Speaker ethnic identification for continuous speech in Malay language using pitch and MFCC

Rafizah Mohd Hanifa[1], Khalid Isa[2], Shamsul Mohamad[3]
[1,2,3]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Malaysia
[1]Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Nov 14, 2019
Revised Jan 23, 2020
Accepted Feb 1, 2020

Keywords:

Ethnic identification
Feature extraction
Malay language
Mfcc
Support vector machine

ABSTRACT

Voice recognition has evolved exponentially over the years. The purpose of voice recognition or sometimes called speaker identification, is to identify the person who is speaking. This can be done by extracting features of speech that differ between individuals due to physiology (shape and size of the mouth and throat) and also behavioral patterns (pitch, accent and style of speaking). This paper explains an approach of voice recognition to identify the ethnicity of Malaysian people. Pitch and 13 Mel-Frequency Cepstrum Coefficients (MFCCs) are extracted from 52 recorded continuous speech in Malay for use as features to train the classifiers using Tree, Naïve Bayes, Nearest Neighbors and Support Vector Machine (SVM) and another 10 recorded speeches are used for testing. The results reveal that the use of a combination of pitch and 13 coefficients for features extraction and training the data using SVM provide better accuracy (57.7%) than the use of only 13 coefficients (53.8%).

Corresponding Author:

Rafizah binti Mohd Hanifa,
Center for Diploma Studies,
Universiti Tun Hussein Onn Malaysia,
Johor, Malaysia.
Email: rafizah@uthm.edu.my

1. INTRODUCTION

Humans have dreamed of creating robots to interact socially just as humans interact with each other. Applications based on social robots, a kind of humanoid robot, have emerged as a platform with huge potential in the field of human-robot interaction (HRI). Sophia, Jia Jia, Erica [1], Nadine, Pepper [2], Nico [3] and Frog [4] are some examples of robots that had been enhanced with human-like traits to make the communication between robot and human better and easier [5]. Unfortunately, most of the aforementioned social humanoid robots are based on some key languages such as English, Mandarin and Japanese. Since each language reflects the culture of a particular social group, the humanoid must be sensitive to the pitch and intonation of each language in order for it to interpret correctly as well as give appropriate response when communicating with users. Malaysia is a multi-racial country consisting of many ethnic groups such as the Malay, Chinese, Indian and Bumiputera which can be further classified as Iban, Kadazan, Melanau, Murut, Bidayuh and Bajau [6]. Malay language is the national language and it is spoken by the various ethnic groups but their pronunciations may slightly differ when it is not their native language [7]. Thus, this research attempts to identify whether the ethnicity of the speakers can be detected based on the features extracted from the recorded audio. By knowing the ethnicity of the speaker would make it easier to interpret and respond to the different ethnic groups.

The term speech recognition and voice recognition have often been used interchangeably although they are actually different. Speech recognition is concern with the words being spoken whereas voice recognition aims to recognize the speaker rather than the words [8]. Voice recognition is also called speaker identification since each individual's speech is unique based on his/her physiology and behavioral patterns. Pitch, speaking style and accent are some features that contributed to the differences [9].

In voice recognition, feature extraction is one of the procedures that need to be conducted to extract a small amount of data from the speaker's voice signal and these data are kept as a database [10]. Feature extraction is important in the digital waveform to reduce the variability in the continuous speech [11]. There are many different techniques that can be used for feature extraction such as the Linear Predictive Coding (LPC), Perceptual Linear Coding (PLC), Mel-Frequency Cepstrum Coefficient (MFCC), etc [12]. MFCC is mainly designed using the knowledge of human auditory system [13-17]. This paper is organized as follows. Section 2 explains in depth the research methodology adopted in conducting the experiment. Section 3 discusses the results and finally, Section 4 draws the conclusion and highlight avenues for future work.

2. RESEARCH METHOD

The basic representation of speaker identification system consists of pre-processing, feature extraction, and classification as shown in Figure 1 [18]. Interference due to noise often occurs during speech recording causing the performance to be degraded. Thus, before feeding the speech signal to the feature extraction phase, the noise in the signal must be reduced as it is impossible to completely remove it from the signal. In this research, PRAAT is employed to reduce the noise from the original signals based on spectral subtraction method which is one of the first algorithms used for the enhancement of mono channel speech [19]. The basic concept of spectral subtraction is to obtain clean speech by subtracting the noise spectrum from the noisy speech spectrum [20]. Figure 2 shows the result for one speech which has undergone the process.

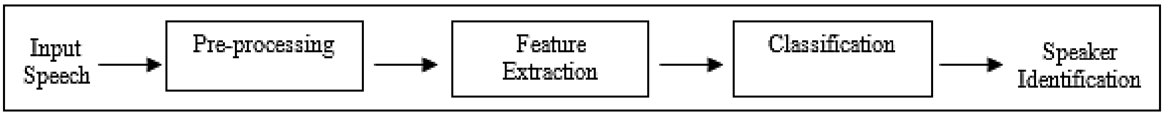
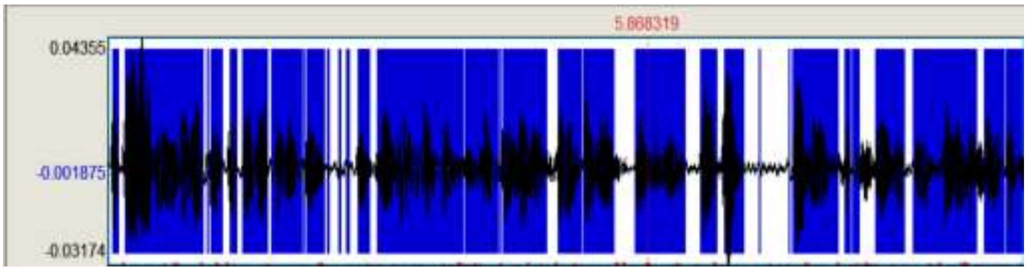
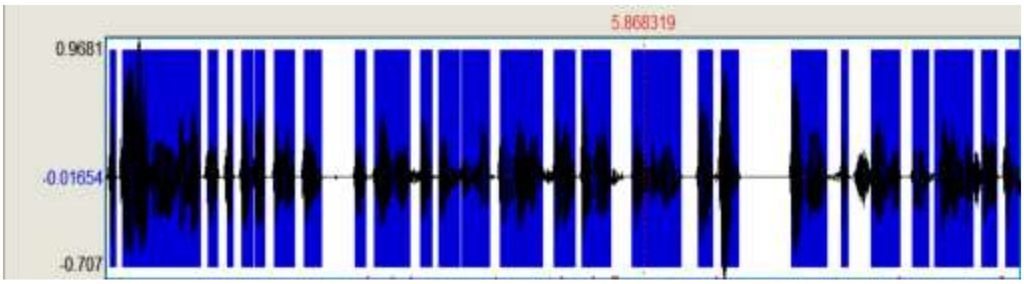


Figure 1. Basic diagram for speaker identification



(a) Original signal



(b) Denoised signal

Figure 2. Before and after noise reduction signal, (a) Original signal and (b) Denoised signal

One of the most important parts of machine learning is feature extraction whereby raw data is turned into information that is useful for machine learning algorithms by eliminating the redundancy present in many types of measured data [21]. Besides, it is also important to avoid having too many features as they require more computational resources during the training stage and might cause overfitting. Thus, this research uses the MFCC technique to capture the main characteristics of 52 continuous speeches due to its high accuracy and high performance. The basic concept of MFCC technique is shown in Figure 3 [22].

Basically, speaker identification involves two phases: training phase and testing phase [23]. Figure 4 shows the block diagram of the methodology used in this work. The denoised speech corpus from 52 recorded speeches in Malay language acquired from local news websites are developed using the audio speeches collected by T.P. Tan [24]. Table 1 shows the demographic profiles of the respondents.

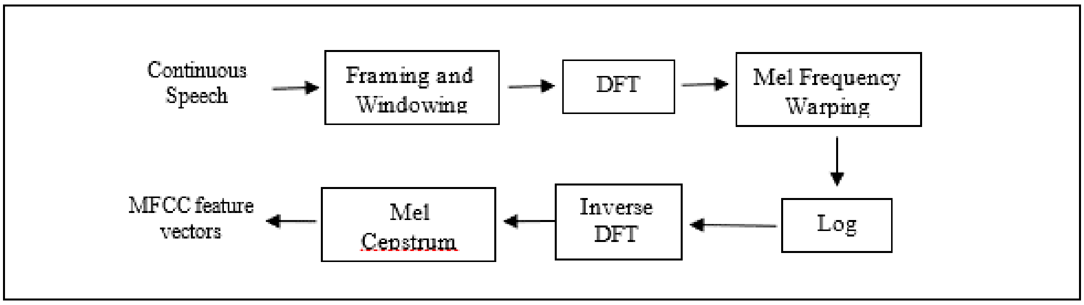


Figure 3. Steps to compute MFCC feature vectors

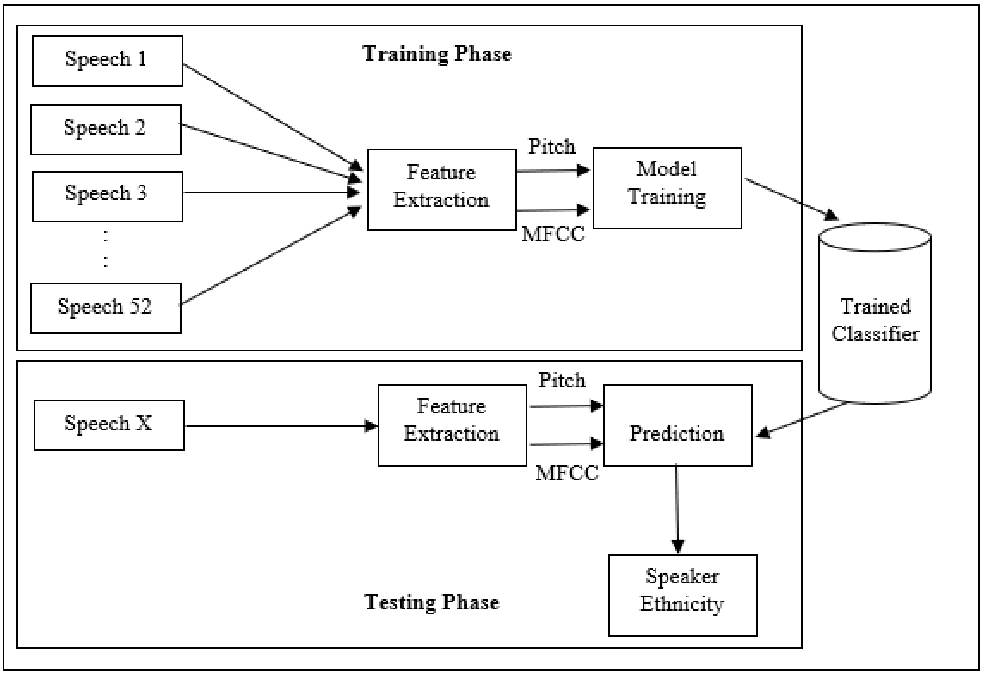


Figure 4. Block diagram of the proposed methodology

Table 1. Demographic of respondents

Ethnic	Gender	Frequency (N=52)	Percentage (%)
Malay	Female	8	15.38
	Male	6	11.53
Chinese	Female	9	17.30
	Male	6	11.53
Indian	Female	8	15.38
	Male	6	11.53
Bumiputera	Female	6	11.53
	Male	3	5.76

Pitch and 13 MFCCs are the features extracted from denoised signals to classify the speaker's ethnicity based on four categories: Malay, Chinese, Indian and Bumiputera. Pitch refers to the frequency of the mechanical movement in the glottis and is very much related to its physical characteristics [25]. Pitch is controlled by the tension in the vocal muscles. The range of frequencies for normal speaking human voice is 70-200 Hz for males and 140-400 Hz for females [26]. PRAAT is used to get the mean pitch for each speech signal. MFCC is one of the popular feature extraction technique used in speech signal [27]. In this work, `mfcc` function in Matlab is used to extract the 13 coefficients. The features of the speech signal are in the form of 13 x N dimensional feature vectors. N is different for each speech depending on the duration of the speech. Thus, the mean for each column of 13 coefficients are determined. These pitch and 13 coefficients (column c0 to c12) are kept in an Excel file as shown in Figure 5 for training purposes and the trained model is then used to test the new data. The same methods are repeated for another 10 new speeches for testing purposes.

Feature Vectors/
Predictors

Class Label/
Response

1	Pitch	c0	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	Ethnic
2	212.7749	-16.1945	3.9530	-0.3544	1.1264	-0.2536	0.6803	-0.8049	0.2359	0.0172	0.3832	-0.0873	0.2806	0.2162	Chinese
3	217.8258	-15.9150	3.0383	-0.0363	1.2534	0.3147	0.1787	-0.3900	0.2695	0.0692	0.0157	0.1103	-0.0271	0.1458	Chinese
4	210.8919	-14.2964	2.3465	0.7771	0.7664	-0.1883	1.0129	-0.2472	0.2025	-0.0925	0.2565	-0.0917	0.3375	0.0474	Chinese
5	204.2245	-15.7130	2.7607	0.3737	0.8784	-0.2200	0.5283	-0.3981	0.4447	0.1270	0.2952	-0.4919	0.3503	0.1433	Chinese
6	208.7963	-15.1094	3.0742	-0.3530	0.8218	-0.1191	0.0011	-0.4591	0.3166	-0.2324	0.3567	-0.2284	-0.0320	0.3494	Chinese
7	212.5492	-16.4010	3.4466	-0.1718	0.6694	0.4492	0.0172	-0.1018	0.0003	0.0901	0.3108	-0.1617	-0.0403	0.0607	Chinese
8	229.9852	-14.5778	2.4319	0.3456	0.7689	0.1455	0.1558	-0.1290	0.0745	0.0013	0.0498	0.0154	0.0726	0.4161	Chinese
9	180.7146	-17.0603	5.3994	0.1629	1.4777	-0.1863	0.5861	-0.6690	0.5249	0.1918	0.3490	-0.1591	-0.0503	0.1312	Chinese
10	166.0858	-20.7777	5.2351	2.8130	0.1594	-0.7012	0.8113	-0.0522	-0.2170	0.3422	-0.0283	0.0469	0.0358	0.1513	Chinese
11	162.9645	-15.7856	2.9632	0.4865	1.0858	-0.0926	0.6982	-0.4161	0.1835	0.0157	0.1414	-0.1782	0.1855	0.0389	Chinese
12	161.2017	-15.6412	3.0924	0.2366	1.1281	-0.0694	0.4659	-0.0827	0.0827	0.1664	-0.0045	-0.1190	-0.2744	0.1337	Chinese
13	117.1732	-14.3864	3.7176	-0.1100	1.2668	-0.1581	0.9195	-0.8555	0.4567	-0.2349	0.3855	0.1232	-0.3173	0.3632	Chinese
14	150.6414	-16.2654	2.6931	0.0207	1.0146	-0.1121	0.3674	-0.1653	0.3627	-0.4804	0.4391	-0.1377	-0.2597	0.2330	Chinese
15	96.9292	-16.7121	3.0063	0.4613	1.0371	-0.2439	0.7293	-0.5060	0.0543	0.0220	-0.0273	0.0044	-0.1308	0.1085	Chinese
16	223.9309	-16.3687	2.9726	1.0073	0.7187	-0.3113	0.6115	-0.4984	0.4992	-0.1230	-0.0428	0.3757	-0.0691	0.1004	Malay
17	236.9767	-15.6918	2.0092	0.8535	1.0158	-0.0877	0.0883	-0.2465	0.0968	-0.0992	0.1569	0.3483	0.1321	0.0905	Malay
18	183.5125	-24.9809	3.3133	1.0224	1.3532	0.0498	0.1029	-0.0126	0.3975	-0.0123	0.2412	-0.0538	-0.2986	0.2112	Malay
19	204.0728	-23.8167	3.1526	0.4147	1.1142	-0.3138	0.1903	-0.0732	0.2771	-0.2111	-0.1389	0.1063	0.0969	0.2261	Malay
20	204.3083	-14.7101	3.0739	0.2503	1.3449	-0.1217	0.7746	-0.4006	0.7534	-0.1417	0.3151	-0.1003	-0.0919	0.4171	Malay
21	204.2883	-13.5839	3.4552	0.1399	1.4719	-0.3868	0.6749	-0.4059	0.5844	-0.2161	0.3223	-0.2665	0.1574	0.2897	Malay
22	207.5525	-15.2019	3.2852	0.1482	1.6576	-0.7047	0.5902	-0.2620	0.2155	0.0241	0.4030	-0.2355	0.1848	0.3764	Malay

Figure 5. Features compilation in Excel for training

3. RESULTS AND ANALYSIS

Classification Learner apps is a machine learning provided by Matlab to train models to classify data. The advantages of this app include allowing user to select features, specify validation schemes, train model and even assess the results. Automated learning can be done to get the best type of classification model. To perform this, a known set of input data (observations) and known responses to the data (classes) are needed. These data are used to train a model that generates predictions for the response to new data. This can be done by exporting the model to the workspace to create the trained model. In this research, we performed two models before deciding on the best model to be used.

3.1. Model 1-13 coefficients

In Model 1, the input data uses only 13 coefficients extracted with the exclusion of pitch during the training phase. The Tree, Naïve Bayes, Nearest Neighbors and SVM are used as classifiers to determine which one gives the highest accuracy. Table 2 shows the result on accuracy for each classifier and it can be seen that the Linear SVM provides the highest accuracy at 53.8%.

Figure 6 shows the confusion matrix for Model 1 indicating areas where the classifier has performed poorly. The row shows the true class and the columns show the predicted class. Since this work uses cross-validation, the confusion matrix is calculated using the predictions on the held-out observations. The diagonal cells show where the true class and predicted class match. As can be seen, the Indian ethnic group has the highest number of matching followed by the Malay, Chinese and lastly, the Bumiputera group.

The lowest number for the Bumiputera group may be attributed to the small sample of such speakers collected for this research. The trained model is then exported to the workspace in Matlab to be tested with 10 new data speeches to determine the number of correct predictions. Table 3 shows the predicted results using Linear SVM for Model 1 where 10 speeches are compared to the class label of testing data. Only ethnicity of six speeches (60%) were predicted correctly.

Table 2. Results of accuracy with 13 features

Type of Classifier		Accuracy (%)
Tree	Fine Tree	30.8
	Medium Tree	30.8
	Coarse Tree	36.5
Nearest Neighbors	Fine KNN	40.4
	Medium KNN	40.4
	Coarse KNN	25.0
	Cosine KNN	42.3
	Cubic KNN	40.4
	Weighted KNN	42.3
Naïve Bayes	Gaussian Naïve Bayes	51.9
	Kernel Naïve Bayes	48.1
Support Vector Machine	Linear SVM	53.8
	Quadratic SVM	46.2
	Cubic SVM	50.0
	Fine Gaussian SVM	32.7
	Medium Gaussian SVM	46.2
	Coarse Gaussian SVM	28.8

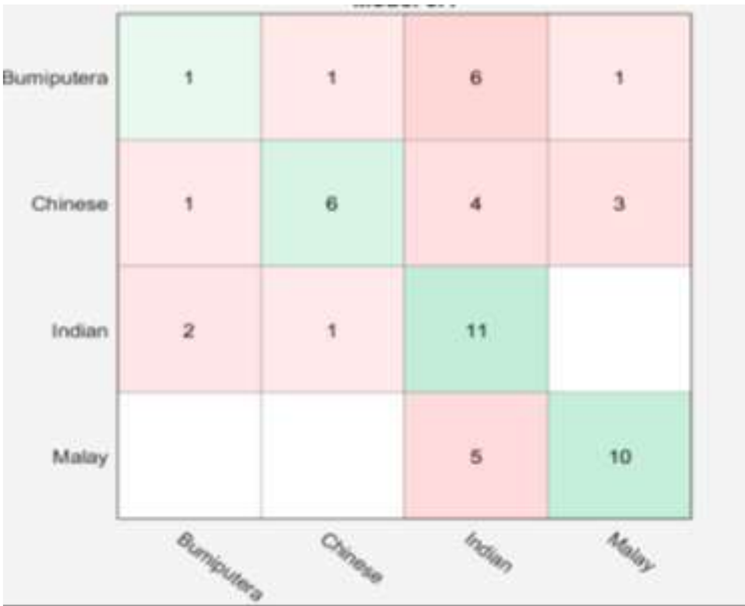


Figure 6. Confusion Matrix for Model 1

Table 3. Predicted result using Linear SVM for Model 1

Class Label of Testing Data	Predicted Result using Linear SVM
Chinese	Indian
Chinese	Malay
Bumiputera	Indian
Indian	Indian
Indian	Indian
Malay	Malay
Chinese	Chinese
Indian	Indian
Malay	Malay
Bumiputera	Indian

3.2. Model 2-Pitch + 13 Coefficients

In Model 2, the input data uses pitch and 13 coefficients. The Tree, Naïve Bayes, Nearest Neighbors and SVM are used as classifiers to see which gives the highest accuracy. Table 4 shows the results of accuracy for each classifier. Again, Linear SVM gives the highest accuracy at 57.7%.

Figure 7 shows the confusion matrix for Model 2. The Indian ethnic group still maintains the highest number of matching followed by the Malay and Chinese groups which share the same number of matching. Interestingly, the number of matching for the Bumiputera group has now increased by 1 using Model 2 compared to Model 1. The increment might be due to the additional features included in Model 2. Table 5 compares the class label of testing data with the predicted result for Model 2. Seven speeches (70%) were predicted correctly in Model 2.

Table 4. Results of accuracy with 14 features

Type of Classifier		Accuracy (%)
Tree	Fine Tree	30.8
	Medium Tree	30.8
	Coarse Tree	32.7
Nearest Neighbors	Fine KNN	44.2
	Medium KNN	38.5
	Coarse KNN	25.0
	Cosine KNN	34.6
	Cubic KNN	48.1
	Weighted KNN	40.4
Naïve Bayes	Gaussian Naïve Bayes	40.4
	Kernel Naïve Bayes	42.3
Support Vector Machine	Linear SVM	57.7
	Quadratic SVM	48.1
	Cubic SVM	48.1
	Fine Gaussian SVM	28.8
	Medium Gaussian SVM	50.0
	Coarse Gaussian SVM	28.8

Figure 7. Confusion Matrix for Model 2

Table 5. Predicted result using Linear SVM for Model 2

Class Label of Testing Data	Predicted Result using Linear SVM
Chinese	Indian
Chinese	Malay
Bumiputera	Bumiputera
Indian	Indian
Indian	Indian
Malay	Malay
Chinese	Chinese
Indian	Indian
Malay	Malay
Bumiputera	Indian

4. CONCLUSION

The research conducted has shown that the combination of pitch and MFFCs as features in speech provide better accuracy in predicting the speaker's ethnicity. Among the classifiers used in this research, Linear SVM provides the most accurate trained model. Future work may consider other feature such as formants and intensity to get a better trained model in predicting speaker's ethnicity.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research under the Tier-1 Research Grant (U926).

REFERENCES

[1] D. F. Glas, *et al.*, "ERICA: The ERATO intelligent conversational android," *IEEE International Symposium Robot and Human Interactive Communication*, pp. 22-29, 2016.
[2] F. Tanaka, *et al.*, "Pepper learns together with children: development of an educational application," *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 270-275, 2015.
[3] P. A. J. Ven, "Implementing EyePi architecture on R3D3," *Bachelor's thesis, University of Twente,* 2017.
[4] J. Diprose, *et al.*, "Designing an API at an appropriate abstraction level for programming social robot applications", *Journal of Visual Languages & Computing,* vol. 39, pp.22-40, 2017.
[5] K. Kantharak, *et al.*, "Design and development of service robot based human-robot interaction (HRI)," *IEEE 2017 International Conference on System Science and Engineering,* pp. 293-296.
[6] S. Nagaraj, *et al.*, "Counting ethnicity in malaysia: the complexity of measuring diversity," *Social Statistics and Ethnic Diversity. IMISCOE Research Series, Springer, Cham,* 2015.
[7] R. M. Hanifa, K. Isa, and S. Mohamad, "Voiced and Unvoiced separation in malay speech using zero crossing rate and energy," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 16, no. 2, pp.775-780, 2019.
[8] C. Kikel, "Difference between voice recognition and speech recognition," [Blog] Total Voice Technologies, 2019. Available at http://www.totalvoicetech.com [Accessed 17 October 2019].
[9] R. M. Hanifa, K. Isa and S. Mohamad, "Malay speech recognition for different ethnic speakers: an exploratory study," *IEEE Symposium on Computer Applications & Industrial Electronics*, pp. 91-96, 2017.
[10] A. Jain and O.P. Sharma, "A Vector quantization approach for voice recognition using mel frequency cepstral coefficient (MFCC): A Review," *International Journal of Electronics and Communication Technology,* vol. 4, no. 4, pp. 27 -29, 2013.
[11] C. Vimala and V. Radha, "Suitable feature extraction and speech recognition technique for isolated tamil spoken words," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 1, pp. 378-383, 2014.
[12] A. Verma and I. Kaur, "Automatic speech recognition using mel-frequency cepstrum coefficient (MFCC and Vector Quantization (VQ) techniques for continous speech," *International Journal of Advanced and Applied Sciences,* vol. 5, no. 4, pp.73-78, 2015.
[13] P. Pal Singh, and P. Rani, "An approach to extract feature using MFCC," *IOSR Journal of Engineering,* vol. 4, no. 8, pp.21-25, 2014.
[14] Z. Wanli, and L. Guoxin, "The research of feature extraction based on MFCC for speaker recognition," *International Conference on Computer Sciences and Network Technology*, pp.1074-1077.
[15] G. B. Janvale, B. A. Marathwada, and S. Gambhire, "Speech feature extraction using mel-frequency cepstral coefficient (MFCC)," *Emerging Trends in Computer Science, Communication and Information Technology*, pp. 503-506, 2010.
[16] M. Cutajar, *et al.*, "Comparative study of automatic speech recognition techniques," *IET Signal Processing,* vol. 7, no. 1, pp. 25-46, 2013.
[17] A. Hassan Mansour, G. Alabdeen Salh, and H. Zeen Alabdeen, "Voice recognition using back propagation algorithm in neural networks," *International Journal of Computer Trends and Technology,* vol. 23, no. 3, pp. 132-139, 2015.
[18] R M. Gamit, K. Dhameliya, and S. N. Bhatt, "Classification techniques for speech recognition: a review," *International Journal of Emerging Technology and Advanced Engineering,* vol. 5, no. 2, pp.58-63, 2015.

[19] N. Upadhyaya, and A. Karmakar, “Speech enhancement using spectral subtraction-type algorithms: a comparison and simulation study,” *Eleventh International Multi-Conference on Information Processing,* pp. 574-584, 2015.

[20] V. Ekaterina and S.Boris “Enhanced spectral subtraction method for noise reduction with minimal speech distortion”, *17th International Conference on Systems, Signals and Image Processing,* 2010.

[21] “Mastering Machine Learning: A Step-by-Step Guide with MATLAB”, [e-book] https://www.mathworks.com/campaigns/offers/mastering-machine-learning-with-matlab.html

[22] S. B. Magre, P. V. Janse and R. R. Deshmukh, “A review on feature extraction and noise reduction technique”, *Int. J. of Advanced Research in Computer Science and Software Engineering,* vol. 4, no. 2, pp. 352-356, 2014.

[23] T. R. Suchitha, and A.T. Bindu, “Feature extraction using MFCC and classification using GMM,” *International Journal for Scientific Research and Development,* vol. 3, no. 5, pp.1278- 1283, 2015.

[24] T.P. Tan, and B. Ranaivo-Malançon, “Malay grapheme to phoneme tool for automatic speech recognition,” *Third International Workshop on Malay and Indonesian Language Engineering,* 2009.

[25] M. Jiménez Hernández, “A tutorial to extract the pitch in speech signals using autocorrelation,” *Open Journal of Technology & Engineering Disciplines*, vol. 2, no. 1, pp.1-11, 2016.

[26] The Human Voice. Retrieved from https://www.colorado.edu/physics/phys1240/phys1240_sm09/index_files/L17

[27] B. J. Ramgire, and S. M. Jagdale, “A survey on speaker recognition with various feature extraction techniques”, *International Research Journal of Engineering and Technology,* vol. 3, no. 4, pp.709-712, 2016.

BIOGRAPHIES OF AUTHORS

Rafizah Mohd Hanifa obtained her bachelor degree in Computer Science from Universiti Sains Malaysia (USM) in 1999, followed by master degree in Information Technology at the Faculty of Information Technology, Universiti Utara Malaysia (UUM) in 2001. She is currently a Ph.D candidate at the Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM). Her research interests include speech technology, artificial intelligent, information system and augmented reality.

Khalid Isa graduated from Universiti Teknologi Malaysia, in 2001 with a BSc in Computer Science. He then pursued his MSc. in Computer System Engineering at Universiti Putra Malaysia, graduating in 2005. In 2014, he completed his PhD degree in Electrical and Electronic Engineering at Universiti Sains Malaysia, with specialization in Computational Intelligence and Underwater Robotics. He is currently a Senior Lecturer in the Computer Engineering Department, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM). He is the Chairman of IEEE Oceanic Engineering Society (OES) Malaysia Chapter. His research interests are underwater robotics, computational intelligence, bio-inspired engineering, internet of things, image processing, control, and autonomous system.

Shamsul Mohamad obtained his BSc in Computer Science from Universiti Teknologi Malaysia in 1999 and MSc in Computer Science from Universiti Sains Malaysia in 2004. Currently, he is a Senior Lecturer at the Computer Engineering Department, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM). His research interests include crowd simulation, artificial intelligence and Internet of Things.