is 86.15% as presented in Figure 3. When consonants from group 1 were used for testing, the accuracy is 78.4% which is very good for a system covering such large number of pronunciation mistakes. So the average accuracy for method-1 is 82.27%. When method-2 was tested with group 2 phonemes, the accuracy is again 86.15%. The accuracy of the method-2 falls considerably when group 1 consonants were used for testing. The accuracy for group 1 consonants is 61% which is poor for a mispronunciation detection system. The average accuracy for method-2 is 73.57%. The recall rate of the method-1 and method-2 is 71% and 82.1% respectively. This proves that method-1 is more effective as compared to method-2. The problem with method-2 is that all the group 1 consonants have similar ending segments of the sounds. So acoustic features extracted from those similar segments are same for all consonants. These segments dominate the overall acoustic features, making it difficult for the classifiers to classify sounds which are very similar. These results prove the effectiveness of the proposed technique.

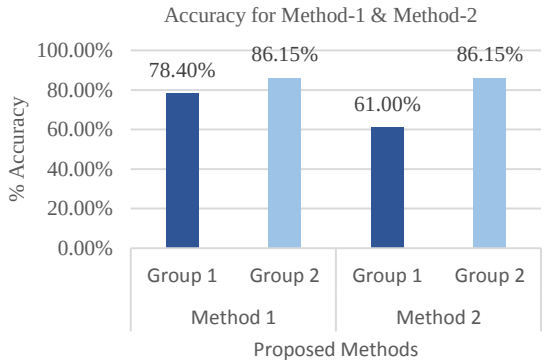

Figure 3. Accuracies for different percentages of phonemes' length.

Table 5. Classification results for discriminative classification technique.

Discriminative classification technique				
Techniques	**Method 1**	**Method 2**	**Abdou *et al.* System [1]**	**Al-Hindi *et al.* system [2]**
Accuracy	82.27%	73.57%	52.2%	92.95%

Acomparison with existing Arabic mispronunciation detection systems has also been presented in Table 5. Al-Hindi *et al*. [2] developed a system to detect mispronunciations for only 5 Arabic phonemes. The overall accuracy of the system is on the higher side as compared to proposed approach. It is mainly because of the reason the system covers pronunciation mistakes for only 5 Arabic consonants. Another reason for such a high accuracy is that they used a very well established ASR system for mispronunciation detection. While on the other hand, the proposed system covers the pronunciation mistakes for all 28 consonants. To cover such large number of phonemes in one system is very difficult. The average accuracy for Method-1, which uses discriminative acoustic features is 82.27%. The accuracy is very good for a system which onlyuses acoustic features. Abdou *et al*. [1] covered a large number of pronunciation mistakes

in one system. The overall accuracy of this system is 52% which is not very high. Even though they used statistical models, but covering such a large number of mistakes, makes it difficult for the system to detect pronunciation mistakes. The problem behind covering a large number of pronunciation mistakes is the diverse nature of pronunciation mistakes. As mostly a separate classifier is required to cater a single pronunciation mistake. The proposed system use only a single classifier for each group of consonants. The accuracy of the system is satisfactory, as it's a new technique and it can be further used a baseline for APF based mispronunciation detection system.

The biggest advantage of this research is the availability of the non-native data for Arabic mistakes. Most of the time non-native data is insufficient to train classifiers. To overcome this issue, researchers create artificial mistakes in the corpus. In this research, all the data was recorded from non-native speakers which were used to reliably train classifiers. Another advantage of APF based mispronunciation detection system is that it uses acoustic-phonetic features other than MFCC. MFCCs are more general acoustic features which are largely used in GOP based systems. While other acoustic-phonetic features are more specific to individual phones or pronunciation mistakes. Therefore, when there is a large mismatch between training and testing data, these specific features can outperform the MFCCs. Another advantage of APF based technique is that one can easily develop system while focusing on the specific APF by using existing knowledge. As APF are more specific for every mistake, it is very easy to design such mistake specific classifiers. Such system also suffers from a serious disadvantage, a separate classifier is needed to develop for every pronunciation mistake. ANN can be used to solve this issue to some extent, it creates separate nodes for each phone, making it handle each phone separately. But still, there is a need to develop such a system which can identify the pronunciation mistake and then automatically extract suitable features specific for those mistakes.

4.4. Subjective Results

It has been observed, through domain knowledge, that there exist a set of consonants that sounds very similar. These phonemepairs include /ت/ and /ط/, /ح/and /هـ/, /ح/ and /خ/, /ق/ and /ك/. It has also been reported by Alsulaiman *et al*. [5] that Pakistani speakers often confuse theseafore mentioned consonants pairs and make pronunciation errors. These consonants pairs, when pronounced by non-native speakers, sounds very similar. It was expected, the results of our proposed system should be even higher for group 1 consonants because of the discriminative acoustic features. But the accuracy of the proposed system is not excellent. The only reason might be that these set of consonants share almost the same starting segments too. Therefore, theclassifier is not been able to differentiate them properly.

In order to verify the results of our proposed technique, same five language experts were asked to classify these confusing phonemes. These language experts were asked to listen to these confusing phonemes by different non-native speakers and rate these phonemes as correct or incorrect. In order to get more realistic results, they were not informed about which phoneme they were listening. These results show that even the language experts face difficulty to properly classify these confusing phonemes as shown in Table 6. It is clearly evident from the results that most of the language experts were not able to differentiate between these contrast phoneme pairs when pronounced by non-native speakers. The average accuracies for /ت/, /ط/, /ح/, /خ/, /ق/, /ك/ and /هـ/ are 68%, 59%, 53%, 82%, 64%, 70% and 44% respectively. These subjective results confirmed the objective results of our system and also perfectly correlate the findings of Alsulaiman *et al*. [5]. When human judges find it difficult to differentiate between these phonemes when incorrectly pronounced, it is very difficult for the classifier to differentiatecorrectly. These phonemes are termed as Bad Phonemes. The term "Bad Phonemes" is used for those phonemes which are severely affecting the accuracy of the system [6].

Table 6. Average % accuracy for bad phonemes classification by 5 judges.

Phoneme	ت	ط	ح	خ	ق	ك	هـ
Judges Scoring	68%	59%	53%	82%	64%	70%	44%

4.5. Discussion

The key findings of our work are summarized here:

1. This research work produced comparable results to the existing systems that are based on highly sophisticated statistical models. In this work, discriminative part of the Arabic consonants are identified and the feature vector is extracted from these parts of the consonants. This helped algorithm to produce good results.
2. A set of acoustic-phonetic features is experimented to identify the features that are discriminative for mispronunciation classification. Thirteen discriminative features are derived for mispronunciation training systems and they produced a very good result. The results suggest that these features can be used for mispronunciation detection systems as a baseline.
3. Arabic phonemes are categorized into two classes based on their acoustic patterns. This strategy played an important role in increasing the accuracy of the system and keeping the computational cost

within reasonable limits. This approach can be used as a baseline in future for developing computer-assisted language learning systems for Arabic.

4. The proposed system was also tested for confusing phonemes (named as bad phonemes), the system produced comparable to the accuracy of human experts.

5. Conclusions and Future Work

In this paper, an acoustic phonetic feature based mispronunciation detection system is developed. A set of discriminative acoustic features for pronunciation are identified and used instead of existing statistical features for mispronunciation detection of isolated Arabic consonants. This paper suggested a novel discriminative classification approach for Arabic consonants by considering similarities in their acoustic patterns. The accuracy of proposed algorithm is increased by grouping Arabic phonemes into two groups on the basis of their similar sounds. Furthermore, the discriminative part is considered only for feature extraction of group 1 consonants. This system is tested on amedium level corpus consisting of 5600 consonants. This can be further extended for a large corpus.

There can be many future avenues of this research work. In future, a platform is required which can automatically extract the acoustic features for a specific pronunciation error. A generic language independent system can also be another future avenuein this area.

References

[1] Abdou S., Rashwan M., Al-Barhamtoshy H., Jambi K., and Al-Judaibi W., "Enhancing the Confidence Measure for an Arabic Pronunciation Verification System," *in Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, pp. 6-8, 2012.

[2] Al Hindi A., Alsulaiman M., Muhammad G., and Al-Kahtani S., "Automatic Pronunciation Error Detection Of Nonnative Arabic Speech," *in Proceedings of IEEE/ACS 11th International Conference on Computer Systems and Applications*, Doha, pp. 190-197, 2014.

[3] Ali H., Ahmad N., Zhou X., Ali M., and Manjotho A., "Linear Discriminant Analysis Based Approach For Automatic Speech Recognition of Urdu Isolated Words," *in Proceedings of in International Multi Topic Conference*, Jamshoro, pp. 24-34, 2013.

[4] Almehmadi T. and Htike Z., "Vehicle Classification System Using Viola Jones and Multi-Layer Perceptron," *The International Arab Journal of Information Technology*, vol. 13, no. 6A, pp. 965-971, 2016.

[5] Alsulaiman M., Ali Z., Muhammad G., Al Hindi A., Alfakih T., Obeidat H., and Al-Kahtani S., "Pronunciation Errors of Non-Arab Learners of Arabic Language," *in Proceedings of International Conference on Computer, Communications, and Control Technology*, Langkawi, pp. 277-282, 2014.

[6] Cucchiarini C., Strik H., and Boves L., "Quantitative Assessment of Second Language Learners' Fluency By Means of Automatic Speech Recognition Technology," *The Journal of the Acoustical Society of America107*, vol. 107, no. 2, pp. 989-999, 2000.

[7] Franco H., Neumeyer L., Ramos M., and Bratt H., "Automatic Detection of Phone-Level Mispronunciation for Language Learning," *in Proceedings of 6th European Conference on Speech Communication and Technology*, Budapest, pp. 851-854, 1999.

[8] Franco H., Neumeyer L., Digalakis V., and Ronen O., "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Communication*, vol. 30, no. 2, pp. 121-130, 2000.

[9] Ito A., Lim Y., Suzuki M., and Makino S., "Pronunciation Error Detection Method Based on Error Rule Clustering Using A Decision Tree," *in Proceedings of 9th European Conference on Speech Communication and Technology*, Lisbon, pp. 173-176, 2005.

[10] Metwalli S., "Computer Aided Pronunciation Learning System Using Statistical Based Automatic Speech Recognition Techniques," Ph.D. Thesis, Cairo University Giza, 2005.

[11] Odriozola I., Navas E., Hernaez I., Sainz I., Saratxaga I., Sánchez J., and Erro D., "D.: Using An ASR Database to Design A Pronunciation Evaluation System in Basque," *in Proceedings of 8th Internet Conference on Language Resources and Evaluation*, Istanbul, pp. 4122-4126, 2012.

[12] Strik H., Truong K., De-Wet F., and Cucchiarinia C., "Comparing Different Approaches for Automatic Pronunciation Error Detection," *Speech Communication*, vol. 51, no. 10, pp. 845-852, 2009.

[13] Truong K., Automatic Pronunciation Error Detection in Dutch as a Second Language: An Acoustic-Phonetic Approach, MA Thesis, Utrecht University, 2006.

[14] Wei S., Hu G., Hu Y., and Wang R., "A New Method for Mispronunciation Detection Using Support Vector Machine Based on Pronunciation Space Models," *Speech Communication*, vol. 51, no. 10, pp. 896-905, 2009.

[15] Weigelt L., Sadoff S., and Miller J., "Plosive/Fricative Distinction: The Voiceless

Case," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2729-2737, 1990.
[16] Witt S. and Young S., "Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, no. 2, pp. 95-108, 2000.
[17] Zahid S., Hussain F., Rashid M., Yousaf M., and Habib H., "Optimized Audio Classification And Segmentation Algorithm by Using Ensemble Methods," *Mathematical Problems in Engineering*, vol. 2015, pp. 1-11, 2015.

Muazzam Maqsood is currently doing his Ph.D. in Software Engineering from University of Engineering and Technology, Taxila. He has completed his MS degree in 2013 from University of Engineering and Technology, Taxila. His research interests include Speech Processing, Machine Learning, Recommender System and Image Processing.

Adnan Habib completed his MS (Electrical Engineering) in 2004 and Ph.D. (Electrical Engineering) in 2007 from University of Engineering and Technology, Taxila, Pakistan. He is currently serving as Head of Department of Computer Science in UET Taxila Pakistan. His research interests include Speech Processing, Image and Video Processing, Software Development, Artificial Intelligence and Artificial Neural Networks.

Tabassam Nawaz received his MS Computer Engineering in 2005 from CASE (Center for Advanced Studies in Engineering), Islamabad, Pakistan and subsequently, completed his Ph.D. in 2008. He is currently serving as a Head of Department of Software Engineering. His research interestsinclude Image and video processing, Software development, Artificial Intelligence and web development.