International Journal of u- and e- Service, Science and Technology
Vol.9, No. 1 (2016), pp.221-234
http://dx.doi.org/10.14257/ijunesst.2016.9.1.24

Development of Malay Word Pronunciation Application using Vowel Recognition

M.Y. Shahrul Azmi

Technopreneurship Incubation Centre
Universiti Utara Malaysia
shahrulazmi@uum.edu.my

Abstract

In Malaysia, many researchers focus on developing speaker independent systems for training or articulation therapy or to assist language learners to learn about Malay Language or Bahasa Malaysia. Accuracy, noise robustness and processing time are concerns when developing speech therapy systems. In this study, a Malay word pronunciation test application was developed using the first 3 format and fundamental frequencies in an effort to improve pronunciation in Malay. This application was developed using Matlab and uses a vowel recognition algorithm classified using MLP classification technique. The application was developed and tested on UUM undergraduate students. For vowel classification, when fundamental frequency was added, 3-format feature vowel classification rate increased by 1.55% for male gender and 1.48% for female. When combined both genders, a more significant improvement of 1.71% was seen. The developed pronunciation application test results showed that the pronunciation application can assist in testing and improving their Malay word pronunciation. It was also observed that, vowel /i/, /e/, /o/ and /u/ are often mispronounced due to pronunciation habits.

Keywords: *Vowel Recognition, Malay words, Pronunciation Test*

1. Introduction

This article is an extended version of the proceeding article submitted to [1]. Computer based speech therapy and assessment is still new in Malaysia, especially using Malay language or *Bahasa Malaysia*. In this Malaysia language, Malay words are pronounced using a combination of consonant and vowel sounds such as "KATIL" represented by syllable "KA" and "TIL". There are several studies that shows that a speech therapy system that uses vowel phonemes can be used to improve Malay word pronunciation. A hearing impaired person can also be trained to speak *Bahasa Malaysia* properly with a good degree of intelligibility in pronouncing given words. A high degree of standard Malay vowel recognition capability is needed in all of these systems.

Although there are many studies on Malay phoneme recognition, there is still significant work needs to be done. Most of these studies use multiple frame analysis, which is a common method employed by most researchers in the area of Speech Recognition. Accuracy, noise robustness and processing time are still concerns when developing speech therapy systems, especially using *Bahasa Malaysia*. The accuracy aspect involves factors such as age and gender. The size of the vocal tract of different gender and age varies which causes their voice to have different fundamental frequencies. This motivates this study to have an objective of developing a Malay word pronunciation test application in an effort to improve Malay word pronunciation.

ISSN: 2005-4246 IJUNESST

2. Malay Speech Therapy Systems

In Malaysia, Universiti Kebangsaan Malaysia (UKM) has two computer-based speech therapy systems situated in the Clinic of Audiology and Speech Sciences. They are the Kay Elemtrics VisiPitch and IBM Speech Viewer [2]. These systems are used for voice therapy, but not used for training or articulation therapy. Furthermore, these systems use English speech therapy. There are other applications like *OLTK (Optical Logo-Therapy Kit)* [3] and *VATA (Vowel Articulation Training Aid)* [4]. These systems have limitations, and not robust enough to handle real-time identification of vowels. In 2007, Tan et. al [5] developed a Malay Speech Therapy Assistance Tool (MSTAT) which is used to assists therapists in diagnosing children for language disorder and train the children suffering from stuttering problem. It uses speech technologies consisting of speech recognition, Malay Talking Head and Malay text-to-speech system.

A Computer-based Malay Language Articulation Diagnostic System was developed using Hidden Markov Model (HMM) and Mel-Frequency Cepstral Coefficients (MFCCs) [6]. It was developed using a database of Malay words. In 2012, Tan et.al developed a Malay dialect translation and synthesis system, but still at a preliminary stage [7]. The speech synthesis system used here is an HMM speech synthesis system (HTS Speech Synthesis System) at a sampling rate of 22 kHz. The results were promising, but the system does not test on pronunciation. A research was done in 2014 with the objective of developing an ASR system for Malay speaking children [8]. The speech corpus comprises of six children uttering a total of 390 sentences. The parameter training is performed using the HTK toolkit by utilizing an HMM speech acoustic model of Malay speaking children. The system can accurately recognize of up to 76% of test words. Yusof et.al did a study about speech intelligibility of deaf children in Malaysia using a Malay Speech Intelligibility Test (MSIT) system [9]. Researchers from Universiti Malaysia Sarawak did a study on syllabification algorithm based on Malay syllable structure [10]. It was used to build the Iban and Bidayuh syllable list and speech corpus. The accuracy, using Categorical Estimation (CE), gave a mean score of 3.07 out of 5.

3. Vowel Recognition Process

The Malay Word Pronunciation Application engine is based on vowel recognition process. It starts with the data acquisition, next are filtering and pre-processing, frame selection, speech signal modelling, feature extraction and finally vowel recognition processes.

3.1. Data Acquisition

Data collection process was done and taken from 40 Malay students from Universiti Utara Malaysia. The words “ka”, “ke”, “ki”, “ko”, “ku” and “kə” were recorded from speakers representing six vowels of /a/, /e/, /i/, /o/, /u/ and /ə/. In this study, 8000 Hz sampling frequency was used to sample the vowels and up to 10 recordings were taken per speaker depending on situation convenience.

3.2. Feature Extraction

First three formant features using linear predictive coding (LPC) was one of the feature extractions that widely used to classify vowels. Formant values can vary widely from person to person, and all voiced phonemes have formants even if they are not as easy to recognize. There are standard algorithms to compute the first three formants as explained below. Formant feature extraction typically has several steps:

Step 1: Get a section of vowel

```
sq=0;
ii=num2str(i);
openwave=files(i).name;
[y,fs,bits]=wavread(openwave);
fs=8000;
[x]=size(y);
starter=round(x/5);
stopper=round(4*x/5);
sig=y(starter:stopper);
```

From the coding above, first a section of vowels have to be extracted. The frame size chosen was 60% waveform length with the centre frame located at the centre of the waveform.

Step 2: Computation of formant candidates for every frame, and
Step 3: Determination of the formant track, generally using continuity constraints.

For steps 2 and 3 were explained here with the Matlab coding. One way of obtaining formant candidates at a frame level is to compute the roots of a p^{th} order LPC polynomial. There are standard algorithms to compute the complex roots of a polynomial with real coefficients. Each complex root z_i can be represented as $z_i = \exp(-pb_i + j2pf_i)$ where fi and b_i are the formant frequency and bandwidth respectively of the *ith* root. Real roots are discarded and complex roots are sorted by increasing *f*, discarding negative values. The remaining pairs (f_i,b_i) are the formant candidates. In the experiments, the value for p used is 10. LPC coefficients are computed from 250-millisecond Hamming windows, using the autocorrelation method. Here, the first four formants were calculated and only three formants are used.

```
t=(0:length(x)-1)/fs;
ncoeff=2+fs/1000;
a=lpc(sig,ncoeff);
r=roots(a);
r=r(imag(r)>0.01);
clear ffreq
ffreq=sort(atan2(imag(r),real(r))*fs/(2*pi));
if length(ffreq)<4
  ffreq(4)=0;
elseif length(ffreq)<3
  ffreq(3)=0;
end
w1=[w1;ffreq(1) ffreq(2) ffreq(3) ffreq(4)1 1 1 0 0];
end
```

There are slight differences between male and female speaker. The formant frequencies of female are higher than male speaker. The results are saved in excel file, as shown in Appendix B1 and B2.

3.3. Fundamental Frequency (F_o)

In this section, fundamental frequency (f_o) is obtained. This technique developed for an adaptive feature extraction that is more accurate. The algorithm used to obtain the fundamental frequency is shown below.

```
sq=0;
ii=num2str(i);
openwave=files(i).name;
[x,fs]=wavread(openwave);
ms20=fs/50;              % minimum speech Fx at 50Hz
r=xcorr(x,ms20,'coeff');
ms2=fs/500;              % maximum speech Fx at 500Hz
ms20=fs/50;              % minimum speech Fx at 50Hz
r=r(ms20+1:2*ms20+1);
[rmax,tx]=max(r(ms2:ms20));
Fxval=fs/(ms2+tx-1);
w1=[w1; Fxval 1 1 0 0 1];
```

Fundamental frequencies of the female speaker are higher than male speaker as shown in Fig.1. The fundamental frequency from this experiment can be compared with the theory. From the experiments, ranges of the fundamental frequency obtained are higher than the theory.

Table 1. Fundamental Frequency Ranges and Average

	Experiment		Theory	
Gender	**Male**	**Female**	**Male**	**Female**
Range	**93 - 190**	**160 – 296**	**85 – 180**	**165 – 255**
Average	**136**	**231**	**120**	**210**

3.4. Multi-layer Perceptron (MLP)

In this study, an artificial neural network will be used to classify the feature. This classifier was chosen based on their popularities in speech recognition researches. The features in this study are classified using Multi-layer Perceptron (MLP) tool built-in. Multi-layer Perceptron is a feedforward neural network with one or more hidden layers.

The input and output attributes used in this study was data set of male, female and combine that must be normalised. The input variables consist of three for first experiment, and four for second experiment and one output variable. The input variables are formant 1, formant 2, formant 3 and fundamental frequency. The output variable corresponds to five Malay vowels. Training and testing data sets will be chosen by doing 10-fold of the data to make it random data. Data are partitioned into training, 70% and test, 30%. The neural networks are trained using training data. Test data are used to confirm the performance of the network. In this study, an accuracy and root mean square (RMS) error will be measured. The same data will be used to repeat the experiments that have been 10-fold cross validation. The best model with highest accuracy and have the lowest error rates that will be selected. The process of the MLP will be explained clearly in the rest of this chapter. Fig.1 shows the block diagram of multi-layer perceptron process.

Cross validation steps for Neural Network using MLP for 10 folds

Step 1: Normalize data from 0.1 – 0.9

The data will be converted into appropriate forms for mining. For example attribute data maybe normalized to fall between small ranges such as 0.1 to 0.9. Normalize data can be measured using min-max normalization formula as follow:

$$v' = \frac{v - min}{max - min}(new_{max} - new_{min}) + new_{min} \quad (1)$$

Step 2: Randomize data and save file in .csv delimited

Step 3: Built MLP model

A model is built describing a predetermined set of data classes or concepts. The model is used for classification.

Step 4: Insert the data into input development data interface.

Step 5: Split data into 2 sets

Set 1 (70%) – Training Set
Set 2 (30%) – Testing Set

Step 6: Adjust the MLP network depend on the requirement.

The network consists of three layers: input, hidden and output layer. Each of the three activation functions: sigmoid, tanh and linear functions is employed in developing the models. Since in this study, continuous data is used, sigmoid function is more suitable to use. For each activation function employed, the number of hidden nodes are changed subsequently starting with three, four, five and six nodes. The hidden nodes can be calculated using equation (2).

$$no.of\ hidden\ nodes = \sqrt{input\ x\ output} \quad (2)$$

The conjugate gradient and steepest descent are being used for the learning algorithm.

Step 7: Train each classifier model using same training set.

Step 8: Test model with testing set. Compute Classification rate.

The percentage of test set samples that are correctly classified. Accuracy of the model based on training sets and test sets. Training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to overfit the data (i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set overall). Therefore, a test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set. They are independent of the training tuples, meaning that they are not used to construct the classifier.

Step 9: Repeat step 1-7 for next run until 10-fold cross validation.

Step 10: Compute average classification rate. End

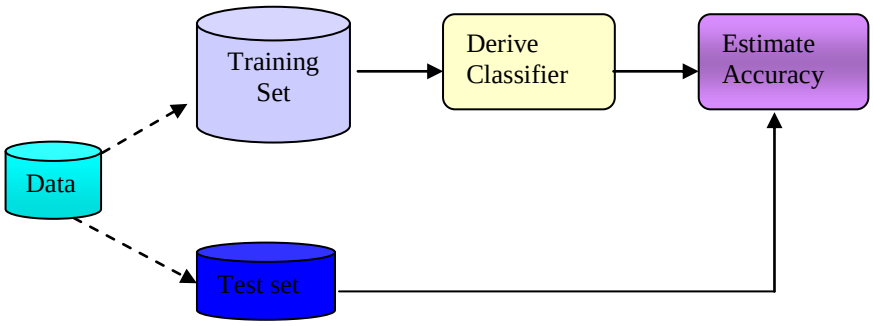

Figure 1. Overall Process of MLP Building

4. Vowel Classification Results

The features from proposed method are classified using Neural Network which is Multi-layer Perceptron (MLP). The reason for choosing this classifier was because it was among the most widely used classifiers by speech recognition researchers. This classification results were divided into three parts, which are male, female and combine of male and female. The experiments were repeated by adding the fundamental frequency, (f_o) as another input variable. This classification results were compared between the first experiment, without fundamental frequency and the second experiment, with fundamental frequency.

Classifications results were based on cross validation techniques where the database is randomly divided into training and testing sets in the ratio of 7:3. This was done for each cross validation run where 70% training set will be used in training the classifier model and the other 30% of the data was treated as unseen testing inputs. A total of 10-fold cross validations tests were done and their averaged classification results were computed averaged for classifier.

4.1. Result Without F_o

In this section, the first experiment was done with three inputs of formants 1, 2 and 3 without the fundamental frequency. Table 2, Fig.2 and Fig.3 shows the result of classification rate without the fundamental frequency information.

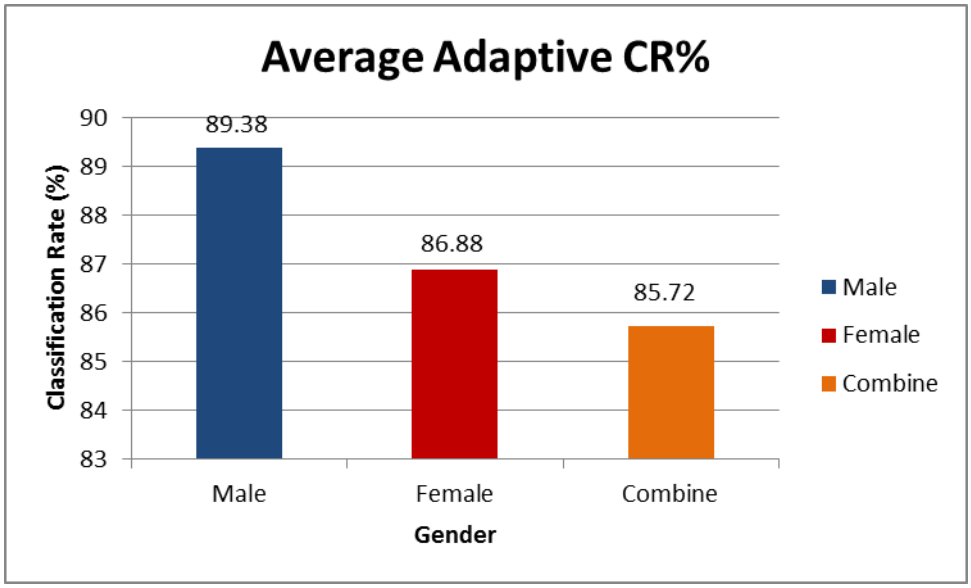

Figure 2. Comparison Average Classification Rate without f_o

Table 2. Average of NN/MLP Formant without f_o

Gender	a	e	i	o	u	CR%
Male	99.18	88.66	100	65.59	95.84	89.38
Female	98.47	78.59	98.17	71.14	88.99	86.88
Combine	99.43	85.81	99.37	57.82	79.36	85.72

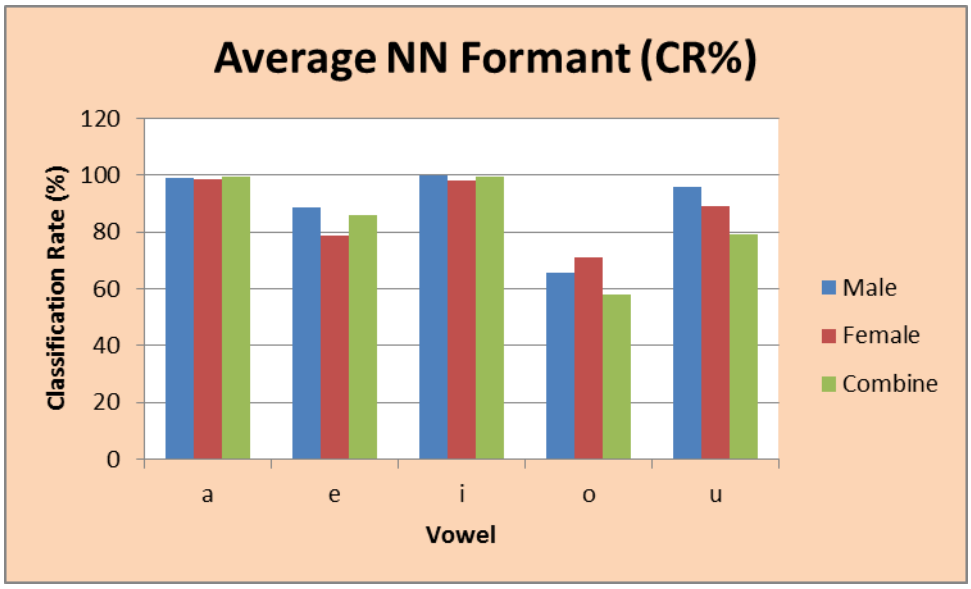

Figure 3. Comparison of Average NN/MLP Formant Classification Rate without f_o

Fig.3 shows the classification rate between male, female and combine of five Standard Malay (SM) vowels which are /a/, /e/, /i/, /o/ and /u/. Male gave the best accuracy, 89.38% followed by female of 86.88% and combine with 85.72%. Classification by individual genders is more accurate than the combine of both genders.

Table 3. Vowel Classification result without fo

	Best Recognition Performance for Vowel		Worst Recognition Performance for Vowel	
Gender	**Vowel**	**CR%**	**Vowel**	**CR%**
Male	**/i/**	**100**	**/o/**	**65.59**
Female	**/a/**	**98.47**	**/o/**	**71.14**
Combine	**/a/**	**99.43**	**/o/**	**57.82**

Table 3 shows among all the vowel, vowel /i/ was the best classified with 100% accuracy for male gender. Meanwhile vowel /a/ also gave best classified for female and combine with 98.47% and 99.43% respectively. Vowel /o/ gave the worst classification rate for all the gender in which combine gave the worst result of 57.82%.

4.2. Result With Fo

The second experiment was done with four inputs by adding the fundamental frequency and one output.

Weight initialization, Gaussian distribution

Data allocation (70/0/30)

Table 4. Best Fold CR% with f_o

Gender	Learning Algorithms	Accuracy (%)
Male	Conjugate Gradient	90.93
Female	Conjugate Gradient	88.36
Combination	Conjugate Gradient	87.43

Table 4 above shows the result for best accuracy of 10-fold dataset. A total of 10-fold cross validations tests were done and their averaged classification results were computed averaged for classifier. Best fold of male with obtained an overall accuracy of 90.93% followed by female as shown in with an overall accuracy of 88.36%. Combination between male and female gave 87.43% which was 3.50% and 0.93% lower than the male and female results respectively. Fig.4 shows the comparison between these genders in graph.

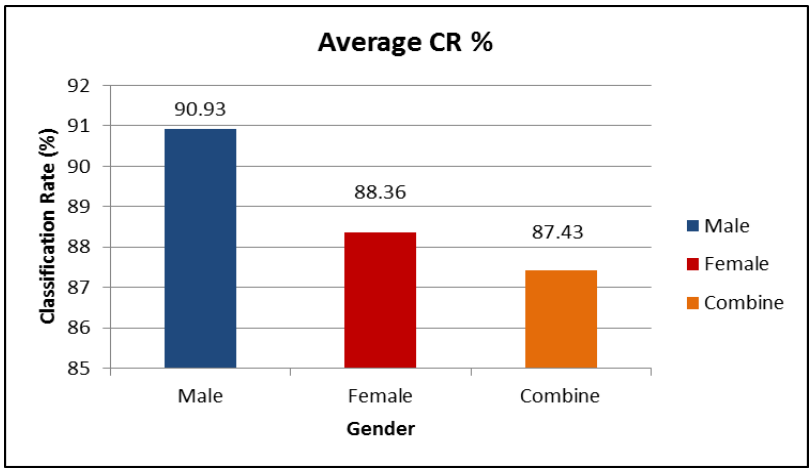

Figure 4. Comparison Average Classification Rate with f_o

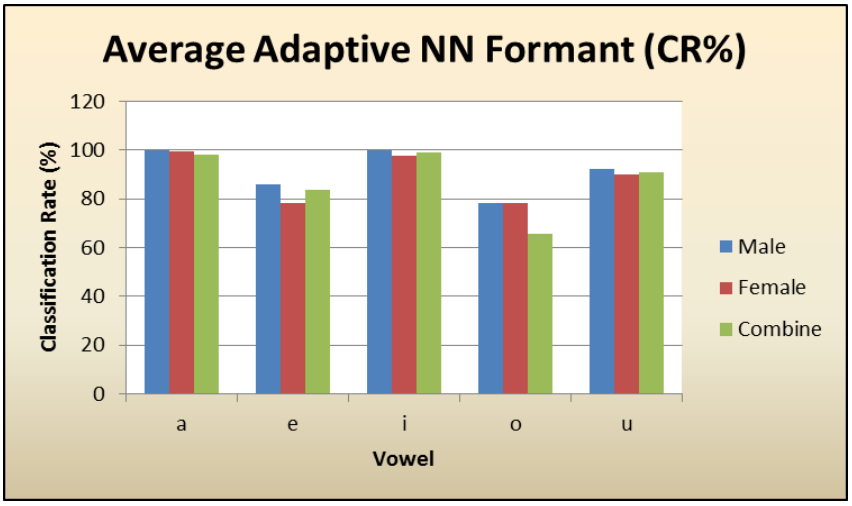

Figure 5. Average of Adaptive NN/MLP Formant with f_o

Table 5. Average of Adaptive NN/MLP Formant with f_o

	Vowel					
Gender	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**CR%**
Male	100	85.82	100	78.50	92.18	90.93
Female	99.58	78.26	97.54	78.37	90.02	88.36
Combine	98.39	83.78	99.05	65.61	91.07	87.43

From the result obtained, the overall results are presented in Fig.5 and Table 5. Fig.5 shows the classification rate between male, female and combine of five Standard Malay (SM) vowels which are /a/, /e/, /i/, /o/ and /u/. Male gave the best accuracy of 90.93% followed by female of 88.36% and combine with 87.43%. Classification by gender is more accurate than the combine of both genders.

Table 6. Vowel Classification result with fo

	Best Recognition Performance for Vowel		Worst Recognition Performance for Vowel	
Gender	**Vowel**	**CR%**	**Vowel**	**CR%**
Male	/a/, /i/	100	/o/	78.50
Female	/a/	99.58	/e/	78.26
Combine	/i/	99.05	/o/	65.61

Table 6 shows among all the vowel, vowel /a/ and /i/ was the best classified with 100% accuracy for male gender. Vowel /o/ gave the worst classification rate with 65.61% for combine and 78.50% for male. Meanwhile for female, vowel /e/ gave the worst result of 78.26%.

4.3. Comparison Results With And Without Fo

In this section, the results obtained are compiled and compared to see the accuracy by genders and by using fundamental frequencies as the 4th input. Table 7 shows that there is significant increase in classification rate when fundamental frequencies are being used for vowel recognition for all categories of genders.

Table 7. Comparison Average of MLP Formant (f0)

Gender	Without f0	With fo
Male	89.38	90.93
Female	86.88	88.36
Combine	85.72	87.43

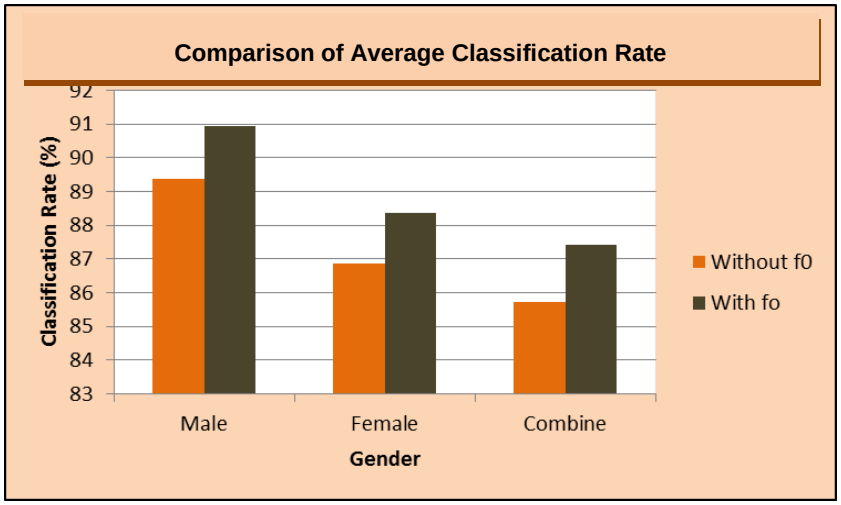

Figure 6. Comparison of Average Classification Rate by Different Genders using Fundamental Frequencies

Table 8. Overall Vowel Classification Improvement

Gender	Improvement of without fo over with fo
Male	1.55 %
Female	1.48 %
Combine	1.71 %

Based on Fig.6 and Table 7, adaptive feature extraction method with fundamental frequency (f_o) performs better in overall vowel classification than feature extraction method. Vowel classification improvement was shown in Table 8. For male gender was improved of 1.55% and female of 1.48%. Biggest improvement was seen for combine where the improvement was 1.71%.

5. Pronunciation Test Interface

The test interface was developed using MATLAB where the user may select either word based interface which is shown in Fig. 7 and avatar based interface as shown in Fig. 8.

5.1. Word Based Interface

For interface 1, the Malay words chosen were 2-syllable words of “katil”, “roti”, “lara”, “buncis”, “potong” and “betik” can be selected representing the six vowels of /a/, /e/, /i/, /o/, /u/ and /ə/. For example, the word “KATIL” means bed in English where the proper pronunciation requires the vowels /a/ and /i/ to be clearly pronounced. The algorithm calculates the accuracy of pronunciation based on these two vowels and lower accuracy will be given if the pronunciation differs. The lowest accuracy was recorded based on the first uttered word in which the speakers are untrained. Then the speakers are trained on how to pronounce the words properly. Next they will try uttering the words again 4 times and the best accuracy result is then taken.

5.2. Virtual Avatar Interface

University Virtual Interview Application (UVIA) is a virtual interview application that uses a virtual avatar that acts like an agent to interact with users. In 2013, a working prototype was developed by Shahrul Azmi from Universiti Utara Malaysia (UUM) and used to prepare students for job interviews. A study done by Raudhoh in 2014 shows that UVIA is able to improve the self confidence level of its users to face real job interviews. Currently, Dr. Shahrul is working on adding more functions such as stress level monitoring and automatic scoring. UVIA has been improved to pronunciation accuracy of the users. The avatar will assist the user to pronounce the given words properly by interacting and encouraging the users.

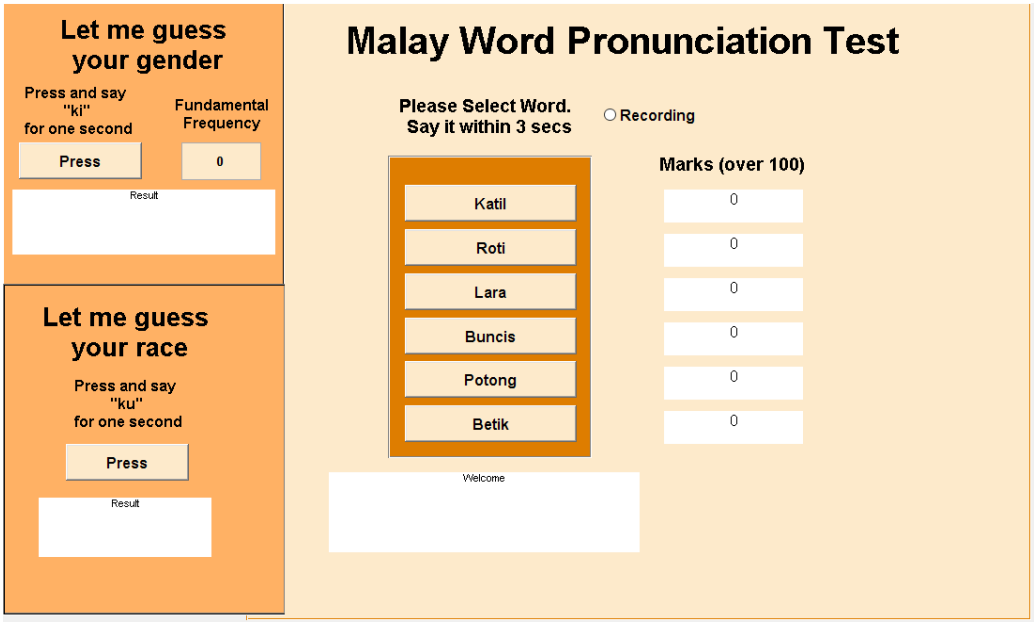

Figure 7. MATLAB Screenshot of Testing Interface

Figure 8. Screenshot of UVIA Testing Interface

5.3. Pronunciation Application Test Results

The application is tested on another 40 Malay students from Universiti Utara Malaysia. The results of the pronunciation test are shown in Table 10. The average lowest accuracy is the average accuracy of the first trial of each speaker pronunciation the given words.

Table 10. Pronunciation Test Results

	Accuracy		
Words	**Average Lowest**	**Average Highest**	**Average Tries to reach average highest**
Katil	56%	86%	4
Roti	65%	92%	4
Lara	71%	92%	2
Buncis	52%	84%	4
Potong	70%	91%	2
Betik	58%	88%	3

The first accuracy results were taken based on the first attempt. Then the speaker will be trained on how to pronounce properly and tested again in the next 5 more attempts. Based on the results obtained, for the word "katil" or "Ka" and "til" meaning bed, the average lowest accuracy were 56% accuracy for the first trial and 86% for after improvements were done in subsequent pronunciations. On the average, 4 times are needed to obtain the highest average accuracy. For this word, the syllable "til" was supposed to be pronounced like the word "till" in English, but often pronounced inaccurately as "tail" in English. After the correction in pronunciation, the speakers are able to improve their pronunciation and obtained an improved score of 86%. For the word Roti, the speakers often mispronounced "Ro" as "Rue" instead of "Roo". For the word Lara, not much problem in pronouncing it correctly but the lower initial accuracy was due to the speaker spoke the word softly and sounded less confident. For the word "Buncis", the syllable "cis" was often mispronounced as "cess" instead of "cheese" in English. Not much problem seen when pronouncing the word "potong". For the word "Betik", the syllable "tik" was often mispronounced as "take" instead of "tick" in English.

6. Conclusions

This paper presents a new Malay Word Pronunciation Application. This application was developed using Matlab and uses a speech recognition algorithm based on Formant features and fundamental frequency (F_0) and MLP classification technique. The application was developed and tested on UUM undergraduate students. For vowel classification, male gender showed an improvement of 1.55% and female of 1.48%. Furthermore, the best improvement was seen for combine gender data where the improvement was 1.71%.

For the pronunciation testing, the actual result obtained was significantly lower than 75% for the first try. This is because of the lack of emphasizing on proper pronunciation on the given words due to daily mispronunciation which is often happening around them. The clarity of the pronounced words may lower the accuracy measured by the application

which is mostly due to nervousness and lack of confidence. After training, the speakers are able to pronounce accurately. Overall, this application is able to help individuals to learn to pronounce Malay words properly and clearly.

Acknowledgments

The author would like to thank Universiti Utara Malaysia for providing the grant to do this study. The author would also like to appreciate all the colleagues and students who have supported in giving good insights and suggestions.

References

[1] M.Y. Shahrul Azmi, "Malay Word Pronunciation Application for Pre-School Children using Vowel Recognition", The 8th International Conference on u- and e- Service, Science and Technology, **(2015)**; Jeju Island, South Korea.
[2] H. Ting, J. Yunus, S. Vandort and L. Wong, "Computer-based Malay articulation training for Malay plosives at isolated, syllable and word level", Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, **(2003)**.
[3] A. Hatzis, "Optical Logo-Therapy (OLT): Computer-based audio-visual feedback using interactive visual displays for speech training, **(1999)**.
[4] A. Zimmer, "VATA: An improved personal computer-based vowel articulation training aid", Old Dominion University, **(2002)**.
[5] T. S. Tan, A. K. Ariff, C. M. Ting and S. H. Salleh, "Application of Malay speech technology in Malay speech therapy assistance tools", **(2007)**.
[6] M. N. Mazenan and Tan, T. S., "Malay Alveolar Vocabulary Design for Malay Speech Therapy System". Recent Advances in Electrical Engineering Series, vol. 11, **(2013)**.
[7] T. P. Tan, S. S. Goh and Y. M. Khaw, "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System", International Conference on IEEE Asian Language Processing (IALP), November, **(2012)**, pp. 109-112.
[8] F. D. Rahman, N. Mohamed, M. B. Mustafa and S. S. Salim, "Automatic speech recognition system for Malay speaking children", IEEE Third ICT International Student Project Conference (ICT-ISPC), **(2014)**, pp. 79-82.
[9] Z. M. Yusof, R. Hussain and M. Ahmed, "Malay Speech Intelligibility Test (MSIT) for Deaf Malaysian Children", International Journal of Integrated Engineering, vol. 5, no. 3, **(2014)**
[10] S. F. Juan, V. Edwin, C. Y. Cheong, L. J. Choi and A. W. Yeo, "Adopting Malay Syllable Structure for Syllable Based Speech Synthesizer for Iban and Bidayuh Languages", IEEE International Conference on Asian Language Processing (IALP), **(2011)**, pp. 216-219.
[11] Ø. Birkenes, "A Framework for Speech Recognition using Logistic Regression", Unpublished PhD thesis, Norwegian University of Science and Technology, **(2007)**.
[12] J. Hillenbrand and T. Nearey, "Identification of resynthesized/hVd/utterances: Effects of formant contour", The Journal of the Acoustical Society of America, vol. 105, vol. 6, **(1999)**, pp. 3509-3523.
[13] So, Y., & Cary, N., "A tutorial on logistic regression", SAS White Papers, **(1995)**.

Author

M.Y. Shahrul Azmi, He obtained his bachelor degree in Electrical Engineering from Missouri University of Science and Technology, USA (previously University of Missouri-Rolla) in 1994. He later pursue his MSc. In Information Technologi from Universiti Sains Malaysia and obtained his PhD in Mechatronics Engineering from Universiti Malaysia Perlis in 2010. He served as an engineer at Robert Bosch and Seagate Company from 1995 until 1999. Currently, he is a faculty member at School of Computing in Universiti Utara Malaysia. His research interests include Speech Recognition, IT Entrepreneurship, Artificial Intelligence and Manufacturing.