A Comparative Study of Classifier Based Mispronunciation Detection System for Confusing Arabic Phoneme Pairs

Article · June 2017

CITATIONS
9

READS
341

5 authors, including:

Muazzam Maqsood
COMSATS University Islamabad, Attock Campus
76 PUBLICATIONS **1,444** CITATIONS
SEE PROFILE

Hanna Habib
Ain Shams University
13 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Syed Anwar
University of Engineering and Technology, Taxila
133 PUBLICATIONS **3,738** CITATIONS
SEE PROFILE

Mustansar ali Ghazanfar
University of East London
58 PUBLICATIONS **1,359** CITATIONS
SEE PROFILE

The Nucleus ***54****, No. 2 (2017) 114-120*

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)
ISSN 2306-6539 (Online)

A Comparative Study of Classifier Based Mispronunciation Detection System for Confusing Arabic Phoneme Pairs

M. Maqsood[1,3,*], H. A. Habib[2], S. M. Anwar[1], M. A. Ghazanfar[1] and T. Nawaz[1]

[1]*Department of Software Engineering, UET, Taxila, Pakistan*

[2]*Department of Computer Science, UET, Taxila, Pakistan*

[3]*Department of Computer Science, COMSATS Institute of Information and Technology, Attock, Pakistan*

muazzam.maqsood@uettaxila.edu.pk; adnan.habib@uettaxila.edu.pk; s.anwar@uettaxila.edu.pk; mustansar.ali@uettaxila.edu.pk tabassam.nawaz@uettaxila.edu.pk

ARTICLE INFO

Article history:
Received : 02 November, 2016
Accepted : 20 June, 2017
Publsihed: 30 June, 2017

Keywords :
Computer assisted language learning systems
Acoustic-phonetic features
Arabic phonemes
Mispronunciation detection
Machine learning classifiers

ABSTRACT

Pronunciation training systems detect mispronunciations from language learner's speech and provide useful feedback. Mispronunciation detection systems can either be developed using Confidence Measures (CM) or using classifiers with Acoustic Phonetic Features (APF). This paper presents an APF based computer assisted pronunciation training (CAPT) system for most confusing Arabic phoneme pairs (/ط/ vs /ت/)and (/ ح/ vs /خ/ or /هـ/) developed for subjects of Pakistani origin. A super-vector is formed based on APF consisting of Mel-frequency cepstral coefficients (MFCCs) along with its first and second derivative, energy, zero-cross, spectral features and pitch. A large dataset has been recorded from 200 speakers of Pakistani origin learning Arabic as their second language. Four different machine learning classifiers; Random Forest, Naïve Bayes, Ada-boost and K-NN have been used for mispronunciation detection. A comparison has been conducted between these classifiers and standard Goodness of Pronunciation (GOP) method. The results show that Random Forest outperforms all other methods by a significant margin.

1. Introduction

Advancements in Artificial Intelligence and Machine Learning have led to automation in many fields. Computer language learning (CALL) system is one such research area. Learning new languages has become a requirement to communicate with people around the globe [1]. There is a great need to develop intelligent systems to help users in learning new languages. These systems can provide a platform to language learners to work on its particular mistakes [2, 3]. These systems can be more useful if they can point out mispronunciations and provide feedback.

Computer-assisted pronunciation training systems can be categorized into two categories; 1) mispronunciation detection and 2) pronunciation scoring [2]. Mispronunciation detection based CALL systems find out mistakes from the utterances of any learner and provide useful feedback related to their particular mistake, whereas pronunciation scoring only rates the proficiency level and cannot tell anything about the pronunciation mistake. Mispronunciation detection is the most useful feature of CALL systems [2, 4].

Mispronunciation detection systems can be formulated using confidence measures (CM) and by classifiers with Acoustic Phonetic Features (APF) [2, 4, 5]. CM based mispronunciation detection systems use Automatic Speech Recognition (ASR) toolkits which are based on well-defined mathematical models [6]. These systems use thresholds derived from the corpus to decide the correctness or incorrectness of the phonemes. Witt [7] introduced Goodness of Pronunciation (GOP) method to detect mispronunciations and it is used as a state-of-the-art system. When mispronunciation detection is considered as a 2-class classification problem, more APF can be used with different classifiers. However, the classifier-based approach can only be used if confusing pronunciation pairs are known [5]. The uttered sound should be properly represented by its acoustic features and any pronunciation mistake should be judged by analyzing the change in these feature values. Therefore, mispronunciation detection problem can be more comprehensively formulated using acoustic-phonetic features [4, 5, 8]. Most mispronunciation detection systems use CM and very little emphasis has been given to APF based mispronunciation detection systems. The reason behind this limited use of APF based systems is that generic discriminative pronunciation Acoustic-

*Corresponding author

phonetic features are still unknown [2]. Therefore, APF for each individual pronunciation error has to be non-statistically identified. However, if discriminative APF can be identified, APF based classifiers can outperform standard GOP based system [4]. These particular features are only applicable to one pronunciation error type. Therefore, an efficient classifier based system is required that can use generic and potentially discriminative APF to handle various pronunciation errors.

Table 1: Segmental pronunciation errors addressed in this research

Arabic Target Consonant with IPA	Mispronounced Arabic Consonant with IPA
/ ت /, /t̪/	/ط / , /t̪ˤ/
/ ح / , /ħ/	/خ / , /ħ/ or /هـ / , /x/

In this paper, a classifier based mispronunciation detection system has been developed for 2 most confusing Arabic pronunciation pairs (/t̪/ vs /t̪ˤ/) and (/h/ vs /ħ/ or /x/) [9] for Pakistani speakers as presented in Table 1. Using domain knowledge, it has been observed that Pakistani speakers often substitute these phonemes from their 1st language. In this research, classifiers are trained to differentiate between the correct pronunciations of confused pronunciation phoneme pairs. A high dimensional APF based super-vector is formed to handle various pronunciation mistakes. Four different classifiers, Random Forest, Naïve Bayes, K-NN and Ada-boost have been tested and evaluated to check their suitability for mispronunciation detection problem. To compare effectiveness of the classifier based system, the standard GOP system was also developed for same confusing pronunciation phoneme pairs. The proposed classifier based system outperformed the standard GOP method. Results have been compared with the existing systems, and the proposed super-vector based classifier approach showed improved results.

This paper is organized as follows; in section 2 literature review has been presented, section 3 explains the methods and materials, section 4 presents results and discussion followed by conclusions and future work.

2. Literature Review

Computer assisted language learning systems have been researched and proposed in recent years mostly for languages like Mandarin, English, French, Spanish and Korean [3, 6, 7, 10-15]. These systems are classified in two classes; using Acoustic Phonetic Features (APF) and by Confidence Measure (CM). In the first category, Franco [12] has developed two different techniques for pronunciation detection, firstly only native acoustic models are used to calculate posterior probabilities while the 2nd method uses both native and non-native models for mispronunciation detection using log-likelihood ratios. Another method was proposed by Franco [16] in which they used non-linear methods like classification and regression tree for error detection in pronunciation. This method increases the quality of mispronunciation systems. A decision tree based method is used by Ito [17], where they used different thresholds for different mispronunciations and claimed improved results as compared to global threshold methods. Scaling posterior probability (SPP) for mispronunciation detection was developed by Zhang [6] and achieves considerably good results. Witt and Young [7] introduced a pronunciation scoring algorithms known as Goodness of Pronunciation (GOP) that is a variation of a posterior probability score. This GOP algorithm is now considered as the benchmark algorithm for CALL systems [4, 11, 14, 16-22]. Wang [23] proposed a hybrid structure that incorporates GOP scores and different patterns to find pronunciation mistakes. A mispronunciation detection system for five Arabic phonemes has been proposed using King Saud University (KSU) [1] database. These phonemes are mostly mispronounced by people of Pakistani and Indian origin. The GOP score was used to classify speech as correctly or incorrectly pronounced. Their algorithm was tested for each phoneme separately and produced very good average accuracy for the system. Another TAJWEED training system was developed to find out mistakes from continuous Arabic speech [24]. They recorded some predefined words and these words were converted intophonemes. A Hidden Markov Model (HMM) based classifierwas trained on these features to detect mispronunciations [25]. A confidence measure based pronunciation training system named HAFZSS was developed to provide feedback for Arabic. Mispronunciation was detected by calculating CM score from the difference of classified mannered features and reference manner features. These CM scores were used by the HMM model to classify the phonemes as correct or incorrect.

In the second category, a wide range of acoustic features were used for mispronunciation detection. Truong [8] defined different pairs for mispronunciation detection using linear discriminant analysis and decision trees. These methods use some formants and duration based acoustic features. Different comparisons were done between these two broad categories by Strik [4] and they showed that acoustic phonetic based LDA methods are better than CM based GOP. LDA-based methods can distinguish between fricative /x/ from velar plosives/k/. Doremaien [26] proposed a classifier based mispronunciation detection system for the Dutch language. Wei [2] also proposed a classifier based pronunciation training system using pronunciation space models.

3. Materials and Methods

3.1 Feature Extraction

Feature extraction process takes signal as an input and converts it to different required features. Different researchers have used a different set of acoustic and

Table 2: Dataset details for labeled phonemes andspeakers

No. of speakers			
Adult male	Adult female	Children	Total
100	50	50	200
No. of labelled phonemes			
2500	1500	1000	5000

features. Different speech features were extracted from audio signals consisting of MFCCs along with first and second derivative, Linear Prediction Coefficient (LPC), Perceptual Linear Prediction, zero-crossing, energy, pitch, entropy and spectral features along with 6 statistical features including mean, slope, standard deviation, periodic frequency, periodic entropy and periodic amplitude.

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are most commonly used in speech processing and its applications because of their ability to differentiate between different sounds. MFCCs are a short term spectral feature that can be calculated using a band-pass filter [27].

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

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

In this equation, K represents the number of band pass filters S_k represents the output power of the k^{th} filter and L represents MFCCs.

Linear Predictor Coefficients (LPC): Linear Predictor Coefficients (LPC) represents the auto-regressive model of speech. Speech is divided into small frames. The mathematical representation of all-pole vocal tract transfer can be represented as:

$$H(z) = \frac{G}{1-\sum_{i=1}^{P} a_i z^{-i}} \quad (2)$$

Where a_p represents the prediction coefficients, z is a polynomial, while G represents the gain and $H(z)$ is a z-transform of the filter. LPC can be obtained by minimizing the mean square error between actual and estimated samples using autocorrelation [28].

Perceptual Linear Prediction (PLP): The Perceptual Linear Prediction (PLP) models human speech using psychophysics of hearing. It removes irrelevant information thus improves speech recognition. It is identical to LPC except that it transforms spectral statistical features to develop mispronunciation detection systems. The most discriminating pronunciation features are yet to be identified [2]. In this research, a super-vector was formed consisting of generic APF for Arabic mispronunciation detection system. A hamming window of size 25ms was used with 10ms shift to extract the audio characteristics to match human auditory system. PLP makes computations on three levels; i) critical band resolution curve, ii) equal loudness curve, and iii) intensity-loudness power-law relation [29].

Zero-Crossing Rate: Zero-crossing is a feature to represent how many times a signal has changed its sign. [30]. It is calculated as:

$$ZCR = \frac{1}{2(M-1)}\sum_{n=1}^{M-1}|sgn[x(n+1)] - sgn[x(n)]| \quad (3)$$

Here sign[…] shows the sign function and the discrete signal with values ranging from n=1,……,M is represented by x(n).

Spectral Features: Spectral features are frequency domain feature besides fundamental frequency. Formants are the most commonly used features to differentiate between vowels and consonants.

Pitch: Pitch is defined as the rate at which vocal folds vibrate when sound is produced. Pitch is also used as one of the most discriminative features in speech and emotion recognition systems [31].

Short-Time Energy: Energy is also considered as a good feature for speech recognition systems [31]. It has also been used in mispronunciation detection systems and can be calculated as:

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

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

3.2 Classifiers

There are many classifiers used in speech processing applications. Mispronunciation detection systems are usually based on CM, where HMM is the obvious choice. In this paper, APFs were used to develop mispronunciation system. Therefore, different classifiers were tested to evaluate the performance for mispronunciation detection systems.

Naïve Bayes: A Bayesian classifier that is based on Bayes theorem with a strong independence assumption [32, 33]. Using Bayes theorem the probability of a single phoneme residing in a particular class can be calculated as :

$$\mathrm{P}\,(Ci/\,X) = P\,(X|Ci)\,P\,(Ci)/\,P\,(X) \quad (5)$$

Naïve Bayes is different from Bayesian in a sense that it assumes conditional independence of each attribute. It means that no attribute is dependent on any other attribute. *K-NN:* K-NN is considered as a simple and

Table 3: The percentage accuracies (%) for each target phoneme

	Naïve Bayes	KNN	Ada-boost	Random Forest	GOP
/t̪/	75.66	80.1	76.7	94.8	71.5
/t̪ˤ/	74.5	77	72.5	96.9	73.8
/h/	68.1	87.1	81.14	94.1	80.5
/ħ/	67.43	87.39	79.8	92.7	71.4
/x/	84.3	90.5	86.9	98.5	85.5

fundamental instance based machine learning classification algorithm [34]. It is used to classify an instance based on the similarity to its neighbors. By default, linear search is used to find similarity between instances but there are some other searching methods also available. These searching methods use distance as a selection parameter where Euclidean distance is mostly used as a default.

Ada-boost: Ada-boost is an adaptive machine learning classifier and uses different weak classifiers to form a strong classifier. It is adaptive in a sense that you can tweak the weak classifier to handle those instances which are misclassified by previous weak classifiers. It is sometimes sensitive to outliers and noisy data but it can handle the over-fitting problem which is faced by many other learning algorithms.

Random Forest: Random forest is an ensemble method, based on many decision trees[35, 36]. Random forest uses bootstrap aggregation and random feature selection to construct a collection of decision trees. The final predicted class is obtained by combining the predictions of all the trees.

4. Experiments and Results

4.1 Dataset

There are many CALL systems available for different languages like English, Mandarin, Dutch and French but very little emphasis has been given to Arabic. Therefore, there is no state of the art dataset available for Arabic mispronunciation detection. In this research, a dataset was recorded from 200 speakers of Pakistani origin, learning Arabic as their second language. These speakers were asked to read phonetically rich Arabic sentences.

This dataset covers both types of speakers; highly proficient as well as those learners who have just started learning Arabic. The dataset was recorded in an office environment using a simple microphone. All phonemes were segmented automatically using HTK toolkit [37].

The labeling process was carried out by five Arabic language experts. All Arabic phonemes were labeled by each language expert separately. All language experts classified these Arabic phonemes into correct or incorrect classes. A certain label was assigned to a phoneme ifat least three language experts agreed on the same class. Details of speaker'sand labeled phonemes are presented in Table 2.

4.2 Evaluation Metrics

To evaluate CALL systems, different evaluation matrices were used for accuracy, precision, recall and Mean Absolute Error (MAE). In this paper, accuracy was used to evaluate the results.

Accuracy can be defined as follows :

$$\text{Accuracy} = \frac{N_R}{N_D} \times 100\% \qquad (6)$$

Where N_R and N_D represent the number of true mispronunciations detected and the total number of mispronunciations detected by the system, respectively.

5. Results and Discussion

This section presents the results for two pronunciation contrast pairs that includes (/t̪/ vs /t̪ˤ/) and (/h/ vs /ħ/ or /x/) phonemes. Four different classifiers; Naïve Bayes, K-NN, Ada-boost, and Random Forest were tested for each target phoneme. The dataset has been divided according to 80-20 rule, where 80% of the data was used for training while 20% of the data was used for testing purpose. Equal numbers of instances were used for each phoneme and all classifiers were used with default settings. A baseline GOP system was also developed similar to [1], for the same pronunciation pairs in order to compare the effectiveness of the proposed classifier system.

The performance of all 4 classifiers; Naïve Bayes, K-NN, Ada-boost, Random Forest and GOP method has been evaluated for the 1st confusing pair (/t̪/ vs /t̪ˤ/). The results for each target phoneme are presented in Table 3. The average accuracies are found to be 75.1%, 77%, 72.5%, 95.9% and 73.8% respectively. The performance of the same 4 classifiers was evaluated for the second confusing pair (/h/ vs /ħ/ or /x/). The accuracies for second confusing pair were found to be 73.3%, 88.3%, 82.6%, 95.1% and 79.1%, respectively.

The results for both confusing pairs are presented in Fig. 1. The results show that the classifier based approach efficiently handles both confusing pairs. It can be inferred from the results that Random forest outperformed

Table 4: Comparison of accuracies of proposed system with existing CAPT systems

Techniques	Language	Method	Non-standard dataset details			Average accuracy
			No. of target phonemes	No. of speakers	No. of phonemes	
Abdou et al. [25]	Arabic	GOP	NA	43	2742	92.58%
Al Hindi et al. [1]	Arabic	GOP	5	32	905	92.95%
Strik et al. [4]	Dutch	Classifier+APF	2	31	13290	90.1%
Troung et al. [8]	Dutch	Classifier+APF	2	80	1364	87.18%
Proposed System	Arabic	Classifier+APF	5	200	5000	95.4%

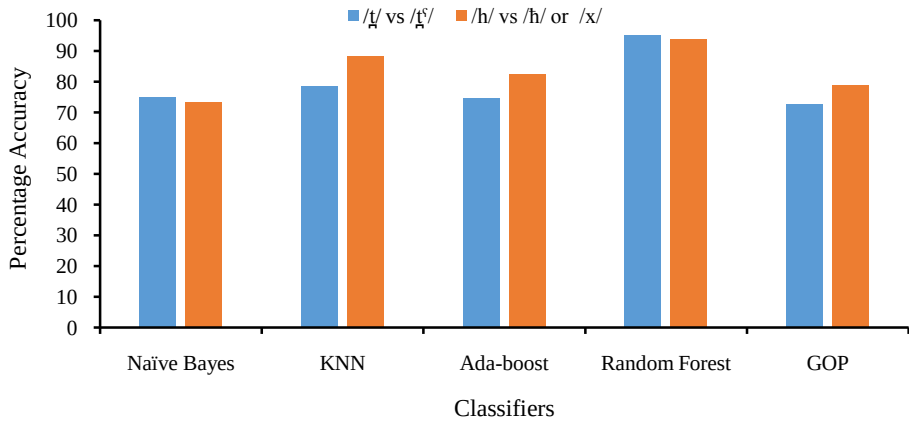

Fig. 1: Accuracy for each confusing phoneme using different methods

all other classifiers by a large margin that shows it can handle this complex problem well. On the other hand, Naïve Bayes produced the worst results. The pattern in results shows that a simple classifier such as Naïve Bayes cannot cope with the complexity of the problem while Ada-boost also performed reasonably well because it uses many weak classifiers to build a strong classifier. The excellent results for Random Forest are also due to its ensemble nature. Mispronunciation problem is a complex problem and therefore, more complex classifiers are required to handle this problem. A classifier based mispronunciation detection suffers a major drawback that it requires a separate classifier for each pronunciation contrast pair. On the other hand, GOP based systems can only predict the proficiency level of the speaker and cannot point out the pronunciation error. Therefore, to point out pronunciation error type, a mispronunciation detection system is required that can automatically selects discriminative pronunciation features.

The performance of the proposed system has also been compared with existing systems. These existing systems include both Arabic mispronunciation detection systems [1, 24] and classifier based mispronunciation detection systems [4, 8]. To this end, accuracy is used to compare the performance of the mispronunciation detection systems. Details of datasets used by each researcher are presented in Table 4. The datasets used in these systems are not publically available.

Therefore, comparisons were made using all the parameters that are generally used in CALL systems. These parameters include; language, method, the number of target phonemes, no. of speakers, total no. of phonemes and average accuracy. It can be seen from Table 3 that the proposed system performs better than the existing systems. The effectiveness of the proposed system is because of the use of generic and potentially discriminative features for pronunciation errors. The number of target phonemes is generally less because of the difficulty level associated with it. Only Al-Hindi [1] have developed a system that covers 5 target phonemes using state of the art GOP technique. In this paper, we have also covered 5 target phonemes using classifier-based approach and achieved better performance. It can be concluded from the results that classifier based mispronunciation detection systems can outperform standard GOP based methods. The better performance of the proposed systems as compared to non-statistical APF based classification methods shows that generic pronunciation features are more effective in such systems.

6. Conclusion and Future Work

This paper presented a classifier-based approach to differentiate between the correct pronunciation of the most confusing Arabic pairs (/t̪/ vs /t̪ˤ/) and (/h/ vs /ħ/ or /x/) for Pakistani Arabic speakers. An APF based super-vector was formed consisting of MFCCs, zero-crossing,

pitch, spectral features and energy features. Four different classifiers were used (Random Forest, Naïve Bayes, K-NN and Ada-boost) on a dataset of Pakistani speakers learning Arabic as their second language. A baseline GOP based system was also developed for comparison. The results show that Random Forest gives the best average accuracy for both confusing Arabic pairs as compared to Naïve Bayes, K-NN, Ada-boost and GOP. A comparative analysis was also conducted with existing state of the art systems and the best performing classifier (Random Forest) outperformed existing Arabic language learning systems with significant margins. This simple super-vector based approach proved very useful but there is a strong need to develop a platform where discriminative APF related to a specific confusing pair can automatically be selected. As the classifier based approach requires a single classifier for each phoneme pair, another future avenue for this area can be to reduce the number of classifiers required to develop such systems.

References

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

[2] S. Wei, G. Hu, Y. Hu and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models", Speech Communication, vol. 51, pp. 896-905, 2009.

[3] F. Zhang, C. Huang, F.K. Soong, M. Chu and R. Wang, "Automatic mispronunciation detection for Mandarin", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 5077-5080, 2008.

[4] H. Strik, K. Truong, F. De Wet and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection", Speech Communication, vol. 51, pp. 845-852, 2009.

[5] S. M. Witt", Automatic error detection in pronunciation training: Where we are and where we need to go", Proc. IS ADEPT, vol. 6, 2012.

[6] W. Hu, Y. Qian, F. K. Soong and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, vol. 67, pp. 154-166, 2015.

[7] S.M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", Speech communication, vol. 30, pp. 95-108, 2000.

[8] K. Truong, A. Neri, C. Cucchiarini and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach", InSTIL/ICALL Symposium, 2004.

[9] H. M. A. Tabbaa and B. Soudan, "Computer-Aided Training for Quranic Recitation", Procedia-Social and Behavioral Sciences, vol. 192, pp. 778-787, 2015.

[10] I. Amdal, M. H. Johnsen and E. Versvik, "Automatic evaluation of quantity contrast in non-native Norwegian speech", SLaTE, pp. 21-24, 2009.

[11] C. Cucchiarini, H. Strik and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology, "The Journal of the Acoustical Society of America, vol. 107, pp. 989-999, 2000.

[12] H. Franco, L. Neumeyer, M. Ramos and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning", EUROSPEECH, 1999.

[13] A.M. Harrison, W.-K. Lo, X. Qian and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training", SLaTE, pp. 45-48, 2009.

[14] A. Neri, C. Cucchiarini and H. Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training", IRAL-International Review of Applied Linguistics in Language Teaching, vol. 44, pp. 357-404, 2006.

[15] S.-Y. Yoon, M. Hasegawa-Johnson and R. Sproat, "Landmark-based automated pronunciation error detection", Interspeech, pp. 614-617, 2010.

[16] H. Franco, L. Neumeyer, V. Digalakis and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality", Speech Communication, vol. 30, pp. 121-130, 2000.

[17] A. Ito, Y.-L. Lim, M. Suzuki and S. Makino, "Pronunciation error detection method based on error rule clustering using a decision tree", pp. 173-176, 2005.

[18] C. Hacker, T. Cincarek, A. Maier, A. HeBler and E. Noth, "Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children", IEEE Int. Conf. on Acoustics, Speech and Signal Processing-ICASSP'07, pp. IV-197-IV-200, 2007.

[19] A. M. Harrison, W. Y. Lau, H. M. Meng and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer", INTERSPEECH, pp. 2787-2790, 2008.

[20] X. Qian, H. M. Meng and F. K. Soong, "The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training", INTER SPEECH, pp. 775-778, 2012.

[21] X. Qian, F. K. Soong and H. M. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)", INTERSPEECH, pp. 757-760, 2010.

[22] O. Ronen, L. Neumeyer and H. Franco, "Automatic detection of mispronunciation for language instruction", EUROSPEECH, 1997.

[23] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training", IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 5049-5052, 2012.

[24] S.M. Abdou, S.E. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, et al., "Computer aided pronunciation learning system using speech recognition techniques", INTERSPEECH, pp. 849-852, 2006.

[25] S. Abdou, M. Rashwan, H. Al-Barhamtoshy, K. Jambi and W. Al-Jedaibi, "Enhancing the Confidence Measure for an Arabic Pronunciation Verification System", Proc. of the Int. Symp. on Automatic Detection of Errors in Pronunciation Training, pp. 6-8, 2012.

[26] J. van Doremalen, C. Cucchiarini and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources", IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 580-585.

[27] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum", Acoustics, Speech and Signal Processing, ICASSP'01, pp. 73-76, 2001.

[28] A. Joseph and R. Sridhar, "Performance Evaluation of Various Classifiers in Emotion Recognition Using Discrete Wavelet Transform, Linear Predictor Coefficients and Formant Features", Advances in Computational Intelligence: Proceedings of International Conference on Computational Intelligence, pp. 373-382, 2015.

[29] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition", Int. J. Adv. Res. Engg. & Tech., vol. 1, pp. 1-4, 2013.

[30] D. Shete, S. Patil, and S. Patil, "Zero crossing rate and Energy of the Speech Signal of Devanagari Script", IOSR-JVSP, vol. 4, pp. 1-5, 2014.

[31] S. Zahid, F. Hussain, M. Rashid, M. H. Yousaf and H. A. Habib, "Optimized audio classification and segmentation algorithm by using ensemble methods", Mathematical Problems in Engg., vol. 2015, 2015.

[32] A. I. Al-Shoshan, "Speech and music classification and separation: a review", Journal of King Saud University, vol. 19, pp. 95-133, 2006.

[33] M.A. Ghazanfar, "Experimenting switching hybrid recommender systems", Intelligent Data Analysis, vol. 19, pp. 845-877, 2015.

[34] A.M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," Computers in Human Behavior, vol. 65, pp. 267-275, 2016.

[35] K.J. Archer and R.V. Kimes, "Empirical characterization of random forest variable importance measures", Computational Statistics & Data Analysis, vol. 52, pp. 2249-2260, 2008.

[36] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, and A. s. D. N. Initiative, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease", NeuroImage, vol. 65, pp. 167-175, 2013.

[37] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al., "The HTK book", Cambridge University Engineering Department, vol. 3, p. 175, 2002.