Journal of ICT, 7, pp: 27-40

CLASSIFICATION OF MALAYSIAN VOWELS USING FORMANT BASED FEATURES

S. A. Mohd Yusof[1], Paulraj M[2], S. Yaacob[3]

School of Mechatronics Engineering, Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia

shahrulazmi@uum.edu.my[1]
paul@unimap.edu.my[2]
s.yaacob@unimap.edu.my[3]

ABSTRACT

Automatic speech recognition (ASR) has made great strides with the development of digital signal processing hardware and software, especially using English as the language of choice. Despite of all these advances, machines cannot match the performance of their human counterparts in terms of accuracy and speed, especially in case of speaker independent speech recognition. In this paper, a new feature based on formant is presented and evaluated on Malaysian spoken vowels. These features were classified and used to identify vowels recorded from 80 Malaysian speakers. A backpropagation neural network (BPNN) model was developed to classify the vowels. Six formant features were evaluated, which were the first three formant frequencies and the distances between each of them. Results, showed that overall vowel classification rate of these three formant combinations are comparatively the same but differs in terms of individual vowel classification.

Keywords: Speech recognition, Vowel detection, Formant.

INTRODUCTION

In the human language, a **phoneme** is the smallest structural unit that distinguishes meaning. Normally, languages like English commonly combine phonemes to form words or sentences. In Bahasa Malaysia, children are normally taught to spell the words using a combination of consonants and vowels. In terms of vowel phoneme, Bahasa Malaysia or Malay Language has only six vowel phonemes (/a/, /e/, /i/, /o/, /u/, /e'/) (Maris, 1979), whereas typical American English has 20 vowel phonemes.

English word pronunciation depends on a sequence of phonemes. Audio signals are broken up into acoustic components and translated into phonemes. These phoneme sequences are then compared with words from an English database that can be made up of thousands of words. For Malay words, the approach is different. It is comprised of Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) combinations. It is possible that a Malay word can be spelled out by a computer similar to a human being. We believe that a computer can be taught to spell like a child and is able to translate CV or CVC combinations into proper and understandable words. Our motivation for this work is to see whether a computer system can interact with Malaysians using the Malay Language based on CV or CVC words. Among the other applications that can be developed from this research is a system that can assist individuals such as children and those who are new to Bahasa Malaysia to learn to speak the language with proper pronunciation. There are two objectives of this paper. The first is to study the differences or distance between the first three formant frequencies of the common vowels /a/, /e/, /i/, /o/ and /u/. The second is to classify vowels using a non-linear classifier of Backpropagation Neural Network Model (BPNN) and evaluate the performance of individual vowels.

Research Background

Vowels are voiced sounds produced by passing air through the mouth without any major obstruction in the vocal tract (Rogers, 2000; Stevens, 1998). The peaks of these acoustic spectra are referred to as formants which are the resonant frequencies of any acoustic system. Its acoustic energy concentrates around a particular frequency in the speech wave. In practice, only the lowest three or four formants are of interest (Kent & Read, 2002).

Speech recognition is considered new to most Malaysian researchers where most of their publications had surfaced in the last 10 years. Among the active Malaysian Universities in researching *Bahasa Malaysia* (BM) or Malay Speech Recognition are Universiti Teknologi Malaysia (UTM), Universiti Kebangsaan Malaysia (UKM), Universiti Putra Malaysia (UPM), Universiti Sains Malaysia (USM), and Multimedia University (MMU). For example, UTM did research into Malay plosives sounds (Ting, Yunus, & Salleh; 2002) and Malay numbers (Salam, Mohamad & Salleh, 2001; Sudirman, Salleh, & Salleh, 2006). UTM also did a study on Malay vowels based on cepstral coefficients (Al-Hadaad, Samad, Hussain, Ishak, & Noor, 2009) and fusion of Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) (Ting, & Yunus, 2004).

There was much research in speech recognition done based on vowel recognition. Qin (Yan & Vaseghi, 2003) studied formant features of formant frequency, bandwidth, and intensity to classify accent conversion between British, Americans, and Australian speakers. Carlson (Carlson & Glass, 2006) also analysed Formant Amplitude for vowel classification while Vuckovic (Vuckovic & Stankoric, 2001) researched on automatic vowel classification based on 2-dimensional formant Euclidean distance. There is even an on-going research done in Universiti Malaysia Perlis (UniMAP) to compare vowels obtain from the three main Malaysian races of Malay, Chinese, and Indian. This paper presents the classification of vowels using another formant feature which is formant distance.

In this paper, formant frequencies were obtained with a Linear Predictive Method. In linear predictive (LP) analysis, an all-pole filter with transfer function (1) models the vocal tract transfer function.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (1)$$

where the gain parameter, $\boldsymbol{G}$ *is* a constant and $\boldsymbol{p}$ is the number of poles. $\boldsymbol{S(z)}$ and $\boldsymbol{U(z)}$ are obtained by *Z*-transform from output signal $\boldsymbol{s(n)}$ and input signal $\boldsymbol{u(n)}$ while $\boldsymbol{a_i}$ is the linear prediction coefficients. Spectral envelope can be obtained by means of low-order autoregressive modelling of the audio signal (Hayes, 1996).

METHODOLOGY

This section is broken into four subsections, which are data collection, pre-processing, vocal tract analysis, and formant distance calculation.

Data Collection

Data Collection process was taken from a total of 80 individuals consisting of students and staff from Universiti Malaysia Perlis (UniMAP) and Universiti Utara Malaysia (UUM). The recordings were done using a conventional microphone and a laptop computer with a sampling frequency of 8000Hz. The words "KA, KE, KI, KO, KU" were used to represent the five vowels of /a/, /e/, /i/, /o/ and /u/ because vowels have significantly more energy than consonants. Based on Rabiner & Juang (1993), Hillenbrand, Getty, Clark, and Whealer (1995), Vuckovic and Stankorie, (2001), Huang et al. (2001), the first three formants for vowels are situated within 4 kHz and so are vowel's main

characteristics. For this study, a sampling frequency of 8 kHz was used to sample the vowels. The recordings were done three to four times per speaker. The details of the data collection are listed in Table 1 below.

Table 1: Data Collection Details

Information	1st Data Collection	2nd Data Collection
Sources	40 UniMAP students	20 UUM staff and 20 students
Recorded Utterances	640	445
Sampling Frequency	8000 Hz	8000 Hz
Words Uttered	/ka/, /ke/, /ki/, /ko/, /ku/	/ka/, /ke/, /ki/, /ko/, /ku/

The summary of the entire vowel recognition process is shown in Fig. 1.

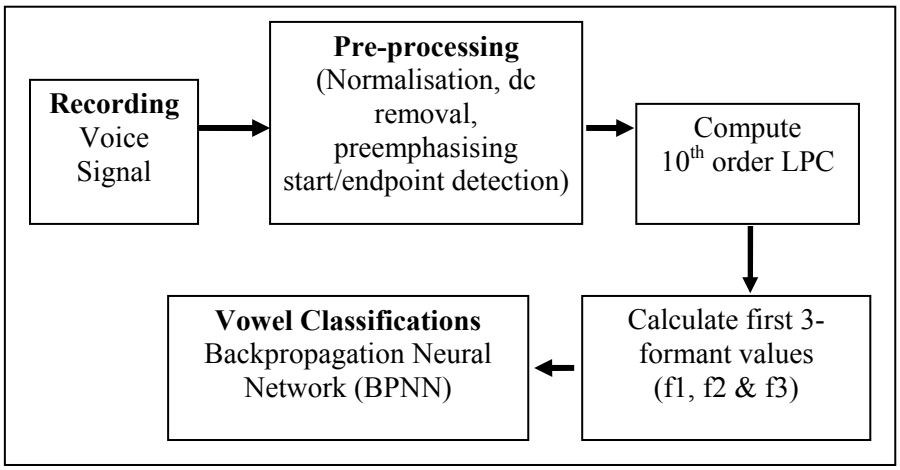


Fig. 1: Vowel Recognition Process

Pre-Processing

The signal was recorded using a laptop and a microphone based on a sampling frequency of 8000Hz. The direct current (dc) portion of the signal which was introduced by the recording equipment was removed and the resultant signal was then normalised. The start and endpoint locations were determined based on waveform energy segmentation method. Two threshold values of signal segment duration and amplitude were used to separate speech signals and noise. The segmentation process is summarised in Fig. 2.

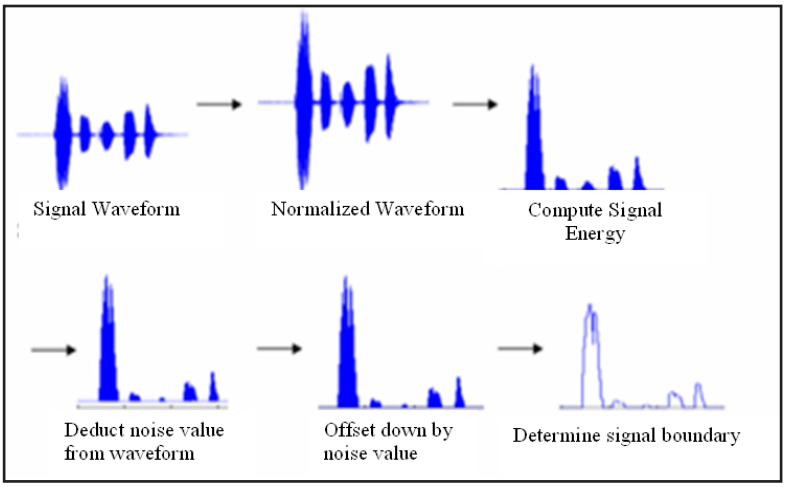
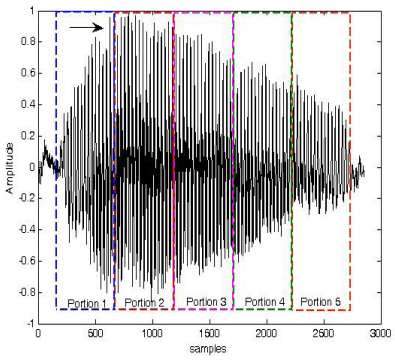


Fig. 2: Vowel Extraction Process

Normally, frame-by-frame analysis is used to analyse the speech signals but in this vowel recognition method, only a single signal frame analysis was used to extract the features. In order to determine the best frame size and location to analyse the waveform, the spectrum was analysed using frame-shifted waveform and frame- expanding waveform methods, as shown in Fig. 3 and 4.



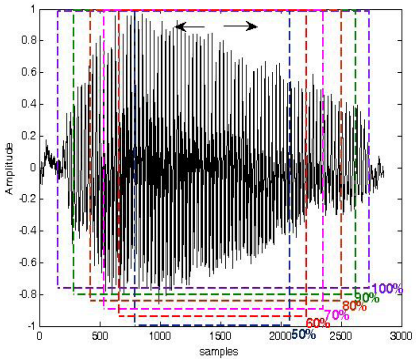


Fig. 3: Frame-shifted Method **Fig. 4: Frame-expanding Method**

The spectrums of the Frame-Shifting analysis showed inconsistent response as the frame moves from left to right by 20% of the total signal duration. On the other hand, the spectrums of the Frame-Expanding analysis showed the same consistent response using different frame size with the centre of the each frame situated at the centre of the waveform. The frame sizes chosen was 70% waveform length with the centre frame located at the centre of the

waveform. When any part of frame is chosen for analysis, the segmentation process may cause some difference between the signal at the beginning and end of the voice segment. This can produce spectral leakages. To reduce the discontinuity, a Hamming window function was applied to bring the signal smoothly to zero at beginning and end points.

The Hamming window is given by the Equation (2).

$$w_H[m] = \begin{cases} 0.54 + 0.46\cos(\pi m / M) & -M \le m \le M \\ 0 & otherwise \end{cases} \quad (2)$$

Next, the signal was pre-emphasised to emphasise the higher frequency component of the signal. Pre-emphasis compensates the effect of the glottal-source and energy radiation from the lips (Kinnunen, 2003). The pre-emphasise filter was implemented by the Equation (3) using a pre-emphasised constant value of 0.95.

$$s'[n] = s[n] - A_c s[n-1] \quad (3)$$

where $s'[n]$ – pre-emphasised signal,
$s[n]$ - original signal, and
A_c – pre-emphasised constant (0.95).

Analysing the Vocal Tract Model

The magnitudes of the 512-point complex frequency response were plotted for each of the vowels. In Fig. 5, all the averaged speakers' spectrum envelope plots are shown for each of the vowels in linear scale, which were modelled using Linear Predictive (LP) Method.

The peaks in the linear-scaled spectrum are more defined than the log-scaled spectrum. It is easily visible how closely the responses for different speakers match up for any of the vowels (see Fig. 5).

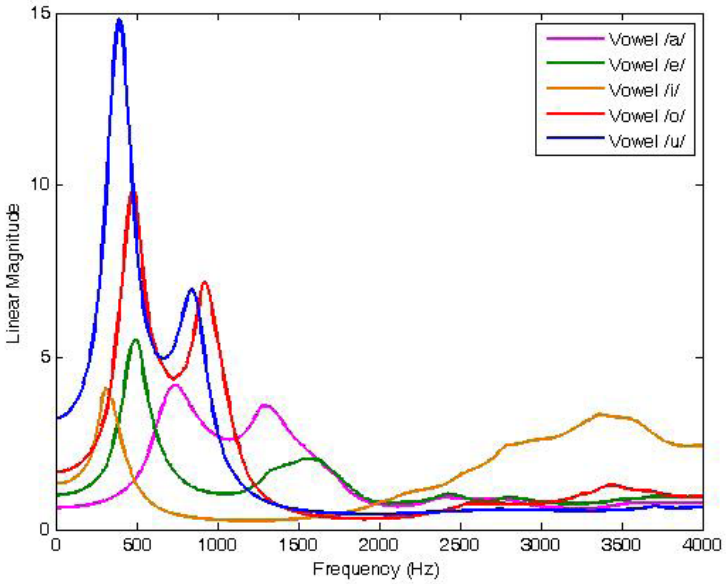


Fig. 5: Linear-scaled Spectrum

Formant Distance Calculation

Formant distance is calculated based on the first three-formant values of f1, f2, and f3, obtained from the LPC approach. Three new values labeled as FD1, FD2, and FD3 were calculated using Equations (4), (5), and (6).

$$FD1=f2-f1 \quad (4)$$

$$FD2=f3-f2 \quad (5)$$

$$FD3=f3-f1 \quad (6)$$

DATA ANALYSIS

Three sets of formant combination were evaluated in terms of classification performance and training time using Backpropagation Neural Network. The Table below shows the combinations of formants in groups called 3F, 4F, and 6F.

Table 2: Formant Groups

Group	Combinations
3F	F1, F2, F3
4F	F1, FD1, FD2, FD3
6F	F1, F2, F3, FD1, FD2, FD3

Classification using Backpropagation Neural Network (BPNN)

In this study, gradient descent with momentum and adaptive learning rate back-propagation (GDX) algorithm was used to classify formant features. GDX is a neural network training function that updates weight and bias values according to gradient descent momentum and adaptive learning rate.

Either one or two hidden layer was used in this study to identify the vowel utterances depending on the method used. The network used are in the form of ***inp_n*** **x** ***hid_n*** **x** ***out_n*** for 1-hidden layered network and ***inp_n*** **x** ***hid_n*** **x** ***hid_n*** **x** ***out_n*** for 2-hidden layered network, where ***inp_n*** is the input neurons, ***hid_n*** is the hidden neurons and ***out_n*** is the output neurons. The vowel /a/, /e/, /i/, /o/ and /u/ are represented by the three bit output neurons valued 001, 010, 011, 100, and 101 respectively. The network was trained using 70% of the data. The weights and biases of the MLP were initialised randomly. The learning rate was set at 0.01 and momentum factor at 0.9. Table 2 and Fig. 11 shows the summary of the results of the classification based on different testing tolerance.

Evaluating Different Network Configuration

Tables 3 to 5 shows the classification rate and training time using different network configuration and settings at testing tolerance of 0.2. The target mean squared error (mse) was 0.02.

Table 3: Classification Rate of 3-formant using Different Network Configurations

Analysis	Network Config.	CR (%)	Train Time (s)
3F	3x10x3	61.82	130
	3x50x3	72.16	185
	3x5x5x3	68.01	209
	3x10x10x3	73.22	230
	3x20x20x3*	**80.66**	**223**
	3x30x30x3	80.01	265
	3x40x40x3	82.16	290

Table 4: Classification Rate of 4-formant Based on Different Network Configurations

Analysis	Network Config.	CR (%)	Train Time (s)
4F	4x10x3	61.62	41
	4x50x3	68.77	105
	4x10x10x3	69.28	64
	4x20x20x3	75.92	97
	4x30x30x3*	**78.22**	**136**
	4x40x40x3	79.33	165

Table 5: Classification Rate of 6-Formant using Different Network Configuration

Analysis	Network Config.	CR (%)	Train Time (s)
5F	6x10x10x3	77.51	250
	6x20x20x3	77.77	272
	6x30x30x3*	**77.37**	**229**
	6x40x40x3	77.29	199

Up to seven different configurations were used. The best configurations obtained based on the classification rate and training time were 3x20x20x3 for 3-Formant, and 3x30x30x3 for both 4-Formant and 6-Formant. Tables 6 to 8 shows the classification rate and training time using mean square error (mse) of 0.01 using best network configuration based on different testing tolerances. A testing tolerance or threshold of 0.2 was selected based on its accuracy and its permissible limit of variation.

Table 6: Classification Rate of 3-formant Based on Best Setting and Different Testing Tolerance

	Testing Tolerance (3x20x20x3 at mse 0.01)									
3F (F1, F2, F3)	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**Time (s)**	**Epoch**	**mse**
Average	77.82	83.37	**86.80**	89.03	90.66	92.16	93.40	1718	93284	0.01
Std. Dev.	1.73	1.50	**1.32**	1.03	0.94	0.83	0.67	198	10814	0.00

Table 7: Classification Rate of 4-formant Based on Best Setting and Different Testing Tolerances

	Testing Tolerance (3x30x30x3 at mse 0.01)									
4F (F1, FD1, FD2, FD3)	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**Time (s)**	**Epoch**	**mse**
Average	76.58	82.74	**86.72**	89.46	91.22	92.69	93.76	802	32112	0.01
Std. Dev.	0.92	0.48	**0.55**	0.47	0.44	0.65	0.49	240	9517	0.00

Table 8. Classification Rate of 6-formant Based on Best Setting and Different Testing Tolerance

	Testing Tolerance (3x30x30x3 at mse 0.01)									
6F (F1, F2, F3, FD1, FD2, FD3)	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**Time (s)**	**Epoch**	**mse**
Average	74.15	81.99	**86.50**	89.67	92.25	94.24	95.67	1696	128810	0.01
Std. Dev.	1.02	0.41	**0.68**	0.58	0.72	0.32	0.41	268	12997	0.00

This GDX trained network gives an average classification rate of 86.80% for 3-Formant, 86.72% for 4-Formant, and 86.50% for 6-Formant combinations, based on 0.2 testing tolerance and mean square error of 0.01 from 10 tries. The Classification performance of all three methods was comparatively the same. Fig. 6 and Table 9 show the classification rate of formant features according to the individual vowels.

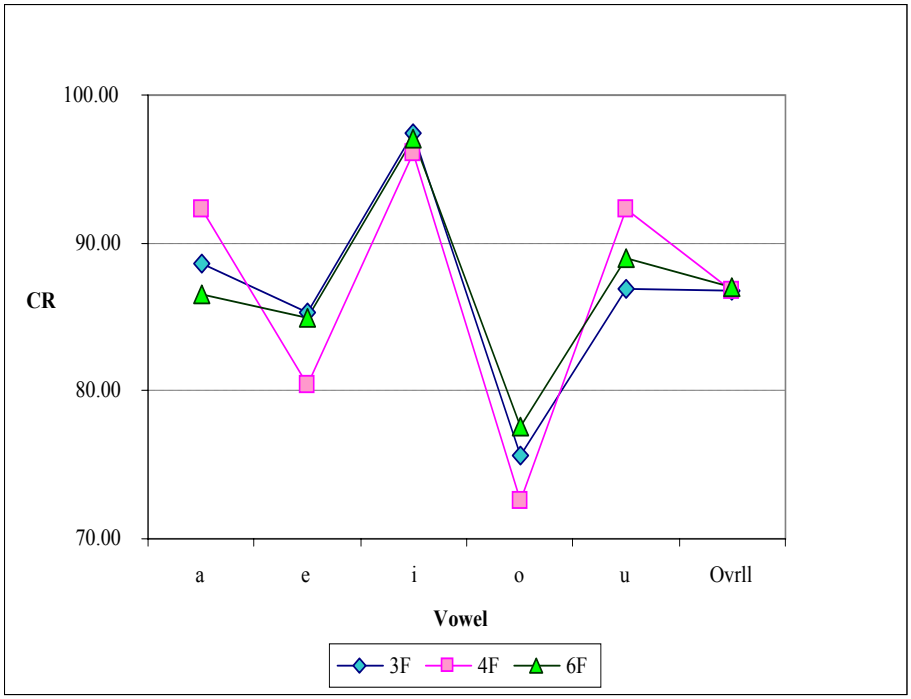


Fig. 6: Vowel Classification at mse =0.01

Table 9: 3-Formant Classification by Vowels

	Vowel					
Method	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Overall**
3F	88.64*	85.28**	97.44**	75.67*	86.92	86.80*
4F	92.32**	80.35	96.08	72.62	92.23**	86.72
6F	86.51	84.93*	97.11*	77.53**	89.03*	87.04**

** - Best Performance * - 2nd Best Performance

Based on Table 9, although the overall performance of BPNN classification was comparatively the same, the classification performance of individual vowels was different. The common 3-Formant group favours both vowel /e/ and /i/, but performs worst for vowel /u/. Group 4F performs best for vowel /a/ and /u/, but did the worst for /e/, /i/, and /o/. Group 6F did best for vowel /o/, but worst for vowel /a/. Overall, above 90% classification can be achieve for vowel /a/, /i/ and /u/ which is good for independent speaker

vowel recognition performance considering the number of features used. On the other hand, classifications of vowel /o/ for all the formant groups were below 78%, which is bad for a speech recognition application.

In terms of training time, 4-Formant method trains the fastest in 802 seconds which is less than half the time taken for both 3-Formant and 6-Formant methods as shown in Fig. 7.

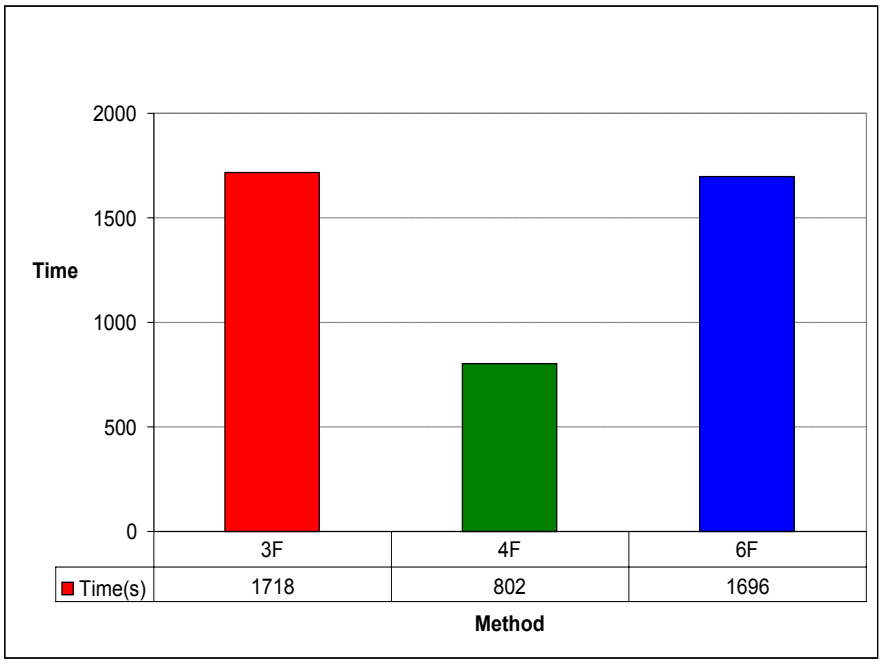


Fig. 7: Classification Accuracy Comparisons

CONCLUSION

In this paper, a useful evaluation of formant features was performed using the first three formants and their distances between each other which fulfilled the first objective. The second objective was also fulfilled where the formant features were classified using a non-linear classifier of BPNN which gives a useful performance result for individual vowel classification. This GDX trained network gives an average classification rate of 86.72% to 87.04% which can be considered comparatively the same. In terms of vowel classification, the best formant combination to detect vowel /a/ and /o/ is F1, FD1, FD2, and FD3. For vowel /e/ and /i/, the common combination of the first three formants performs the best and for vowel /u/, the formant combination of F1,

F2, F3, FD1, FD2, and FD3 performs the best. This study shows promising results of formant distance approach for some of the common vowels of Bahasa Malaysia.

Further study will be done to improve the performance of this feature extraction method in terms of finding better parameters on the vocal tract model to represent the vowels and also in terms of improving the classification network.

REFERENCES

Al-Haddad, S.A.R., Samad, S.A., Hussain, A., Ishak, K.A., & Noor, A.O.A. (2009). Robust speech recognition using fusion techniques and adaptive filtering. *American Journal of Applied Sciences*, *6*(2), 290-295.

Carlson, R., & Glass, J. (2006, October). Vowel classification based on analysis-by-synthesis. *International Conference on Spoken Language Processing*. Canada: Banff, Canada.

Hayes, M.H. (1996). *Statistical digital signal processing and modeling.* John Wiley & Sons.

Hillenbrand J., Getty, L.A., Clark, M.J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of Acoustic Society of America, 97*(5), 3099-3111.

Huang , X. et al. (2001). *Spoken language processing: A guide to theory, algorithm and system development.* Prentice Hall.

Kent, R. D., & Read, C. (2002). *Acoustic analysis of speech.* Singular Thomson Learning.

Kinnunen, T. (2003). *Spectral features for automatic text-independent speaker recognition.* Licentiate's dissertation, University of Joensuu.

Maris, M. Y. (1979). *The Malay sound system*. Malaysia: Fajar Bakti.

Phoneme Chart. *English vowel and consonant sounds*. Retrieved Jan 21, 2009, from http://www.btinternet.com/~ted.power/folkchart.htm.

Yan, Q., & Vaseghi, S. (2003). Analysis, modelling and synthesis of formants of British, American, and Australian accents. *Proceedings on IEEE*

International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003).

Rabiner, L., & Juang, B.H. (1993). *Fundamentals of speech recognition*. Prentice Hall.

Rogers, H., (2000). *The sounds of language*. Essex: Pearson Education.

Salam, M.S.H., Mohamad, D., & Salleh, S.H.S. (2001). Neural network speaker dependent isolated Malay speech recognition system: Handcrafted vs. genetic algorithm. *6th International, Symposium on Signal Processing and its Applications.* Kuala Lumpur, Malaysia.

Stevens, K. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Sudirman, R., Salleh, S.H., & Salleh S. (2006, November). The effectiveness of DTW-FF coefficients and pitch feature in NN speech recognition. *Proceeding of the Third International Conference on Artificial Intelligence in Engineering and Technology*. Kota Kinabalu, Sabah, Malaysia.

Ting H.N., Yunus J., & Salleh, S.H. (2002). Speaker-independent phonation recognition for Malay plosives using neural networks. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2002).* Honolulu, HI, USA.

Ting, H.N., & Yunus, J. (2004). Speaker-independent Malay vowel recognition of children using multi-layer perceptron. *IEEE Region 10 Conference (TENCON 2004).*

Vuckovic, V., & Stankovic, M. (2001). Formant analysis and vowel classification methods. *5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service (TELSIKS 2001).*