Indian Journal of Science and Technology, Vol 8(20), DOI: 10.17485/ijst/2015/v8i20/78959, August 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# Robust Accent Recognition in Malaysian English using PCA-Transformed Mel-Bands Spectral Energy Statistical Descriptors

**M. A. Yusnita[1*], M. P. Paulraj[2], Sazali Yaacob[3], R. Yusuf[1] and M. Nor Fadzilah[1]**

[1]Faculty of Electrical Engineering, Universiti Teknologi MARA Malaysia, Permatang Pauh - 13500, Penang, Malaysia; yusnita082@ppinang.uitm.edu.my
[2]School of Mechatronic Engineering, University Malaysia Perlis, Ulu Pauh - 02600, Perlis, Malaysia
[3]Universiti Kuala Lumpur Malaysian Spanish Institute, Kulim Hitech Park, Kulim - 09000, Kedah, Malaysia

## Abstract

The standard speech feature extractors such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficients (LPC) fail to perform well under noisy conditions. In this paper two noise less-susceptible features are proposed to mitigate the deficiency of MFCC and LPC. Statistical descriptors of Mel-Bands Spectral Energy (MBSE) is applied to the traditional filter-bank analysis, however, this technique increases the feature size. This issue is tackled by proposing a transformation using principle component analysis to generate a new PCA-MBSE feature set. Two types of utterances namely isolated words and continuous speech were elicited from 103 university volunteers in a controlled room to collect speech signals from three main ethnic groups in Malaysia. This study employed two classifiers namely K-nearest neighbors and artificial neural networks to recognize between the Malay, Chinese and Indian accents. Experimental results using independent test samples technique indicated promising accuracy rates of 92.7% and 93.0% using the proposed PCA-MBSE features to recognize between the Malay, Chinese and Indian accents on the male and female datasets respectively. It was found that under severe noisy conditions, the standard MFCC and LPC features started to deteriorate faster than the MBSE-based features. PCA-MBSE features showed the most robust quality where its performance was just slightly deteriorated by 17.1% and 13.6% as compared to MBSE features i.e. 33.1% and 31.3% on the male and female datasets respectively. Further poor results of LPC features were obtained indicating deterioration rates of 40.2% and 32.7%, while that of MFCC features of 35.7% and 36.8% for the male and female datasets respectively. As a conclusion, Malaysian English is a not a uniform English variety colored by its diverse ethnic nuances. Incorporating accent analyzers using the proposed techniques in automatic speech recognition can contribute a substantial improvement in noisy environment.

**Keywords:** Accent Recognition, K-Nearest Neighbors, Linear Prediction Coefficients, Malaysian English, Mel-Bands Spectral Energy, Mel-Frequency Cepstral Coefficients, Principle Component Analysis

## 1. Introduction

Accent in a particular language is a learned or behavioral property rather than physiological or organic factor of human speech. Its presence receives different attitude perceptions towards listeners in many life aspects[1–3] as well as it is a known factor that affects the intelligibility and understanding of speech. Differences in pronunciations between different ethnics or races are fascinating that can relate to unique identity of a particular community in a country as defined by Schneider[4]. This implies that accent can be used as one of the important properties in human speech biometrics, well known as voiceprint in speaker recognition systems and also for corrective strategy in tackling speech variability in Automatic Speech Recognition (ASR) systems. Malaysian English (MalE) is a localized accent burgeoned from the British English and the American English[5]. The phenomenon of MalE arises from the fact that Malaysia is a multi-ethnic and multi-lingual society composed of 28.3 million population of 50.1% Malays, 22.6% Chinese, 6.7%

*Author for correspondence

Indians and others[6]. As a result, Malaysians speak English with local nuances of the Malay (with regional influence), Chinese (Mandarin, Hokkien, Teochew, Cantonese, Hakka) and Indian (Tamil, Telegu, Malayalam) accents[7,8].

Incorporating accent analyzer has proven to greatly improve the ASR performance[9,10]. Extracting accent features using Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficients (LPC) with modeling methods such as hidden Markov model, Gaussian mixture model, support vector machine and Artificial Neural Network (ANN) to recognize English accents, Australian accents, Persian accents and Japanese and Chinese accents have been reported in previous studies. Nevertheless, there are very scarce empirical studies in MalE accents[11]. Albeit the popularly used MFCC shows excellent performance, Tufekci and Gowdy[12] mentioned a few drawbacks of MFCC features and proposed a hybrid between MFCC and discrete wavelet transform to solve phoneme recognition task in noisy space. Another attempt for robustness was demonstrated by Gupta and Gilbert[13] using wavelet coefficients. Whilst Nhat and Lee[14] proposed determination of Mel-filter coefficients by Principle Component Analysis (PCA) to improve the limitation of the conventional filter shape in MFCC. While LPC is considered as older method than MFCC, however, it yielded poorer results[15] despite its simplicity and more robust than the earlier[16–18].

This paper proposes a new feature extractor using filter-bank analysis which uses only simple statistical descriptors of Mel-Bands Spectral Energy (MBSE) which mitigate the problem of noise-susceptibility. However this will increase the size of statistical descriptors by four-fold. The redundancies in the MBSE can further be compressed using PCA. This technique resulted in another new feature set, so called PCA-MBSE. Robustness analysis is conducted to compare the proposed features with the standard MFCC and LPC features. The rest of this paper contains the methods used to extract and reduce features, followed by results and discussion of experiments conducted and will be ended with conclusions of this work.

## 2. Model and Evaluation Methods for Automatic Accent Recognition

This paper considers gender specific Automatic Accent Recognition (AAR) system and investigate its performance using two types of utterances i.e. isolated-words (IWs) and continuous speech in the form of sentences (STs). For all evaluations of the accent recognizer performance based on different feature sets, independent test samples method is used by partitioning the feature database into 60–40 % ages of training and testing datasets respectively using K-Nearest Neighbors (KNN) algorithm. For ANN algorithm, standard practice is to divide the whole data into 60% training dataset, 15% validation dataset and the remaining 25% is assigned as testing dataset[19, 20]. To average out any source of randomness in the learning methods, ten runs per experiment are conducted for the same architecture (any parameter change) and the minimum (min), mean and maximum (max) of all Classification Rates (CRs) are calculated and reported. The description of database of MalE accents[21, 22] used in this research is summarized in Table 1.

## 3. Feature Extraction and Reduction

The feature generation part implements two stages in preparing salient features for recognition stage namely, feature extraction using the proposed statistical descriptors of MBSE and feature reduction using PCA to produce a new feature set so forth referred as PCA-MBSE features.

This new proposed statistical descriptors of MBSE, is formulated from four descriptive statistics namely mean, Standard Deviation (STD), kurtosis (kurt) and the Ratio

**Table 1.** Distribution of speakers in Malaysian English accents database in terms of ethnic group, Gender and type of utterance.

| Ethnic | Gender | No of Speakers | No of utterances | |
|---|---|---|---|---|
| | | | IWs | STs |
| Malay | Male | 16 | 1440 | 738 |
| | Female | 22 | 1620 | 918 |
| | Total | 38 | 3060 | 1656 |
| Chinese | Male | 19 | 1705 | 849 |
| | Female | 15 | 1350 | 765 |
| | Total | 34 | 3055 | 1614 |
| Indian | Male | 16 | 1440 | 816 |
| | Female | 15 | 1350 | 765 |
| | Total | 31 | 2790 | 1581 |
| Total | Male | 51 | 4585 | 2403 |
| | Female | 52 | 4320 | 2448 |
| | Total | 103 | 8905 | 4851 |

of the Standard Deviation to Kurtosis (RSDK) of the log-energies of the Mel-bands. By doing this, nonlinear homomorphic processing of the cepstrals like in MFCC, which is quite sensitive to the background noise can be skipped. The block diagram describing the procedures of PCA-MBSE extraction is depicted in Figure 1.

The log-energy output of each Mel-band as resulted from the third block of the PCA-MBSE processors is calculated as in Equation (1).

$$W(i)=\sum_{f_L}^{f_H}\log|E_m|.H_i\left(\frac{2\pi m}{N_{FB}}\right) \quad (1)$$

where $W(i)$ denotes the output of the Mel-warped log energy for the $i^{th}$ Mel-band, $E_m$ refers to the FFT power spectrum and $H_i(.)$ is the transfer function of the $i^{th}$ Mel-band with $m$ as the FFT sample index and $N_{FB}$ is the number of filters in the filter bank. This summation is done between the lower $f_L$ and upper $f_H$ frequencies of each filter with nonzero coefficients.

The statistical descriptors[23, 24] can be calculated using Equation (2) to Equation (4) to obtain the statistical descriptors of the $i^{th}$ log-energy of the Mel-bands for all the variables defined in Equation (1) above. The mean value is the most common and simplest way to express the central tendency of the log-energy distribution and it is denoted as $\bar{W}$ in Equation (2).

$$\bar{W}(i)=\frac{1}{K}\sum_{k=1}^{K}W_k \quad (2)$$

where for any signal, $k$ is the frame index $k$ = 1, 2, …, K and $i$ denotes the band index $i$ = 1, 2, 3, …, $N_{FB}$ of the Mel-filters.

The standard deviation describes how the log-energy is distributed wherein it measures the spread from the mean value. Firstly, the variance $V$ of this distribution is defined in Equation (3).

$$V(W)=\frac{1}{(K\text{-}1)}\sum_{k=1}^{K}\left(W_k-\bar{W}\right)^2=\overline{\left(W_k-\bar{W}\right)^2} \quad (3)$$

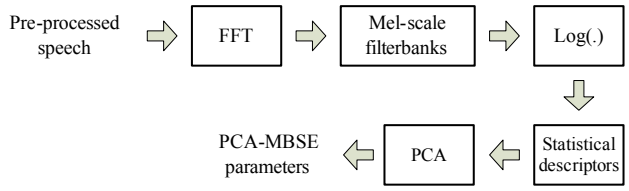



**Figure 1.** Block diagram of PCA-MBSE feature extraction.

Taking the square root of the variance is called STD denoted as $\sqrt{V}=\sigma$. It is the typical deviation that one expects to see from the measured average value.

The other descriptor that provides information concerning the distribution of this log-energy is kurtosis denoted as kurt. It provides the 'peakedness' of the distribution. Positive kurtosis values indicate the distribution is rather peaked, clustered in the center, with long thin tails. On the other hand, the negative values indicate a relatively flat distribution with too many samples in the extremes. Kurtosis is the 4th order central moment and is expressed in Equation (4).

$$kurt(W)=\frac{1}{\sigma^4}\overline{\left(W-\bar{W}\right)^4}-3 \quad (4)$$

## 4. PCA-MBSE Algorithm

Below is the specific algorithm implemented in this paper to extract the MBSE and PCA-MBSE feature vectors of the accented speech signals to produce $l$-dimensional MBSE feature database.

**Step 1:** Initially, silence and unvoiced parts of a speech signal is removed after frame-blocking into the frame length of 32 msec with the frame shift of 16 msec leaving only voiced part for further processing.

**Step 2:** Pre-emphasis filtering is applied to the voiced frames using first-order FIR with emphasis coefficient of 15/16 to compensate the attenuation in the spectral energy by 6 dB per octave. Then, Hamming-windowing is applied to smooth out the signal transition at both edges of a frame.

**Step 3:** Spectrum is computed from the pre-processed frames using Fast Fourier Transform (FFT) algorithm which is very efficient in computing the DFT coefficients.

**Step 4:** In the frequency domain, each region of spectrum of interest is sampled by triangular-shaped windows which are centered linearly in the Mel-scale using the Mel-scale warped filters.

**Step 5:** Logarithmic (log) of spectral energy is calculated from the outputs of the Mel filter bank resulted in Step 4.

**Step 6:** For each signal, four statistical descriptors of the log-energy of the Mel-band are computed from all frames within a signal by using Equation (2) to Equation (4).

**Step 7:** Repeat *Step 6* for all other Mel-bands. Then this will form an $l = N_{FB} \times 4$ dimensions of feature vector.

**Step 8:** Repeat *Step 1* to *7* for the other samples of speech signals and form a matrix of feature vectors with assigned class attribute to each feature vector.

The $l$-dimensional MBSE feature vector can be represented as in Equation (5).

$$x^n(l)=\left[f_{11}, f_{12}, f_{13}, f_{14}, f_{21}, f_{22}, f_{23}, f_{24} \ldots, f_{ij}, \ldots, f_{IJ}\right]^{\mathrm{T}} \quad (5)$$

where $x$ is the feature vector, $f(i, j)$ is the statistical descriptors of Mel-bands spectral energy of dimension $l = 1, 2, 3, \ldots, \mathrm{I} \times \mathrm{J}$ .

The indices $i$ and $j$ represent the band number $i$ = 1, 2, 3, …, $\mathrm{N_{FB}}$ and the statistical descriptors $j$ = {1 for mean, 2 for STD, 3 for kurt and 4 for RSDK} respectively, evaluated for $n^{\mathrm{th}}$ signal and T denotes transposition.

For extracting PCA-MBSE features, continue these two steps from the above Step 7.

**Step 9:** Apply PCA Algorithm[25] to the resulted matrix in Step 8 above.

# 5. Results and Discussion

This paper reports two types of speech material used to elicit accented speech from male and female speakers namely, IWs and STs. For Automatic Accent Recognizer (AAR) designed using KNN, different accent groups were labeled as 1 for the Malay class, 2 for the Chinese class and 3 for the Indian class respectively. As for ANN, the input layer consisted of either one of this number of input neurons namely n = 72 (MBSE) and n = 72 and 52 (PCA-MBSE) which corresponds to the number of input features. The output layer was consisted of m = 3 nodes wherein activation of a node represented an accent individually. In order to evaluate the efficacy of MBSE- and PCA-MBSE-based features for modeling accent recognizer, the analysis and discussion are divided into four experiments as follows:

## 5.1 Varying K-Parameter of KNN

In this experiment, the effect of K-parameter of KNN was investigated by fixing the Mel-filters $\mathrm{N_{FB}}$ to 20 and distance metric to Cityblock. Table II tabulates the resulted statistical CRs namely min CR, mean CR and max CR of the 72-MBSE using KNN model by varying the nearest neighbors K = 1 to 10, measured for the 60-40 percentage of independent test samples on the IWs speech and the STs speech of different genders respectively. The performances are compared in Figure 2. From this graph, the results suggested that K = 1 or 2 resulted the best performance for this database and the performance deteriorated as K-parameter increases. It is worth noting that, referring to K=2 of mean CRs, the formulated STs speech gained higher accuracy rates of between 1.1% and 1.6% as compared to the IWs speech. Meanwhile, the male speakers had higher accuracy rates of between 1.6% and 2.1% as compared to the female speakers. The highest CRs achieved for these four speech test scenarios were highlighted in Table 2.

## 5.2 Varying KNN Distance Metric

In this experiment, the effect of different choices of distance metric of the KNN parameter was investigated using K-parameter of 2 for the 72-MBSE features. Each dataset of four speech test scenarios were tested against four different distance metrics and the results of mean CRs are depicted in Figure 3. From the results, it is worth noting that for all test scenarios, Cityblock distance was leading in the performance, followed by cosine and correlation distances and the worst case was yielded using Euclidean distance. The improvements made in the mean CR across four speech test scenarios using Cityblock distance varied from 4.8% to 7.8% as compared to cosine distance, 5.4% to 7.9% as compared to correlation distance and 7.1% to 9.2% as compared to Euclidean distance. The details of statistical CRs of all testing conditions across different distance metric are tabulated as in Table 3 with the best distance metric highlighted.

## 5.3 Varying Principle Components

In this experiment, the dimension of 72-MBSE feature set was reduced using PCA to transform the features into a new PCA-MBSE feature set. The PCA-transformed

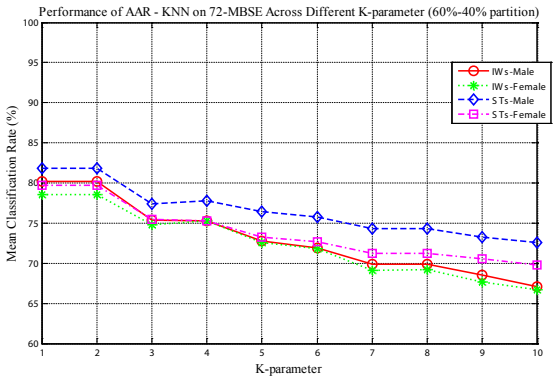



**Figure 2.** Performance of the 72-MBSE of four speech test scenarios across different K-parameter.

**Table 2.** Performance of AAR on 72-MBSE-based KNN model across different K-parameter

| K-parameter | IWs Speech | | | | | | ST Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | | Female | | | Male | | | Female | | |
| CR Statistics | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max |
| 1 | 78.5 | 80.2 | 82.2 | 76.7 | 78.6 | 80.0 | 78.8 | 81.8 | 84.6 | 76.7 | 79.7 | 82.1 |
| 2 | 78.5 | 80.2 | **82.2** | 76.7 | 78.6 | **80.0** | 78.8 | 81.8 | **84.6** | 76.7 | 79.7 | **82.1** |
| 3 | 73.7 | 75.4 | 77.0 | 72.1 | 74.8 | 76.0 | 74.7 | 77.4 | 79.8 | 74.1 | 75.5 | 77.3 |
| 4 | 73.7 | 75.3 | 76.6 | 73.2 | 75.3 | 76.9 | 75.3 | 77.8 | 80.1 | 73.1 | 75.3 | 78.8 |
| 5 | 70.4 | 72.8 | 73.9 | 70.4 | 72.6 | 73.7 | 73.4 | 76.4 | 79.4 | 70.8 | 73.3 | 76.2 |
| 6 | 70.2 | 71.9 | 72.8 | 69.2 | 71.8 | 73.7 | 73.4 | 75.8 | 78.0 | 70.3 | 72.7 | 77.2 |
| 7 | 68.2 | 69.9 | 71.1 | 66.4 | 69.1 | 71.6 | 72.1 | 74.3 | 76.3 | 68.5 | 71.2 | 74.4 |
| 8 | 67.7 | 69.9 | 71.1 | 67.4 | 69.2 | 71.1 | 71.2 | 74.3 | 76.9 | 68.3 | 71.2 | 74.7 |
| 9 | 66.1 | 68.5 | 70.0 | 66.0 | 67.7 | 69.0 | 70.7 | 73.3 | 75.3 | 68.3 | 70.6 | 74.1 |
| 10 | 65.9 | 67.1 | 68.5 | 65.0 | 66.7 | 68.3 | 70.1 | 72.6 | 75.0 | 66.9 | 69.8 | 73.9 |

**Table 3.** Performance of AAR on 72-MBSE-based KNN model using different distance metric

| Distance metric | IWs Speech | | | | | | ST Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | | Female | | | Male | | | Female | | |
| CR Statistics | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max |
| Euclidean | 71.9 | 73.0 | 74.9 | 69.7 | 71.5 | 73.6 | 70.9 | 72.6 | 74.5 | 69.3 | 71.0 | 73.2 |
| Cityblock | 78.5 | 80.2 | **82.2** | 76.7 | 78.6 | **80.0** | 78.8 | 81.8 | **84.6** | 76.7 | 79.7 | **82.1** |
| Cosine | 73.7 | 75.4 | 77.0 | 71.6 | 72.8 | 73.8 | 73.2 | 75.2 | 77.7 | 69.7 | 71.9 | 74.7 |
| Correlation | 73.4 | 74.8 | 76.4 | 71.2 | 72.7 | 73.9 | 72.5 | 75.0 | 76.6 | 70.3 | 71.8 | 74.0 |

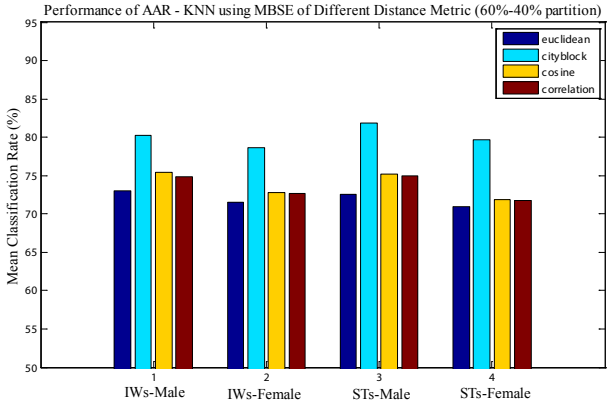



**Figure 3.** Performance of the 72-MBSE dataset across different distance metric of KNN under four speech test scenarios.

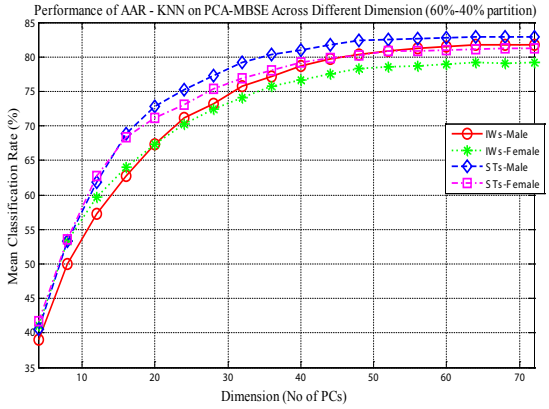



**Figure 4.** Performance of PCA-MBSE test dataset across different dimension (PCs) of four speech test scenarios using KNN model.

MBSE feature set was varied from the original size of 72 and stepped down to 68, 64, 60, … 4 number of Principle Components (PCs) to determine the performance on four speech test scenarios. Figure 4 shows the performance by fixing K-parameter to 2 using Cityblock distance for varying PCs from 72 to 4. It shows that the performance was gradually dropping as the number of PCs reduced from 72 to 32. After 32 PCs, the performance dropped drastically for every four subsequent reduction in the PCs. Table 4 tabulates statistical CRs of all testing conditions at selected PCs and the best matches of performance to that of the original MBSE were highlighted in this table as the suggested dimension of PCA-MBSE. The transformed 72-PCA-MBSE exhibited only a small increment in the mean CRs of 0.6% to 1.5% as compared to the original MBSE across these four testing conditions.

**Table 4.** Performance of AAR on PCA-transformed MBSE-based KNN model across different PCs

| PCs | IWs Speech | | | | | | ST Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | | Female | | | Male | | | Female | | |
| CR Statistics | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max |
| 12 | 55.1 | 57.2 | 58.9 | 58.0 | 59.7 | 61.1 | 59.6 | 61.9 | 63.1 | 60.5 | 62.7 | 65.6 |
| 32 | 73.2 | 75.7 | 77.0 | 72.4 | 74.1 | 75.8 | 76.9 | 79.2 | 81.1 | 74.4 | 76.9 | 78.8 |
| 52 | 79.0 | 80.8 | **82.1** | 76.6 | 78.6 | **80.4** | 80.4 | 82.5 | **85.8** | 79.0 | 80.8 | **82.5** |
| 72 | 79.8 | 81.7 | 83.4 | 77.5 | 79.2 | 81.5 | 80.4 | 82.9 | 85.8 | 79.7 | 81.2 | 83.0 |

## 5.4 MBSE-Based ANN Classifier

Next, the formulated feature vectors of 72-MBSE, 72-PCA-MBSE and 52-PCA-MBSE were tested using ANN of two-layer feed-forward multilayer perceptron. The number of hidden neurons p used in both IWs and STs speech was set to 35 units and the learning and momentum rates utilized here were $\alpha = 0.5$ and $\beta = 0.9$ throughout all testing conditions. Figure 5 compares the performance (mean CRs) of these feature vectors using maximum (max) criterion and Threshold and Margin (T&M) criterion across four speech test scenarios. In addition, Table 5 and Table 6 tabulate the details of statistical CRs using these methods. These results

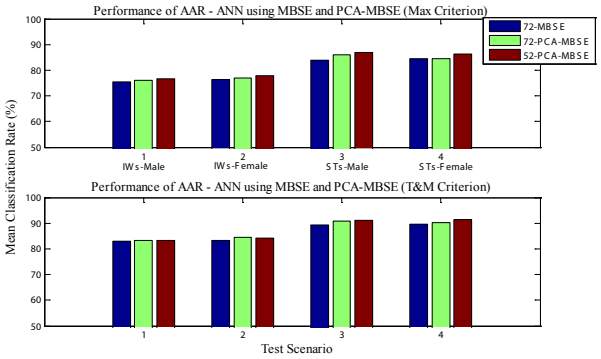



**Figure 5.** Performance of two different output neuron assignment methods using MBSE-Based and PCA-MBSE-based AAR-ANN of four speech test scenarios.

**Table 5.** Performance of AAR on MBSE-based ANN and PCA-MBSE-based ANN models using maximum criterion for the output neuron state

| Speech mode | IWs Speech | | | | | | ST Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | Male | | | Female | | | Male | | | Female | | |
| CR Statistics | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max |
| 72-MBSE | 74.7 | 75.6 | 76.3 | 75.5 | 76.3 | 76.9 | 83.2 | 83.8 | 84.5 | 81.4 | 84.5 | 88.3 |
| 72-PCA-MBSE | 74.1 | 76.2 | 78.5 | 76.1 | 77.1 | 78.1 | 84.0 | 86.0 | 88.5 | 82.8 | 84.6 | 86.3 |
| 52-PCA-MBSE | 73.8 | 76.8 | **78.2** | 75.5 | 77.8 | **80.2** | 85.2 | 86.78 | **88.2** | 85.1 | 86.3 | **87.4** |

**Table 6.** Performance of AAR on MBSE-based ANN and PCA-MBSE-based ANN models using threshold and margin criterion for the output neuron state

| Speech mode | IWs Speech | | | | | | ST Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | Male | | | Female | | | Male | | | Female | | |
| CR Statistics | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max | Min | Mean | Max |
| 72-MBSE | 81.9 | 82.9 | 84.6 | 82.24 | 83.2 | 84.7 | 88.4 | 89.4 | 89.9 | 88.0 | 89.5 | 90.2 |
| 72-PCA-MBSE | 81.1 | 83.2 | 85.3 | 84.05 | 84.6 | 85.8 | 89.4 | 90.9 | 92.0 | 89.6 | 90.3 | 91.1 |
| 52-PCA-MBSE | 81.3 | 83.4 | **84.8** | 81.55 | 84.3 | **85.6** | 89.6 | 91.1 | **92.7** | 89.7 | 91.4 | **93.0** |

emphasized the best reduced dimension of PCA-MBSE as more than just as accurate as the original MBSE. From the tabulated results, it is worth noting that T&M criterion produced better performance than that of max criterion approximately by 4.3% to 6.6% increase in the mean CR with reference to the 52-PCA-MBSE. In addition to this, the unclassified cases according to speech test scenario were approximately 22.7%, 21.9%, 13.2% and 17.8% in average for the IWs-male, IWs-female, STs-male and STs-female respectively using T&M criterion.

Generally, the STs Speech outperformed the IWs speech for both genders approximately by 7.1% to 7.7% of the mean CR on the reduced features using T&M criterion. Unlike using KNN, the performance of different genders on both IWs speech and ST speech were quite comparable. With reference to the 52-PCA-MBSE, the best accuracies were yielded for the STs-female speech i.e. 93.0%, followed by the STs-male i.e. 92.7%, the IWs-female i.e. 85.6% and lastly the IWs-male i.e. 84.8% using a noise-tolerant T&M method of measuring the success.

Next, the performance of MBSE-based and PCA-MBSE-based ANN for AAR can be measured using training time (epoch). Figure 6 shows the min, mean and max epochs resulted from ten run of training each of the test speech scenarios using 72-MBSE, 72-PCA-MBSE and 52-PCA-MBSE. It is worth noting that the training time was fast using Levenberg-Marquardt learning algorithm with overall min and max epoch of 15 to 88 epochs attempted for all speech test scenarios. It was observed that the reduced-dimension of 52-PCA-MBSE completed the training faster averagely by 27.4 to 36.4 epochs as compared to the original 72-MBSE averagely by 40.4 to 70.4 epochs across all testing conditions.

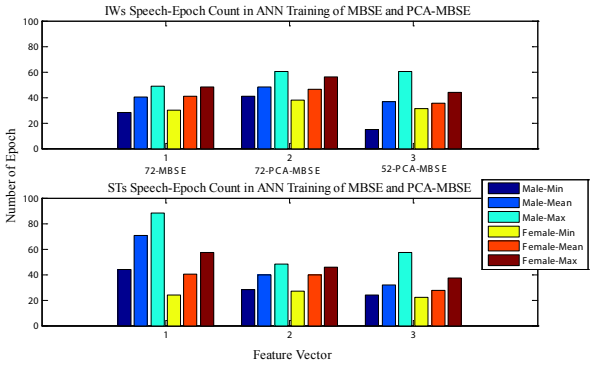



**Figure 6.** Training time performance (epoch) of ANN of the IWs and STs speech scenarios for MBSE and PCA-MBSE-based features.

## 5.5 Performance Comparison under Noisy Conditions

Next, the performance of MBSE and PCA-transformed MBSE are discussed under clean and noisy conditions using KNN classifier in comparison to the MFCC and LPC features. In the quest to test the robustness quality, seven levels of noisiness are presented here. The test datasets which constituted of 40% of the overall audio volume were corrupted with additive white Gaussian noise (AWGN) to stimulate as background noise in the real environment. The performance of different feature sets under different level of Signal-to-Noise (SNR) ratio of 35 dB, 30 dB, 25 dB, 20 dB, 15 dB, 10 dB and down to 5 dB were tested. Comparisons of susceptibility to noise levels for IWs-speech and ST-speech scenarios are depicted in Figure 7 to Figure 10 in terms of degradation in the mean CRs for the male and female datasets respectively.

Generally speaking, from observations made of the aforementioned figures, all feature sets under test were quite less susceptible to AWGN under noisy conditions between 25 dB to 35 dB. For more noisy conditions of

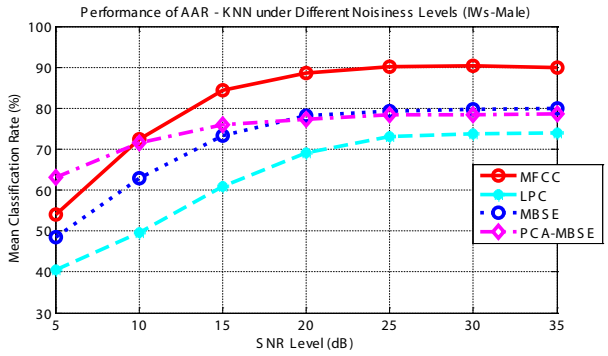



**Figure 7.** Robustness performance of different feature sets for the male-IWs speech under seven noisiness level.

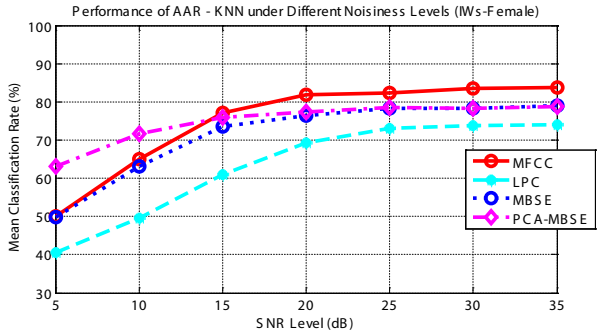



**Figure 8.** Robustness performance of different feature sets for the female-IWs speech under seven noisiness level.

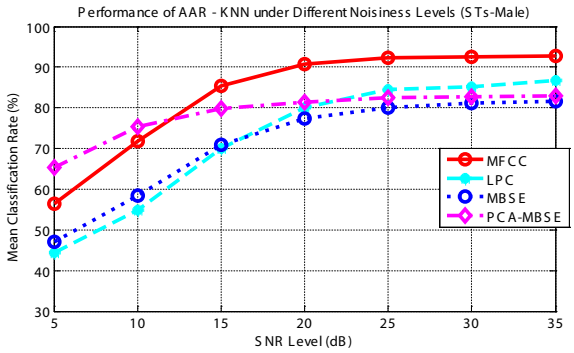



**Figure 9.** Robustness performance of different feature sets for the male-STs speech under seven noisiness level.

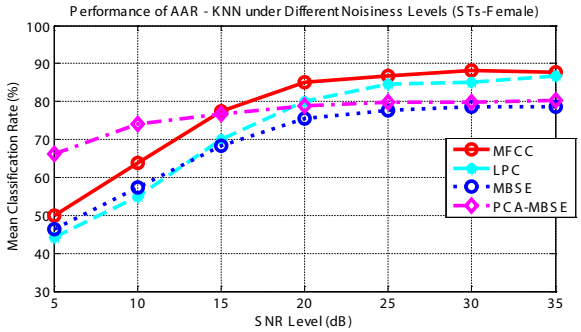



**Figure 10.** Robustness performance of different feature sets for the female-STs speech under seven noisiness level.

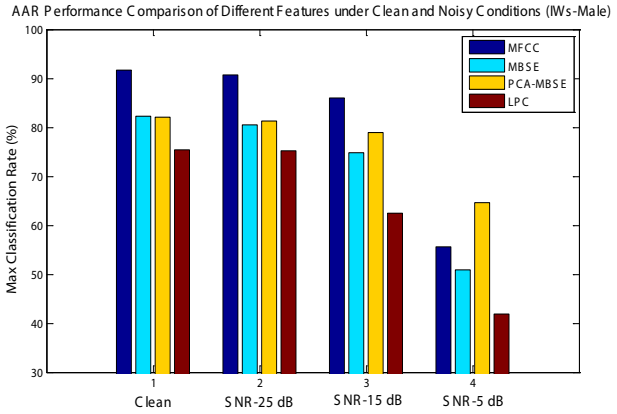
below 25 dB of SNR, MFCC and LPC features started to deteriorate faster than MBSE-based features. Among these four feature sets, PCA-transformed MBSE deteriorated just moderately. Comparing the percentage of drop in the performance of mean CR between 25 dB and 5 dB for the IWs speech, PCA-MBSE only experienced 16.9% and 15.4% drops for the male and female datasets respectively as compared to MBSE of 30.6% and 28.4% drops. The standard LPC features deteriorated badly at 32.4% and 26.9% drops for the respective genders and the results were even worst for the standard MFCC features i.e. 36.2% and 32.3% drops for the male and female datasets respectively.

Similarly for the STs speech, PCA-MBSE only experienced 17.1% and 13.6% drops for the respective genders as compared to MBSE of 33.1% and 31.3% drops. The results were poor for LPC features i.e. 40.2% and 32.7%, while that of MFCC features were 35.7% and 36.8% for the male and female datasets respectively.

Generally both speech modes IWs and STs speech irrespective of gender showed that PCA-MBSE as the most robust features among these four feature sets, followed by MBSE, LPC and MFCC. The overall results are tabulated in Table 7 for comparison. Although MFCC classified better than the proposed MBSE and PCA-MBSE across these four speech test scenarios by approximately 7.6% in average, PCA-MBSE started to outperform MFCC under severe noisy environment of SNR of 10 dB. In most cases, the performance of MBSE-based features surpassed LPC features under both clean and noisy environment. The comparison of accuracy rates achieved using different feature sets can be referred to Figure 11 to Figure 14 for both IWs and STs speech on the male and female datasets for clean and selected SNR levels.

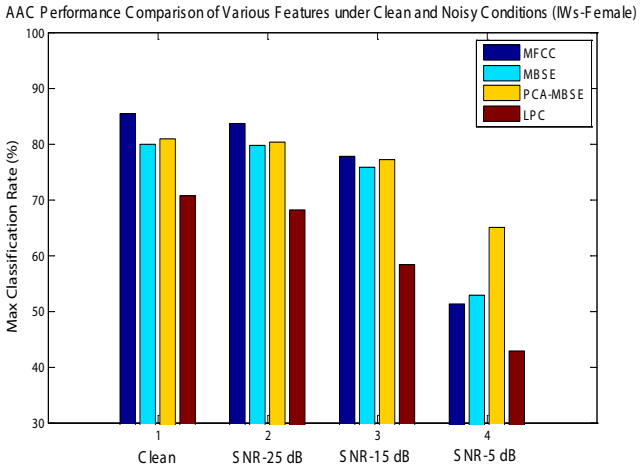
**Table 7.** Performance drop percentage of different feature extractors in mean CR for SNR change between 25-dB and 5-dB

| Feature vector | Mean CR drop from 25 dB to 5 dB (percent) | | | |
|---|---|---|---|---|
| | IWs-Male | IWs-Female | STs-Male | STs-Female |
| MFCC | 36.2 | 32.3 | 35.7 | 36.8 |
| LPC | 32.4 | 26.9 | 40.2 | 32.7 |
| MBSE | 30.6 | 28.4 | 33.1 | 31.3 |
| PCA-MBSE | 16.9 | 15.4 | 17.1 | 13.6 |



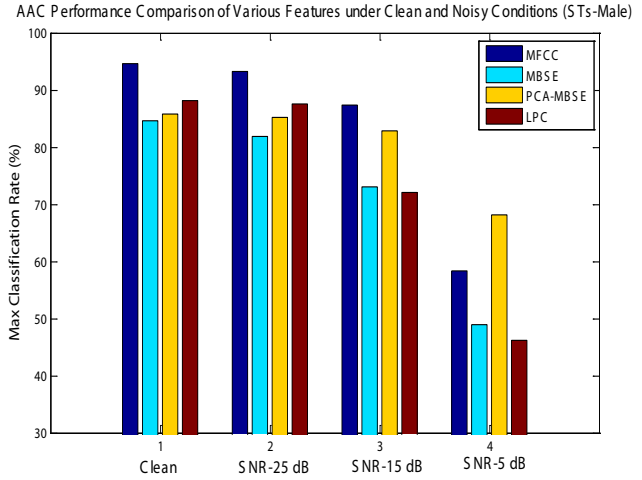

| SNR | MFCC | MBSE | PCA-MBSE | LPC |
|---|---|---|---|---|
| **Clean** | 91.7 | 82.2 | 82.1 | 75.4 |
| **25 dB** | 90.7 | 80.4 | 81.3 | 75.1 |
| **15 dB** | 85.9 | 74.8 | 79.0 | 62.5 |
| **5 dB** | 55.6 | 50.8 | 64.7 | 41.9 |

**Figure 11.** Performance (max CR) of different feature sets under clean and different SNR for IWs speech of male speakers.



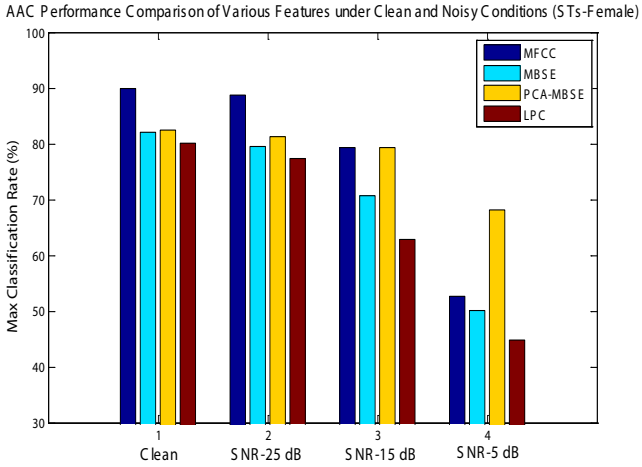

| SNR | MFCC | MBSE | PCA-MBSE | LPC |
|---|---|---|---|---|
| Clean | 85.4 | 80.0 | 80.8 | 70.7 |
| 25 dB | 83.6 | 79.8 | 80.3 | 68.2 |
| 15 dB | 77.7 | 75.7 | 77.1 | 58.4 |
| 5 dB | 51.2 | 52.9 | 65.0 | 42.9 |

**Figure 12.** Performance (max CR) of different feature sets under clean and different SNR for IWs speech of female speakers.




| SNR | MFCC | MBSE | PCA-MBSE | LPC |
|---|---|---|---|---|
| Clean | 94.6 | 84.6 | 85.8 | 88.1 |
| 25 dB | 93.3 | 81.9 | 85.2 | 87.5 |
| 15 dB | 87.3 | 73.0 | 82.8 | 72.1 |
| 5 dB | 58.4 | 48.9 | 68.2 | 46.2 |

**Figure 13.** Performance (max CR) of different feature sets under clean and different SNR for STs speech of male speakers.




| SNR | MFCC | MBSE | PCA-MBSE | LPC |
|---|---|---|---|---|
| Clean | 89.9 | 82.1 | 82.5 | 80.1 |
| 25 dB | 88.7 | 79.5 | 81.3 | 77.3 |
| 15 dB | 79.4 | 70.6 | 79.3 | 62.8 |
| 5 dB | 52.6 | 50.1 | 68.2 | 44.9 |

**Figure 14.** Performance (max CR) of different feature sets under clean and different SNR for STs speech of female speakers.

## 6. Conclusion

This paper has presented a new proposal for extracting and compressing accent features of MalE accented speech using MBSE and PCA-MBSE formulated feature extractors. In current situation, there is still lack of empirical studies to prove that Malaysian speakers of different ethnics speak with their mother tongue influential accents rather than just using bias human observations and perspectives. This research corroborates these assumptions with promising results that accents can be detected accurately from their speech at the best classification rates of 92.7% and 93.0% for the male and female speakers respectively using the proposed PCA-MBSE features. Albeit MFCC classified better than the proposed PCA-MBSE features, approximately by 7.6% in average across four speech test scenarios, under severe noisy conditions however, it is suggested that PCA-MBSE produced a robust quality features for accent recognition of this MalE accents database as compared to the standard MFCC. The proposed features also performed better than the standard LPC features under both clean and noisy conditions. It was found that accent can be detected better using the continuous STs speech than the merely isolated-word

IWs speech in which, the earlier may reflect more natural speech. This paper also suggests that the male speakers possess greater accent severity than the female speakers as evident by consistently better recognition results.

## 7. References


1. Ahmed ZT, Abdullah AN, Heng CS. The role of accent and ethnicity in the professional and academic context. International Journal of Applied Linguistics & English Literature. 2013; 2(5):249–58.
2. Butler YG. How are nonnative-english-speaking teachers perceived by young learners. TESOL Quarterly. 2007 Dec; 41(4):731–55.
3. Mahmud MM, Ching WS. Attitudes towards accented speech among radio deejays in Malaysia. Academic Research International. 2012 May; 2(3):520–30.
4. Schneider EW. Postcolonial English: varieties around the world: Cambridge Univ Press; 2007.
5. Phoon HS. The phonological development of Malaysian English speaking Chinese children: A normative study. Speech Sciences [Ph D thesis]. Christchurch, New Zealand: University of Canterbury. Communication Disorders; 2010:293.
6. Hassan AR. Monthly statistical bulletin Malaysia. Department of Statistics Malaysia. 2012 Feb.
7. McGee K. Attitudes towards accents of English at the British Council, Penang: what do the students want? Malaysian Journal of ELT Research (MELTA). 2009; 5:162–205.
8. Nair-Venugopal S. English, identity and the Malaysian workplace. World Englishes. Oxford, UK: Blackwell Publishers Ltd. 2000 Jul; 19(2):205–13.
9. Arslan LM. Foreign accent classification in American English [PhD thesis]. Department of Electrical and Computer Engineering. Duke University. 1996.
10. Teixeira C, Trancoso I, Serralheiro A. Accent identification. Fourth International Conference on Spoken Language. Philadelphia, PA. 1996. p. 1784–7.
11. Pillai S, Mohd Don Z, Knowles G, Tang J. Malaysian English: an instrumental analysis of vowel contrasts. World Englishes. 2010 Jun; 29(2):159–72.
12. Tufekci Z, Gowdy JN. Feature extraction using discrete wavelet transform for speech recognition. Proceedings of the IEEE Southeastcon 2000; Nashville, TN. 2000. p. 116–23.
13. Gupta M, Gilbert A. Robust speech recognition using wavelet coefficient features. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01); 2001. p. 445–8.
14. Nhat VDM, Lee S. PCA-based human auditory filter bank for speech recognition. 2004 International Conference on Signal Processing and Communications (SPCOM '04); Banglore, India. 2004 Dec 11-14. p. 393–7.
15. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing. 1980; 28(4):357–66.
16. Rabiner L, Juang BH. Fundamentals of speech recognition. Englewood Cliffs, New Jersey: Prentice Hall. 1993.
17. Rosell M. An introduction to front-end processing and acoustic features for automatic speech recognition. Lecture Notes of School of Computer Science and Communication. Sweden: KTH; 2006.
18. Yusnita MA, Paulraj MP, Yaacob S, Shahriman AB. Classification of speaker accent using hybrid DWT-LPC features and K-nearest neighbors in ethnically diverse Malaysian English. 2012 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE); Kota Kinabalu. 2012 Dec 3-4. p. 179–84.
19. Fahlman SE. An empirical study of learning speed in back-propagation networks. Carnegie Mellon University CMU-CS-88-162. 1988 Sep 1.
20. Looney CG. Pattern recognition using neural network: theory and algorithm for engineers and scientists. New York: Oxford University Press; 1997.
21. Yusnita MA, Paulraj MP, Yaacob S, Yusuf R, Shahriman AB. Analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English. International Journal of Automotive and Mechanical Engineering (IJAME). 2013; 7(1):1053–73.
22. Yusnita MA, Paulraj MP, Yaacob S, Fadzilah MN, Shahriman AB. Acoustic analysis of formants across genders and ethnical accents in Malaysian English using ANOVA. Procedia Engineering. 2013; 64:385–94.
23. Brock IC. Statistical methods of data analysis - WS 00/01. 2003.
24. Pallant J. SPSS survival manual: a step by step guide to data analysis using SPSS for Windows (Version 12). Crows Nest NSW Australia: Allan & Unwin. 2005.
25. Yusnita MA, Paulraj PA, Yaacob S, Shahriman AB. Feature space reduction in ethnically diverse Malaysian English accents classification. 2013 7th International Conference on Intelligent Systems and Control (ISCO); Coimbatore, Tamilnadu. 2013 Jan 4–5. p. 72–8.