SELECTION OF DISCRIMINATIVE FEATURES FOR ARABIC PHONEME'S MISPRONUNCIATION DETECTION

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

Software Engineering Department, U.E.T Taxila *Computer Science Department, U.E.T Taxila
(Corresponding Email Address: adnan.habib@uettaxila.edu.pk)

ABSTRACT: Pronunciation training is an important part of Computer Assisted Pronunciation Training (CAPT) systems. Mispronunciation detection systems recognized pronunciation mistakes from user's speech and provided them feedback about their pronunciation. Acoustic phonetic features plays a vital role in speech classification based applications. This research work investigated the suitability of various acoustic features: pitch, energy, spectrum flux, zero-crossing, Entropy and Mel-Frequency Cepstral Coefficients (MFCCs). Sequential Forward Selection (SFS) was used to find out most suitable acoustic features from the computed feature set. This study used K-Nearest Neighbors *(K-NN)* classifier was used to detect the pronunciation mistakes from Arabic phonemes. This research selected the set of most discriminative acoustic features for each phoneme. *K-NN* achieved accuracy of 92.15% for mispronunciation detection of Arabic Phonemes.

Keywords: Mispronunciation Detection systems, Acoustic Features, Arabic Phonemes, Feature Selection, Sequential Forward Selection (SFS), K-NN.

(Received 26-08-2015 accepted 15-12-2015).

INTRODUCTION

Artificial intelligence and machine learning has been researched to develop automated systems to help people to learn new languages with the help of computers. Computer Assisted Pronunciation Training (CAPT) systems are used as pronunciation training tools to learn new languages. Mispronunciation detection is a process to find out deficiencies in pronunciation and provide useful feedback related to those deficiencies. It is highly desirable to make these language learning systems more reliable so that these systems can be used on large scale (Strik *et al.*, 2007 and Strik *et al.*, 2009).

These language learning systems heavily rely on acoustic features. Different set of pronunciation features are used to train classifiers, these features include pitch, MFCC with their first and second derivative, energy, fundamental frequency, energy and zero-cross (Lu *et al.*, 2003 ; Casey *et al.*, 2008 and Lu *et al.*, 2002). Most of the language learning systems are developed using statistical features and very little emphasis is given to acoustic phonetic features based CAPT systems. It is still a research avenue to find the most suitable pronunciation acoustic features because even after so much research has been done in this field, the set of most discriminative and optimal features for mispronunciation detection are still unknown (Wei *et al*., 2009).

Existing mispronunciation detection systems can be categorized in two classes; mispronunciation detection using statistical features and mispronunciation detection using acoustic phonetic features (Wei *et al.*, 2009). In first category, two different mispronunciation detection techniques are developed. In first technique, posterior probabilities are calculated using native acoustic models only, while in the second technique, the likelihood ratios are calculated using both native and non-native models and these probabilities are used as features for mispronunciation detection (Franco *et al.*, 1999). In another proposed technique used non-linear methods like classification and regression tree for mispronunciation detection, these non-linear methods based technique increased the quality of mispronunciation systems (Franco *et al.*, 2000). A local threshold based decision tree method is developed which produces better results in comparison to those methods using global thresholds for mispronunciation detection (Ito *et al.*, 2005). A mispronunciation detection method based on Scaling Posterior Probability (SPP) is proposed for mispronunciation detection which produced considerable results (Zhang *et al.*, 2008). A system has been developed to detect the mispronunciation for 5 Arabic phonemes taken from KSU dataset (Alhindi *et al.*, 2014). Goodness of Pronunciation (GOP) scores are used for pronunciation scoring (Witt *et al.*, 2000). Another HMM based Tajweed (Recitation of Holy Qur'an) feedback based training system is developed to find out the pronunciation mistakes for some specific mispronounced Arabic phonemes (Metwalli *et al.*, 2005). A CAPT system named HAFZSS has been developed to provide users about their mistakes while reading Arabic. They used HMM based model to find the mistakes in pronunciation, a confidence score is calculated by matching the reference manner features with classified manner features

and then classification is done by using this confidence score (Abdou *et al.*, 2012).

In second category, a wide range of acoustic features are used for the mispronunciation detection. Mispronunciation detection can be formulated more comprehensively using acoustic features but this approach faces a major drawback that the discriminative pronunciation features are still unknown. As set of most suitable pronunciation features are still unknown, there are very less pronunciation training systems available based on acoustic features. A similar acoustic phonetic feature based system is developed using set of both correct and incorrect pronunciation pairs for mispronunciation detection, while classification has been done using linear discriminant analysis and decision trees (Truong., 2004).

Feature selection process plays a vital role in improving the accuracies of machine learning classifiers used for the classification purposes. Feature selection algorithms eliminate features from the vector space, which play no or very little part in discriminating power of the classifier. Mostly, feature vector space is very large and it is highly desirable to reduce the number of features (Bocchieri *et al.*, 1993 and Luukka *et al.*, 2011). An Ada-boost based feature selection method has been developed to find the good pronunciation features and presented the top 15 most discriminative pronunciation features. They achieved the best accuracy of 89% by using top 35 features out of 176 dimensional feature vector (Hacker *et al.*, 2007).

Arabic is 5th largest language in terms of number of speakers, but still very little emphasis has been given for the development of CAPT systems for Arabic language. In this study, an approach is proposed based on acoustic features. A feature selection and classification technique is used to find out the best combination of pronunciation features. Various acoustic features are computed from Arabic phonemes dataset gathered from Pakistani speakers learning Arabic as second language. For feature selection, a modified form of Sequential Forward Selection (SFS) is used along with *K-NN* classifier (Hassan *et al.*, 2009 and Witten *et al.*, 2005). The results are compared with the existing systems which are using advance and complex classification algorithms. These results show that if good features are selected, then even a simple classifier can match to the accuracies of mispronunciation detection systems using complex machine learning classifiers.

The rest of the study is organized as: section 2 describes the proposed approach which includes feature extraction and feature selection; section 3 explains the dataset, evaluation metrics, results, and discussion followed by the conclusion.

MATERIALS AND METHODS

This paper presents an approach to identify the most suitable acoustic features for mispronunciation detection systems. A modified form of SFS in combination with K-NN classifier has been developed to identify the set of most discriminative pronunciation features. The reason to use a simple technique was to keep focus on the primary objective of feature selection.

Feature Extraction: Feature extraction can be explained as a process to convert the signal into series of features. A large set of features based on Low Level Descriptors (LLDs) was calculated which included pitch, minimum energy, spectrum, Entropy, 14 coefficients of MFCC along with its first and second derivative, RMS energy, Statistical features, and zero-cross rate.

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
Rms	Root mean square (rms) energy
Statistical	Mean, , periodic entropy, standard deviation, slope, periodic frequency, periodic amplitude

Statistical features included mean, standard deviation, slope, periodic frequency, and periodic amplitude. The details of LLDs and statistical features are given in table-1.

In this study, 289 acoustic features were calculated from each sample using 44 KHz sampling rate with 16 bit resolution. The sample was divided into frames of equal size by using 25ms hamming window with a 10ms shift as is shown in fig. 1. All acoustic measurements were made using Matlab.

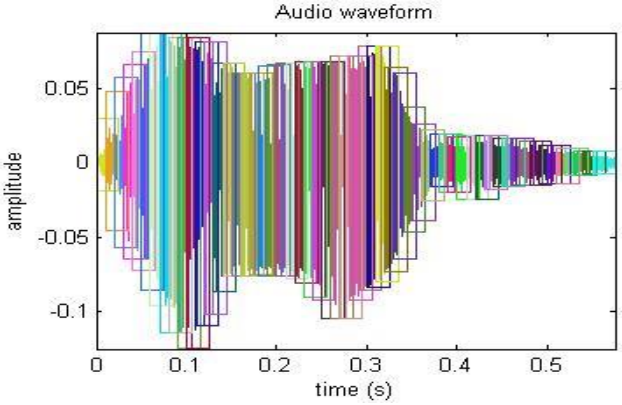


Fig. 1: An input audio signal is divided into short frames using hamming window

i. Zero-Crossing Rate: Zero crossing was a time-domain feature and was widely used in speech and music classification techniques. Zero-crossing rate determined the number of zero-crossing of a signal in a frame. It measured that how many times a signal had changed its sign i.e. movement of signal from positive peak to negative and vice versa. Zero-crossing can be used as a discriminating feature for speech and music. 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[…] represents the sign function and x(n) is the discrete signal with values ranging from n=1,……,M.

ii. Mel-Frequency Cepstral Coefficients (MFCCs): Mel-Frequency Cepstral features are short term spectral based feature and most commonly used in speech recognition. The success of MFCCs in speech recognition was because of its ability to discriminate between different sounds. To calculate MFCCs, divide audio signal into frames and take Short Time Fourier Transform (STFT) of these frames. For each frame, calculate the periodogram estimate of power spectrum. Then use the Mel log scale for power values and take the DCT. The resulting spectrum amplitudes gave the required MFCCs.

These features are extracted using a band-pass filter.

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

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

In this equation, number of band pass filters and MFCCs are represented by K and L respectively.

iii. Spectral Features: Spectral features represent speech signals in frequency domain besides f_0. Formants were the most commonly used type of spectral features and they represented the vocal tract frequencies when we speak. They were widely used in mispronunciation detection to differentiate between vowels and consonants and mostly first two formants were enough to disambiguate a vowel.

i. **Pitch:** Sound was produced when pressurized air coming from lungs passed through vocal folds, the rate at which these vocal folds vibrated was known as pitch of fundamental frequency. Speech recognition and emotion recognition systems commonly used pitch as a feature, due to its discriminative power it could also be used for mispronunciation detection systems.

ii. **Energy:** Energy has been pointed out by different researchers as a good feature for speech recognition and could also be used for mispronunciation detection. Short time energy can be defined as a measure of a total energy spectrum of a frame as given below.

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

Here $x(n)$ represented the input signal, m represented number of frames while $\omega(n)$ showed window used for analysis.

Sequential Forward Selection (SFS): There are many feature selection techniques used for this feature selection but most commonly used feature selection process was Sequential Forward Selection (SFS). It started with the best performing feature and added next best feature and tested the accuracy of the classifier. This algorithm will keep on adding more features until the accuracy of the classifier was increasing and it would terminate as soon as the accuracy of the classifier drops. A modified form of SFS was used because of its underlying greedy assumption. First, SFS was allowed to run until all the features were included. This modification was done because SFS terminated as soon as the accuracy dropped for the first time, making it difficult to guarantee the most optimal solution. Secondly, it was provided with a set of starting features. For this, each individual feature was manually evaluated to check its impact on the accuracy and then MFCCs was provided as a starting point to SFS based on its performance and its universal use in speech classification applications. Pseudo code of SFS is explained below:

Algorithm 1. Sequential Forward Selection (SFS) Algorithm

Input: Set of all features, Y= $y_1, y_2,..., y_d$
Output: a subset of features, Z_k=z_j | j=1,2,...,k; z_j | Y, where k=(0,1,2,...,d)

```
Z_k ← { } ;
OC ← 0 ;
NC ← 1 ;
while NC > OC do ;
OC ← J( Z_k) ;
f+:= argmax J( Z_k+f_i) ;
S^(k+1):= argmax J( Z_k U f_i) ;
NC ←S^(k+1);
k = k + 1;
end while
```

RESULTS AND DISCUSSION

Mispronunciation detection systems was developed for many languages like English, Mandarin, and Dutch etc. A very little emphasis has been given to Arabic language despite being the 5th largest language in terms of speakers.

Table 2: Details of all Arabic Phonemes used in our experiment.

Letter	Name	
أ	*'alif*	أَلِفٌ
ب	*baa'<*	بَآءٌ
ت	*taa'<*	تَآءٌ
ث	*thaa'<*	ثَآءٌ
ج	*jeem*	جِيْمٌ
ح	*haa'<*	حَآءٌ
خ	*khaa'<*	خَآءٌ
د	*daal*	دَالٌ
ذ	*thaal*	ذَالٌ
ر	*raa'<*	رَآءٌ
ز	*zayn*	زَيْنٌ
س	*seen*	سِيْنٌ
ش	*sheen*	شِيْنٌ
ص	*saad*	صَآدٌ
ض	*daad*	ضَآدٌ
ط	*taa'<*	طَآءٌ
ظ	*zaa'<*	ظَآءٌ
ع	*"ayn*	عَيْنٌ
غ	*rayn*	غَيْنٌ
ف	*faa'<*	فَآءٌ
ق	*qaaf*	قَآفٌ
ك	*kaaf*	كَآفٌ
ل	*laam*	لامٌ
م	*meem*	مِيْمٌ
ن	*noon*	نُوْنٌ
ه	*haa'<*	هَآءٌ
و	*waaw*	وَاوٌ
ي	*yaa'<*	يَآءٌ

There were no free available datasets for Arabic phonemes, that's why a dataset for Arabic phonemes has been developed for this research. This dataset for Arabic phonemes was collected from Pakistani speakers learning Arabic as their second language. These recordings were carried out in an office environment using a simple microphone to make it more real time system. Table-2 shows the complete list of Arabic phonemes used in our experiment. A total of 60 speakers participated in recording this dataset which included 30 adult males, 15 adult females, and 15 children.

All the speakers were judged by 2 Arabic language experts for Arabic proficiency level. They unanimously classified forty (40) as proficient and twenty (20) as non-proficient speakers. Dataset consisted of 28 Arabic phonemes and each speaker was asked to read all those phonemes, making 28x60=1680 phonemes in total. Table-3 showed the details related to number of speakers and phonemes in each class.

Table 3: Details for dataset used for this experiment

No. of Speakers				
	Adult Male	**Adult Female**	**Children**	**total**
Native	20	10	10	40
Non-Native	10	05	05	20
Total	30	15	15	60
No. of Phonemes				
	Adult Male	**Adult Female**	**Children**	**Total**
Native	560	280	280	1120
Non-Native	280	140	140	560
Total	840	524	425	1680

Many evaluation metrics have been used to evaluate the results of mispronunciation detection algorithms. These evaluation metrics included accuracy, Mean Absolute Error (MAE), Recall, and Receiver Operating Characteristic (ROC) sensitivity. Accuracy metrics of a mispronunciation detection system showed the frequency of the correctly classified instances by the classifier. MAE was the measure of the deviation of the actual class of the instance and predicted class by the classifier. The aim of the classification algorithm was to minimize the MAE value and increased the accuracy. In this paper Accuracy and MAE was used an as evaluation metric.

Accuracy could be defined as follows:

$$\text{Accuracy} = \frac{N_R}{N_D} \times 100\% \quad (4)$$

Here number of correct mispronunciation detection were represented by N_R and N_D has represented an overall number of detected mispronunciations by that of the system.

$$MAE = \frac{1}{k}\sum_{i=1}^{k} |P_i - A_i| \quad (5)$$

Here k was the total number of samples, P_i represented predicted labelled phonemes, and A_i represented actual labels of the phonemes.

A machine learning algorithm (*K-NN*) was used for classification purpose in combination with Sequential Forward Selection (SFS) technique. All the results in this paper used *K-NN* classifier using 10-fold cross validation. Different values of *k* were tested for *K-NN* and the best results were found for k=9. The motivation behind using such a simple classifier was to keep the focus on the primary objective of this research, which was feature selection.

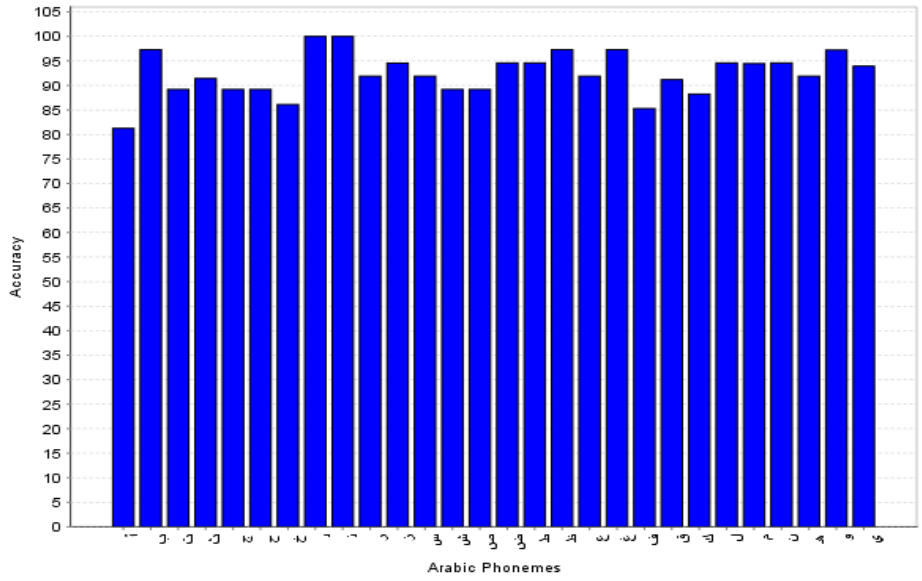
As the dataset consisted of different Arabic phonemes, feature selection process were employed for each phoneme. The results of the feature selection process showed that different set of acoustic features were selected for each individual phoneme, showing the unique acoustic characteristics of these phonemes. As each phoneme could be distinguished using different

acoustic features so the *k-NN* classifier was trained with different feature set for each phoneme.

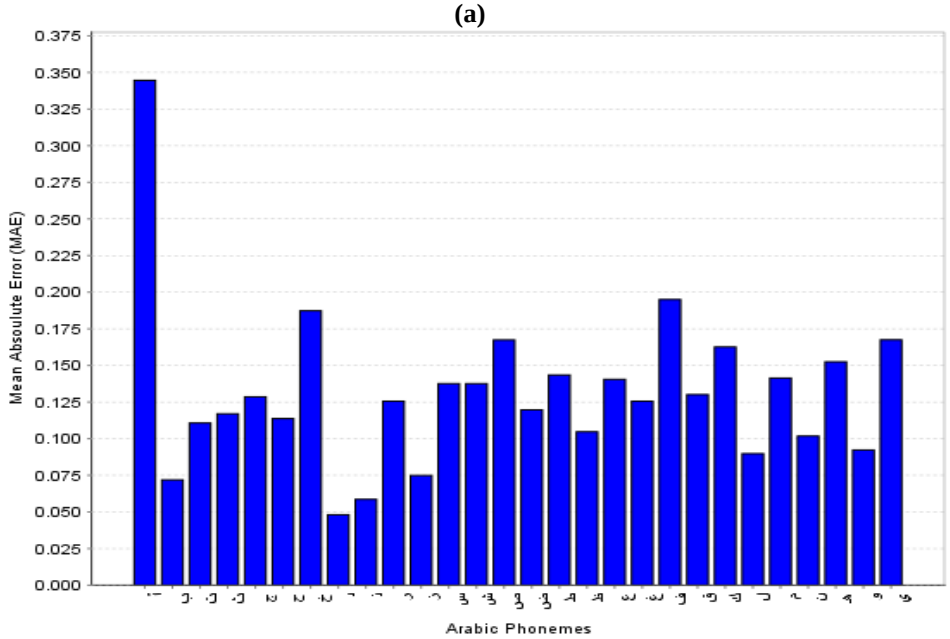
Table 4: Classification results for K-NN classifier used with SFS for k=9

Classification Results		
Feature Set	**Avg. No. of Selected Features**	**Avg. Accuracy with SFS**
289	102	92.15%

As the number of features also differed for each phoneme, average number of features have been presented which were selected by Sequential Forward Selection (SFS). A total of 102 features were selected on average by the SFS process. The accuracy for mispronunciation detection vary because of the difference in the feature vector's length used for each phoneme. So, overall accuracy of the system is presented as weighted average of all the accuracies calculated for all the phonemes. The class-wise average accuracy for proposed system is 92.15%. Table-4 shows the details of classification results of the proposed system.



(a)



(b)

Fig. 2: (a) Accuracy of each phoneme in Arabic Dataset (b) MAE of each phoneme in Arabic dataset

More detailed results for each Arabic phoneme are presented in fig.2, which showed a very high accuracy rates and low Mean Absolute Error (MAE) for each phoneme except for the first phoneme.

A detailed comparison has been presented in table-5 between proposed method and existing systems. These existing systems included different statistical features based CALL systems. A CALL system was developed for English and since then it has been considered as a benchmark (Witt and Young 2000). A relatively large database was used to train the system. GOP score was calculated for each phoneme and a threshold was used to decide about the correctness of that phoneme. Only statistical features were used in this system to decide the pronunciation quality. The accuracy achieved by this system was 80-92%. On the other hand, proposed system only used acoustic features for mispronunciation detection and achieved average accuracy of 92.15%. This too has been achieved without using any well-established mathematical model and ASR system.

A statistical CALL system was develop to detect mispronunciation for velar fricative /x/ and velar fricative /k/. Four different classifiers were designed for mispronunciation detection (Strik *et al.*, 2009). One of these classifiers was trained using acoustic features only and produced the best results. The best accuracy achieved by this system was 81-88%. The proposed system produced better results in comparison to this system. This system only detected a single pronunciation mistake while proposed system has been designed for 28 Arabic Phonemes. A feedback based CALL system for Dutch language was developed to provide pronunciation training for non-native speakers (Cucchiarini *et al.*, 2009). GOP scores were used to detect pronunciation mistakes and produced 86% accuracy. The proposed system produced better results for Arabic phonemes by using only acoustic features.

Another feedback based CALL system was developed for TAJWEED training (Metwalli *et al.*, 2005). Different types of pronunciation mistakes were covered in this system and each mistake was handled with separate classifier. This system covered common recitation mistakes, pronunciation mistakes related to phoneme durations, inter-speaker, and intra-speaker variability. Robust HMM was used as a base classifier to detect recitation mistakes. The overall accuracy was not satisfactory and only produced 52% correctly identified mistakes. On the other hand this proposed system coverd these mistakes and produced excellent results as compared to previous system.

Another CALL system based on statistical features was developed to detect mispronunciation for 6 Arabic phonemes (Alhindi *et al.*, 2014). This system also used GOP scores to detect mispronunciation detection. The class-wise average score for this system was 92.95%. The proposed system almost achieved similar results by using acoustic features. The proposed system was also developed for 28 phonemes in comparison to 6 Arabic phonemes used in this system.

Results showed that proposed system has been able to produce classification results very similar to the best yet published results by using a simple yet effective feature selection approach for Arabic CAPT systems. This performance was achieved by the selection of a most discriminative feature subset and using a simple classifier (*K-NN*) as compared to complex classifiers used for other systems. These results suggested that if feature selection process was carried out properly and good features were selected to train the classifier then even a simple classifier can perform in better way. Another conclusion could be made that may be it was difficult to set the optimal parameters for complex classifiers.

Table 5: Comparison of our proposed technique with existing Arabic CAPT systems

Mispronunciation Detection Systems for Arabic						
Techniques	Proposed Acoustic Feature Selection based Technique	Metwalli *et al.* System	Strik *et al.* System	Alhindi *et al.* system	Witt. System	Cucchiarini *et al.* System
Avg. Accuracy	92.15%	52.2%	81-88%	92.95%	80-92%	86%

In this study, the number of large acoustic phonetic features were computed to analyze their suitability for mispronunciation detection systems. For each Arabic phoneme, a set of discriminative features were selected by Sequential Forward Selection (SFS) process. It was noted that Sequential Forward Selection (SFS) method was allowed to run completely to identify the most discriminative set of pronunciation acoustic features. Typically, SFS process terminated as soon as the accuracy of the system droped for the first time making it hard to test all the features. After modification, SFS process was run through all the features pooled for this research. It was worth pointing out that Sequential Forward Selection process was used in this research to show the importance of feature selection process in this problem.

SFS was a feature selection process with the underlying “greedy” assumption for sequential selection algorithms. In future, we will experiment with further feature selection techniques to solve this problem. K-NN

classifier has been used for mispronunciation detection to demonstrate that even a simple classifier can produce good results by using most discriminative features. In this paper, individual phonemes were experimented. In future, it is planned that continuous speech will be experimented.

Acknowledgement: We hereby, highly acknowledge the funding given by U.E.T Taxila to support this research project.

Conclusion & Future Work: In this research we have presented extensive mispronunciation classification results using acoustic features for Arabic phonemes. A key step to improve the classification results in this research is to use a slightly modified form of Sequential Forward Selection (SFS) technique, to select the best combination of features out of relatively large set of 289 features extracted for this research. A simple classifier K-NN is trained on these features then is able to produce almost similar results as compared to the best published results for Arabic CAPT systems. These results suggest that by selecting the most discriminative pronunciation features, classification results can be improved even by using a simple classifier.

There are many future avenues for the future work, feature selection approach used here is based on a "greedy approach" which usually does not give best results. So it is needed to use a more comprehensive approach for feature selection for CAPT systems. The complication of the problem suggests that a more complex classifier is also required for classification purpose, in this research a simple classifier is used just to show the importance of feature selection step. It is suggested to all the researchers to use a simple classification algorithm before moving to more complex algorithm because sometimes simple classification algorithms can solve the problem very easily.

REFERENCES

Abdou, S., Rashwan, M., Al-Barhamtoshy, H., Jambi, K., and Al-Jedaibi, W. (2012). Enhancing the Confidence Measure for an Arabic Pronunciation Verification System. In Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training June (pp. 6-8).

Alhindi, A., Alsulaiman, M., Muhammad, G., and Al-Kahtani, S. (2014, November). Automatic pronunciation error detection of nonnative Arabic Speech. In Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on IEEE. (pp. 190-197).

Bocchieri, E. L., and Wilpon, J. G. (1993). Discriminative feature selection for speech recognition. Computer Speech & Language, 7(3): 229-246.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96(4): 668-696.

Cucchiarini, C., Neri, A., and Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. Speech Communication, 51(10): 853-863.

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

Franco, H., Neumeyer, L., Digalakis, V., and Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30(2): 121-130.

Hacker, C., Cincarek, T., Maier, A., Hebler, A., and Noth, E. (2007, April). Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on Vol. 4: IV-197. IEEE.

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

Ito, A., Lim, Y. L., Suzuki, M., and Makino, S. (2005). Pronunciation error detection method based on error rule clustering using a decision tree.

Lu, L., Zhang, H. J., and Jiang, H. (2002). Content analysis for audio classification and segmentation. Speech and Audio Processing, IEEE Transactions on, 10(7): 504-516.

Lu, L., Zhang, H. J., and Li, S. Z. (2003). Content-based audio classification and segmentation by using support vector machines. Multimedia systems, 8(6): 482-492.

Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications,38(4): 4600-4607.

Metwalli, S. (2005). Computer Aided Pronunciation Learning System Using Statistical Based Automatic Speech Recognition Techniques (Doctoral dissertation, PhD Thesis, Cairo University, Faculty of Engineering, Department of Electronics and Communication, Egypt).

Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. (2012). A literature review and classification of recommender systems research. Expert Systems with Applications, 39(11): 10059-10072.

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

Strik, H., Truong, K., De Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. Speech Communication: 51(10), 845-852.

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

Wei, S., Hu, G., Hu, Y., and Wang, R. H. (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 Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication, 30(2): 95-108.

Witten, I. H., and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhang, F., Huang, C., Soong, F. K., Chu, M., and Wang, R. (2008, March). Automatic mispronunciation detection for Mandarin. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on pp. 5077-5080. IEEE.