Indonesian Journal of Electrical Engineering and Computer Science
Vol. 10, No. 2, May 2018, pp. 554~561
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v10.i2.pp554-561

Speech Emotion Recognition Using Deep Feedforward Neural Network

Muhammad Fahreza Alghifari[1], Teddy Surya Gunawan*[2], Mira Kartiwi[3]
[1,2] Department of Electrical and Computer Engineering, Kulliyyah of Engineering, Malaysia
[3] Department of Information Systems, Kulliyyah of ICT, International Islamic University Malaysia, Malaysia

Article Info

Article history:

Received Nov 26, 2017
Revised Jan 23, 2018
Accepted Feb 21, 2018

Keywords:

Deep neural network
Mel-frequency cepstral coefficients (MFCC)
Speech emotion recognition (SER)

ABSTRACT

Speech emotion recognition (SER) is currently a research hotspot due to its challenging nature but bountiful future prospects. The objective of this research is to utilize Deep Neural Networks (DNNs) to recognize human speech emotion. First, the chosen speech feature Mel-frequency cepstral coefficient (MFCC) were extracted from raw audio data. Second, the speech features extracted were fed into the DNN to train the network. The trained network was then tested onto a set of labelled emotion speech audio and the recognition rate was evaluated. Based on the accuracy rate the MFCC, number of neurons and layers are adjusted for optimization. Moreover, a custom-made database is introduced and validated using the network optimized. The optimum configuration for SER is 13 MFCC, 12 neurons and 2 layers for 3 emotions and 25 MFCC, 21 neurons and 4 layers for 4 emotions, achieving a total recognition rate of 96.3% for 3 emotions and 97.1% for 4 emotions.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
Kulliyyah of Engineering, Malaysia.
Email: tsgunawan@iium.edu.my

1. INTRODUCTION

Speech Emotion Recognition (SER) can be defined as the identification of the emotional state of the speaker from his or her speech signal [1]. SER is one of the topics in speech processing that has been continuously researched for decades, the simplest attempts dates back from the late fifties [2]. In today's world, SER has shown to be quite a research hotspot, as indicated by the growth of publication papers in each year.

The application of SER can be targeted to several sectors. In banking, an auto caller equipped with SER may assist in detecting the emotion of the customer, generating custom responses based on the result. In education, an e-learning portal with SER can detect the emotions of the user such as frustration and stress, determining whether the studying is conducive or not and give appropriate countermeasures. Yet another application is in transportation, where in the near-future that vehicles are capable of auto-driving, the system can take over the steering wheel in the case where an unhealthy amount of emotion is detected from the driver.

A typical speech emotion recognition is illustrated in Figure 1. The feature extraction marks the start of a SER system. This includes selecting the features appropriate for emotion recognition. Next these features are processed by a classifier. These classifiers are trained by referring to an emotion database. Next, the system will be put into testing by crosschecking with the same database. The processed data obtained will be the determinant of the decision, typically in terms of accuracy and processing time.

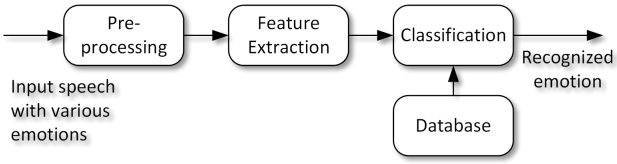


Figure 1. A Typical Speech Emotion Recognition Algorithm

In [3], we conducted a comprehensive literature review on SER in which our proposed system is based on that critical review. Current SER is still in development stage. Although it has been continuously researched in the past decade, the implementation in our daily life and practical uses is still very limited. Many things are left to be desired, in terms of accuracy, processing time and practicality. In this study, we improved the accuracy processing time by optimizing various deep neural network configurations.

2. PROPOSED SPEECH EMOTION RECOGNITION SYSTEM

The SER flow conducted in this research is shown in Figure 2. For the feature extraction, this study has chosen the Mel-frequency cepstral coefficients (MFCC) due to the its nature to be tuned in a scale that is suitable for the human ear [4], best suited for N-way classifiers [2], and is one of the most popular feature to be extracted in SER, such as in [5] and [6]. For the classifier, the chosen algorithm is the deep neural network, a branch of the artificial neural network. It employs less parameters, higher performance compared to VQ model, has very high potential given more hidden layers [7]. Neural networks has gained popularity in SER systems, such as conducted by [8] and [9].

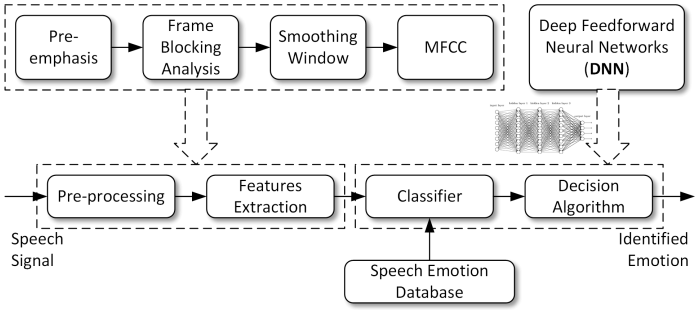


Figure 2. Proposed Speech Emotion Recognition System [3]

2.1 Mel-Frequency Cepstral Coefficients (MFCCs) Feature Extraction

MFCCs use a non-linear frequency scale, i.e. mel scale, based on the auditory perception. A *mel* is a unit of measure of perceived pitch or frequency of a tone. Eq. (1) can be used to convert frequency scale to mel scale.

$$f_{mel} = 177 \ln\left(1 + \frac{f_{Hz}}{700}\right) \quad (1)$$

Where f_{mel} is the frequency in mels and f_{Hz} is the normal frequency in Hz. MFCCs are often calculated using a filter bank of M filters, in which each filter has a triangular shape and is spaced uniformly on the mel scale as shown in Equation (2).

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{k - f[m-1]}{f[m] - f[m-1]} & f[m-1] < k \leq f[m] \\ \frac{f[m+1] - k}{f[m+1] - f[m]} & f[m] < k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (2)$$

where $m = 0,1,\cdots,M-1$. The log-energy mel spectrum is then calculated as follows:

$$S[m] = \ln[\sum_{k=0}^{N-1}|X[k]|^2 H_m[k]] \quad m = 0,1,\cdots,M-1 \tag{3}$$

where $X[k]$ is the discrete Fourier transform (DFT) of a speech input $x[n]$.

Although traditional cepstrum uses inverse discrete Fourier transform (IDFT), mel frequency cepstrum is normally implemented using discrete cosine transform (DCT) since $S[m]$ is even as shown in Eq. (4), as follows:

$$\hat{x}[n] = \sum_{k=0}^{N-1} S[m] \cos\left[\left(m+\frac{1}{2}\right)\frac{\pi n}{M}\right] \quad m = 0,1,\cdots,M-1 \tag{4}$$

Typically, the number of filters *M* ranges from 20 to 40, and the number of kept coefficients is 13. Some research reported that the performance of speech recognition and speaker identification systems reached peak with 32-35 filters [10].

2.2 Deep Neural Networks (DNNs) Classifier

The main principle of DNN is to utilize lower level features learning to update the learning of higher features. There are many available *deep architecture*, such as neural networks with many hidden layers and/or many hidden variables, convolutional neural networks, recurrent neural networks, and deep belief network [11]. In this research, we used deep learning using feedforward neural network architectures with multilayers hidden layers with many hidden variables. Figure 3 illustrates the deep feedforward neural network structure.

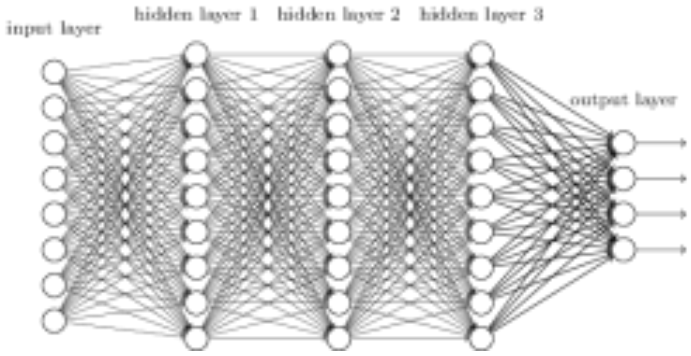


Figure 3. Deep Feedforward Neural Network Structure

2.3 Speech Emotion Database and Recording of New Speech Signals Database for Testing

The database used for training in this research is the open-source Berlin Database of Emotional Speech (Emo-DB) [12], which contains 535 emotional audio files in (.wav) format from 10 different speakers (5 male 5 females). Each speaker utters 10 German sentences under 7 different emotions (anger, boredom, disgust, fear, happiness, sadness, neutral). For this research, only four emotions are selected, including anger, happiness, sadness and neutral.

As additional contribution to the SER research field, this research has attempted to create a new emotional dataset. Inspired by the databases created in [13] and [14], the custom dataset contains elicited English emotional data from 9 speakers in 4 different emotions, happy, angry, sad and neutral. For each emotion, the speaker is prompted to act out 5 lines, for an initial total of 180 emotional lines.

The recording medium used is a conventional phone's recorder and audio transmitted through WhatsApp voice messaging in a noisy environment. The primary reason is to simulate the conversation in a real-life environment. The recording is normally performed in a recording studio where there is minimal noise and interference. But in reality, conversations are held everywhere, therefore samples that have noises are more valuable to train the network.

The second reason is because of the compression factor. In conventional emotional databases, audio quality is one of the preferred factors so that feature extraction is achieved without much difficulties. On the

other hand, in modern Voice-over-Internet-Protocol (VoIP), audio compression is one of the major considerations to ensure that the channel can simulate 'real-time' conversations while sending minimum data over the connection.

To ensure that the quality of the emotional audio is acceptable, each audio files are coded appropriately and randomly selected. For each audio file, the validator will guess what emotion that is conveyed. For each correct answer the audio file will be moved to a validated pool. For validation integrity, the process was conducted in a near silence lab, equipped with a noise-cancelling headphone. From 180 voice lines, 148 are distinguishable and considered to be added to our SER database.

3. RESULTS AND DISCUSSION

The main focus of this section is to investigate the ideal network configuration by adjusting the number of MFCC, neurons in each layer, as well as number of layers. For the first stage, 3 emotions are taken into consideration – happy, angry, and sad. Afterwards, an additional emotion, neutral is added to the system, retrained and tested. The first investigation was the extraction of MFCC from database in terms of processing time. For each coefficient from 1 to 100, the processing time is recorded and repeated for 5 times, then the average was calculated as shown in Figure 4.

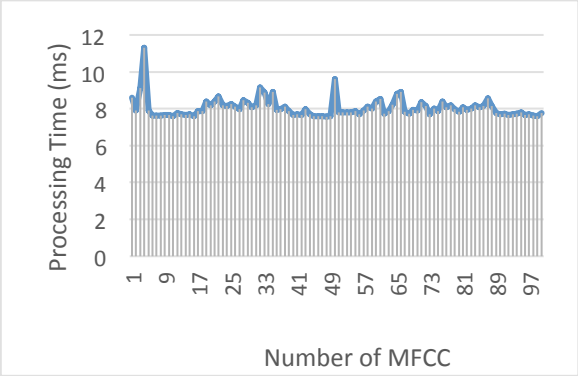


Figure 4. Experiments on Number of MFCC and Its Processing Time

From the results, it can be observed that the MFCC extraction process is lightweight across number of coefficients with small variance and standard deviation of 0.265 and 0.51, respectively. Even going up to 100 MFCC, the processing time is still considerably low at 7.756 ms. Having a negligible processing time means that proposed system do not have constrained on the number of coefficients which enables the focus to be on other parameters.

The next step is investigating the best performance of neural network configuration, in which the following Matlab's function were used, including feedforwardnet(), patternnet(), fitnet(), and cascadedfeedforwardnet(). The default number of neurons is 10 with a single hidden layer. The training, validation and testing ratio is 70%:15%:15%, respectively. The results is shown in Table 1.

Table 1. Results of Various Neural Network Structures

Matlab Function	Variance	Standard Deviation	Best Performance	Worst Performance
feedforwardnet()	0.00050	0.02231	0.96648	0.84575
patternnet()	0.00315	0.05610	0.92778	0.63111
fitnet()	0.003574	0.059782	0.95778	0.64444
cascadeforwardnet()	0.003816	0.006177	0.96333	0.63778

From observation, feedforwardnet() algorithm provides the best overall reliable performance compared to its counterparts, boosting a recognition rate of 0.966 at 13 MFCC while maintaining less variance and standard deviation. This algorithm is chosen to be the foundation of deep neural network. The next stage of optimizing is finding the ideal number of neurons in a layer. For this process, the MFCC is kept constant at 13 and the number of neurons is increased.

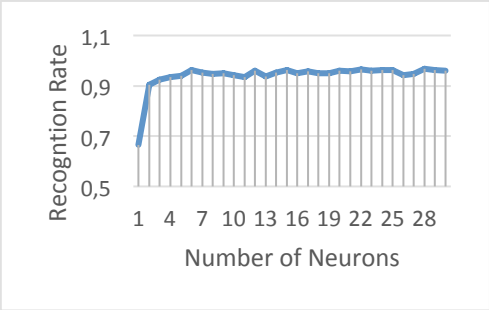


Figure 5. Experiments on Number of Neuron in the Hidden Layer

As shown in Figure 5, the overall results improved in terms of recognition rate, peaking at number of neuron of 12. Finally, the number of hidden layers will be varied between 1 and 4. Using the same number MFCC and neurons, layers are added incrementally by 1 until the results can no longer improve. The increase in layers should significantly impact the processing time, therefore it should now be taken into account. The final results are shown in Table 2.

Table 2. Experiments on Number of Hidden Layers

Number of Layer	Best Performance	Processing Time (s)
1	0.960	3.919
2	0.972	4.128
3	0.968	4.608
4	0.969	4.104

For three emotions evaluated, it has been found that the optimum configuration is 13 MFCC, 12 neurons in 2 hidden layers. The next step is determining whether this configuration is suitable for all MFCC based SER systems. The neutral emotion is added into the training pool. The first result is obtained using the former network configuration of 13 MFCC, 12 neurons in 2 hidden layers resulting a best case recognition rate of 96.67% with processing time of 4.997 s.

The steps of optimization are then repeated – optimize number of MFCC, neuron, and layer, repeating training and testing 5 times each to ensure accuracy. After optimization, the ideal configuration is 25 MFCC, 21 neurons, and 4 hidden layers for 4 emotions achieving a recognition rate of 97.1% in 12.013s. The confusion matrix and network configuration are displayed in Figure 6 and 7.

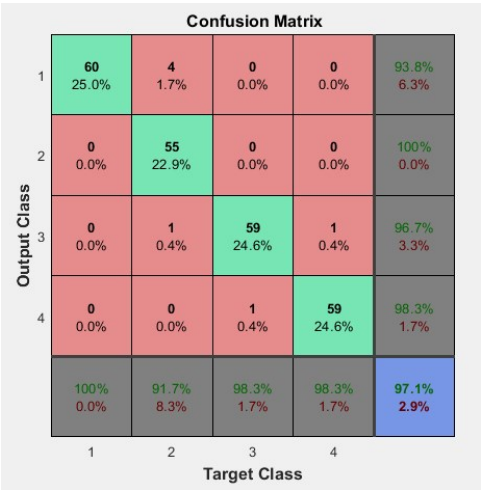


Figure 6. Optimized Network Confusion Matrix

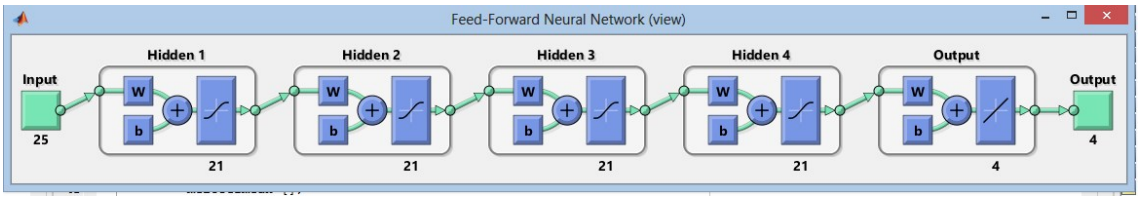


Figure 7. Optimized Deep Feedfoward Neural Network for 4 Emotions

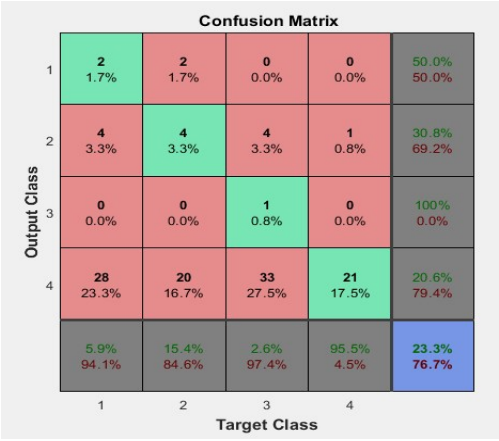
Once the ideal network configuration of deep feedforward neural network is finalized, the final step of the study is validation of the custom emotional database. We used the same trained network configuration but now we used custom database. Note that, as explained in the previous section, the custom database was recorded in the real life setting with environmental noise. As shown in Figure 8(a), the recognition rate is rather low at 23.3%. As the results were rather unsatisfying at 23.3% recognition rate, a new network is created. This network uses the custom emotional database's own audio data for training and testing. The results in Figure 8(b) have shown to be more promising at 70.8% recognition rate.

From the results obtained, one can observe that the newly calibrated network for 4 emotions performed better (97.1%) than the former one trained for 3 emotions (96.67%), at a trade off of a significantly higher processing time. Therefore, if the objective is to maximize recognition rate for each unique number of emotions, it is recommended to recalibrate the ideal number of MFCC, neuron and layer. With that said, the difference in recognition rate is a mere difference of 1%. This indicates that the optimized network configuration can be used for other number of emotions with a certain degree of toleration of error.

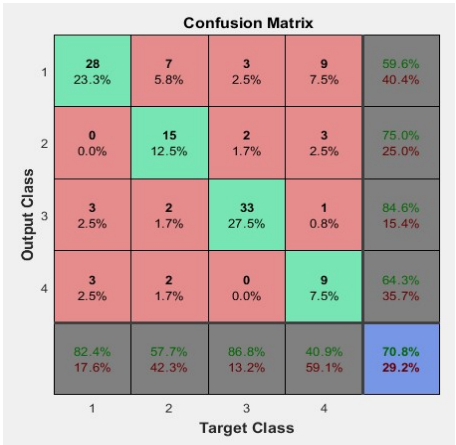
The results obtained in the optimization of network has shown to have a significant improvement compared to the SER research counterparts. The benchmark study conducted by [15] with an unmentioned database has achieved a 92.3% performance while the study by [16] using the same Berlin Emotion Database achieved 65% recognition. Comparatively, the recognition rate achieved using the system proposed in this study is 97.1%.

A factor to consider is the variation of parameters used. In this study, MFCC is the sole input while the study conducted by [15] uses MFCCs, perceptual linear predictive (PLPs), and Filter banks (FBANKs). By theory, adding more parameters should improve the recognition rate at the cost of processing complexity and time. In other words, achieving a high recognition rate using less parameters is desirable.

Another factor that should be understood that the number of emotion analyzed. In this study, only 3 and 4 emotions were analyzed while their study attempts to process 6-7 emotions. Theoretically, increasing the number of emotions to be recognized will deteriorate the accuracy rate. Hence achieving high recognition rate for multiple emotions is commendable. With that said, our result using limited emotions is acceptable as the purpose of the network is to be later employed in real-life situation. The 4 emotions analyzed in this study are the primarily emotions that have practical applications. Maximizing the recognition rate of these emotions are of a higher priority.



(a) Without Network Retraining



(b) With Network Retraining

Figure 8. Performance of the SER System on the Custom Database

For the custom database, the results when using the optimized network only achieved a performance of 23.3%. There are several hypothesis to explain the poor result. The primary reason proposed is due to the language difference. For example, the Arabic language may sound more intense than say, the traditional Sundanese language of Indonesia. In the original network, German emotional sentences were used to train the deep neural network while the custom database uses English. Taking this into account, the results have greatly improved when training using the same dataset, achieving 70.8% in best case performance. This validates the idea of language being a factor in SER. Aside from the language difference, the lossy compression is another likely factor to explain the difference in performance.

4. CONCLUSIONS AND FUTURE WORKS

After extensive experimentation, SER system has been implemented and optimized. The optimum configuration for SER is 13 MFCC, 12 neurons and 2 layers for 3 emotions and 25 MFCC, 21 neurons and 4 layers for 4 emotions, achieving a total recognition rate of 96.3% for 3 emotions and 97.1% for 4 emotions. If the objective is to maximize recognition rate for each unique number of emotions, it is recommended to recalibrate and retrain the network, otherwise the former network is still utilizable given a certain degree of recognition error. In SER, language used is a factor to be considered for recognition rate. Future research include the improvement on deep learning configuration, different database, and real life applications.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Malaysian Ministry of Higher Education (MOHE), which has provided funding for the research through the Fundamental Research Grant Scheme, FRGS15-194-0435.

REFERENCES

[1] A. Joshi, R. Kaur, "A Study of speech emotion recognition methods," *Int. J. Comput. Sci. Mob. Comput.(IJCSMC)*, vol. 2, pp. 28-31, 2013.
[2] M. El Ayadi, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.
[3] T. S. Gunawan, M. F. Alghifari, M. A. Morshidi, M. Kartiwi, "A Review on Speech Emotion Recognition Algorithms," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 6, 2018.
[4] A. Konar, A. Chakraborty, *Emotion Recognition: A Pattern Analysis Approach*, John Wiley & Sons, 2014.
[5] S. R. Bandela, T. K. Kumar, *"Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC,"* in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.
[6] S. T. Saste, S. M. Jagdale, *"Emotion recognition from speech using MFCC and DWT for security system,"* in 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 701-704, 2017.
[7] E. Gopi, *Digital speech processing using Matlab*, Springer, 2014.
[8] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," *IEEE Transactions on Multimedia*, vol. PP, pp. 1-1, 2017.
[9] C. W. Huang, S. S. Narayanan, *"Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition,"* in 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 583-588, 2017.
[10] V. Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, pp. 19-22, 2010.
[11] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436, 2015.
[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," pp. 1517-1520.
[13] D. Pravena, S. Nandhakumar, D. Govind, "Significance of natural elicitation in developing simulated full blown speech emotion databases," pp. 261-265.
[14] A. Chandran, P. Duplex, G. Divu, *Development of speech emotion recognition system using deep belief networks in malayalam language*, 2017.
[15] J. Niu, Y. Qian, K. Yu, "Acoustic emotion recognition using deep neural network," pp. 128-132, 2014.
[16] X. Zhou, J. Guo, R. Bie, *"Deep Learning Based Affective Model for Speech Emotion Recognition,"* in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 841-846, 2016.

BIOGRAPHIES OF AUTHORS

Muhammad Fahreza Alghifari has completed his B.Eng. (Hons) degree in Electronics: Computer Information Engineering from International Islamic University Malaysia (IIUM) in 2018. His research interests are in signal processing, artificial intelligence and affective computing. He received a best FYP award from IEEE Signal Processing – Malaysia chapter. Currently, he is working on the application of speech emotion recognition for suicide prevention.

Teddy Surya Gunawan received his BEng degree in Electrical Engineering with cum laude award from Institut Teknologi Bandung (ITB), Indonesia in 1998. He obtained his M.Eng degree in 2001 from the School of Computer Engineering at Nanyang Technological University, Singapore, and PhD degree in 2007 from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. His research interests are in speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He is currently an IEEE Senior Member (since 2012), was chairman of IEEE Instrumentation and Measurement Society – Malaysia Section (2013 and 2014), Associate Professor (since 2012), Head of Department (2015-2016) at Department of Electrical and Computer Engineering, and Head of Programme Accreditation and Quality Assurance for Faculty of Engineering (since 2017), International Islamic University Malaysia. He is Chartered Engineer (IET, UK) and Insinyur Profesional Madya (PII, Indonesia) since 2016.

Mira Kartiwi completed her studies at the University of Wollongong, Australia resulting in the following degrees being conferred: Bachelor of Commerce in Business Information Systems, Master in Information Systems in 2001 and her Doctor of Philosophy in 2009. She is currently an Associate Professor in Department of Information Systems, Kuliyyah of Information and Communication Technology, International Islamic University Malaysia. Her research interests include electronic commerce, data mining, e-health and mobile applications development.