Bulletin of Electrical Engineering and Informatics
Vol. 8, No. 4, December 2019, pp. 1324~1332
ISSN: 2302-9285, DOI: 10.11591/eei.v8i4.1646

On the use of voice activity detection in speech emotion recognition

Muhammad Fahreza Alghifari[1], Teddy Surya Gunawan[2], Mimi Aminah binti Wan Nordin[3], Syed Asif Ahmad Qadri[4], Mira Kartiwi[5], Zuriati Janin[6]
[1,2,3,4]Department of Electrical and Computer Engineering, International Islamic University Malaysia, Malaysia
[2]Visiting Fellow, School of Electrical Engineering and Telecommunications, UNSW, Australia
[5]Information Systems Department, International Islamic University Malaysia, Malaysia
[6]Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia

Article Info

Article history:

Received Mar 19, 2019
Revised May 25, 2019
Accepted Jun 26, 2019

Keywords:

Deep neural network
Speech emotion recognition
Voice activity detection

ABSTRACT

Emotion recognition through speech has many potential applications, however the challenge comes from achieving a high emotion recognition while using limited resources or interference such as noise. In this paper we have explored the possibility of improving speech emotion recognition by utilizing the voice activity detection (VAD) concept. The emotional voice data from the Berlin Emotion Database (EMO-DB) and a custom-made database LQ Audio Dataset are firstly preprocessed by VAD before feature extraction. The features are then passed to the deep neural network for classification. In this paper, we have chosen MFCC to be the sole determinant feature. From the results obtained using VAD and without, we have found that the VAD improved the recognition rate of 5 emotions (happy, angry, sad, fear, and neutral) by 3.7% when recognizing clean signals, while the effect of using VAD when training a network with both clean and noisy signals improved our previous results by 50%.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
International Islamic University Malaysia,
Jalan Gombak, 51300 Selangor, Malaysia.
Email: tsgunawan@iium.edu.my

1. INTRODUCTION

Speech emotion recognition (SER) is the ability of a system to recognize human emotions from speech. Typically this can be performed in two ways, textual/context analysis where whatever the speaker says is transcribed first into text then performing linguistic analysis, or analyzing the speech signal patterns itself. The latter will be the focus on this paper. The challenge of SER comes from the ability to achieve a high recognition rate while having limited resources (time, processing power) or interference (noisy background).

The recent trend of performing SER is to employ machine learning for the system to learn the speech emotion feature patterns. There are many popular machine learning models such as support vector machine (SVM) [1-3], Hidden Markov Model (HMM) [4, 5], or artificial neural networks in many forms such as convolutional neural networks (CNN) [6, 7] or recurrent neural network (RNN) [8]. A neural network can be considered deep when it uses more than a single layer. Although the methodology may vary, the general flow of SER can be viewed in Figure 1.

The speech signals used to train a neural network comes from a dataset where emotional speech spoken are recorded and labeled. Typically a clean (minimum noise) dataset is recorded in a studio with

elicited emotional speech. However there are also noisy audio datasets which imitates the condition in natural environment. To handle this, usually a preprocessing step is performed before training.

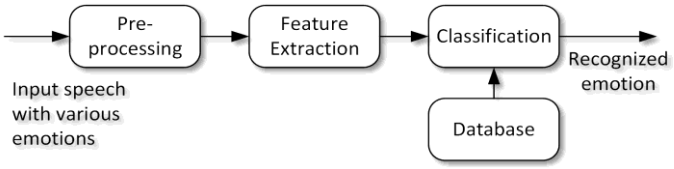


Figure 1. A typical speech emotion recognition algorithm

Voice activity detection (VAD) is one prominent method of pre-processing speech data. A VAD functions to detect presence or absence of human voice in a signal [9]. Researches on practical applications of VAD and how to implement them are numerous, such as the paper published by [10] which proposes the implementation of real-time VAD smartphone application or the algorithm proposed by [11] for assisting those with hearing problems. The personality identification research by [12] was improved by utilizing VAD.

In this research, the main usage of VAD is to improve the results of SER. Other similar research using VAD to improve SER such as [13] yielded an accuracy rate of 96.97%. This is achieved however for only 3 emotions, angry neutral and sad. Another research conducted by [14] incorporated the arousal-valence theory and obtained a recognition rate of 95% for 4 emotions, happy, fear, anger, and sad. Other research, such as [15], investigated the effects of VAD by introducing 5 different types (Voice babble, Factory noise, HF radio channel, F-16 fighter-jets, and Volvo 340) to the clean audio signals. Their proposed VAD system have shown a mean improvement of 5.54% when across the 5 noises when tested for the Berlin Emotion Database (EMO-DB).

While research of VAD in SER has already been performed, there are still room for improvement in terms of accuracy or the amount of emotions detected. Hence in this paper we contribute to the field in two ways. The first is showcasing our proposed SER system results by using VAD, benchmarking with other papers using similar methodology. The second is investigating the effects of VAD when mixing a clean emotion dataset with a noisy one for training and testing a deep neural network. The rest of the paper is as following. Section 2 briefly outlines our steps taken to perform SER. Section 3 displays the results obtained and a discussion as well as benchmarking. Section 4 concludes by summarizing the content of this paper.

2. RESEARCH METHOD

The research conducted in this paper continues our previous research in [16], but now with the addition of VAD. Our system has 4 steps to perform SER. The VAD is performed in the preprocessing stage, where the audio files are passed through Sohn's VAD algorithm to remove the silent segments. After, the MFCC of the speech signals are extracted for feature extraction. The MFCC is then used to train and test the deep neural network. For this system, we adopt 2 datasets, the German Emotion Database EMO-DB and a custom made low quality database. The overview of the system can be view in Figure 2.

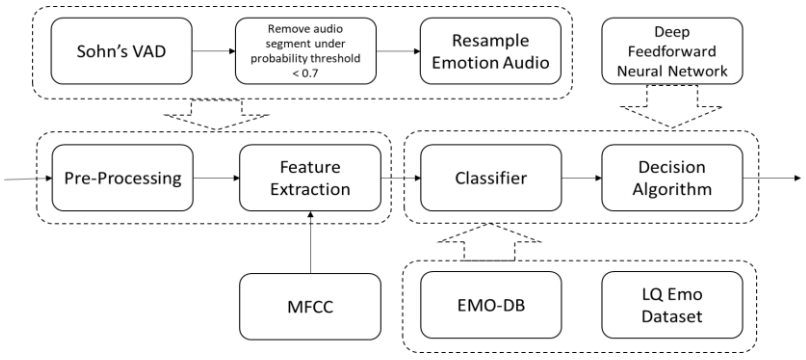


Figure 2. Proposed SER with VAD methodology

As mentioned previously, the investigation has been divided into two parts. The first is investigating the effects of VAD on a clean signal dataset, then comparing the results obtained with and without VAD recognition. The second part is testing the VAD on noisy speech input. The following subsections elaborates in more detail each step.

2.1. Voice activity detection pre-processing

A typical VAD algorithm is shown in Figure 3. For this project, the VAD is performed by using the Sohn's VAD algorithm which integrates a decision-directed parameter estimation and HMM-based hang-over scheme to improve the results [17]. This is implemented as 'vadsohn' using VoiceBox toolbox [18]. The algorithm uses a mix of The VAD output is a probability decision on a frame-by-frame basis, with a threshold probability of 0.70, as shown in Figure 4.

Each audio files are then resampled at original sampling frequency of 16 kHz concatenating only the signals with voice signal detected, to then be passed to the feature extractor. Figure 5 shows a sample of VAD pre-processing performed on a sample audio from the EMO-DB and LQ Emo Dataset.

Figure 3. A typical voice activity detection algorithm

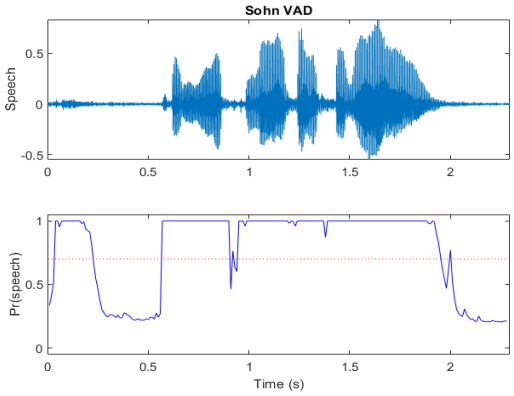
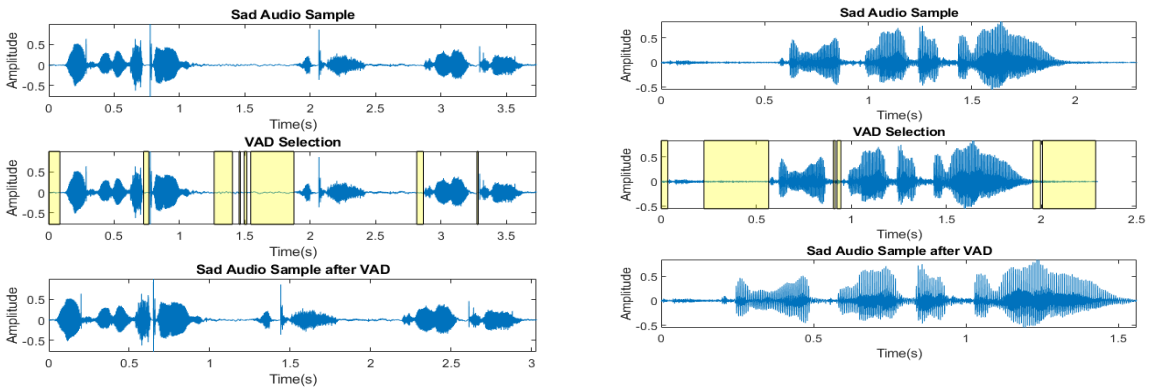


Figure 4. Sohn's VAD Algorithm



(a) (b)

Figure 5. VAD pre-processing step, (a) EMO-DB, and (b) LQ Audio Database

2.2. Mel-frequency cepstral coefficients (MFCCs) feature extraction

For the reasons outlined in [19] Mel-Frequency Cepstral Coefficients to be the sole feature for the classification. 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. The (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 (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:

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

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 (4), as follows:

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

MFCC is one of the most popular speech feature to be utilized for SER, as shown in research [1-7]. This common usage enables us to benchmark the results. For this step, the melcepst module from VOICBOX is used to extract the MFCC. We have used the default parameters, namely Hamming window in time domain with a triangular shaped filter in the mel domain.

2.3. Deep feedforward neural network classification

In this paper, we utilize the deep neural network algorithm for the system to learn and classify the emotions. MATLAB neural network pattern recognition tool is used with 70/15/15 training/validation/testing ratio. As the purpose of this paper is to investigate the effects of VAD on the recognition rate, the network size is kept at a constant rate of 2 hidden layers with 10 neuron each, using a varying number of MFCC as the features, as shown in Figure 6. For each iteration, the training/validation/testing data is randomized.

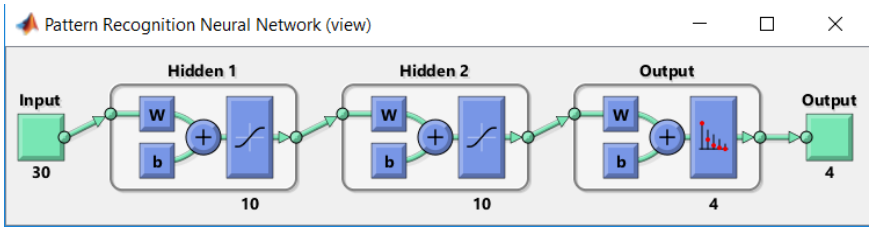


Figure 6. Deep feedforward neural network configuration

2.4. Dataset

This research uses the Berlin Emotion Database (EMO-DB) [20] as the primary dataset. The EMO-DB contains simulated emotional voices from 5 female and 5 male actors in 10 German utterances (5 short and 5 longer sentences) in 7 emotions. For this experiment, we use only 60 for each emotion so that the

training would be equal across emotions. As we are analyzing up to 5 emotions, the total voice lines used is 300 samples.

For the second part of the research, we reused the Low Quality Emotional dataset (LQ Emo Dataset) collected from [16], which contains 148 low quality emotional speech in .ogg format. The data collection was performed using a common mobile phone recorder under noisy environment, transmitted with WhatsApp audio message. WhatsApp is using the Opus codec, a lossy audio coding format, for voice media streams at either 8 kHz or 16 kHz [21]. The database contains the short utterance in 4 emotions which has the most potential applications in fields such as customer service, which is happy, angry, sad, and neutral. Each audio file contains 5 daily conversational lines in English, such as 'hello, good morning" and the average length of the files are 2.26 seconds.

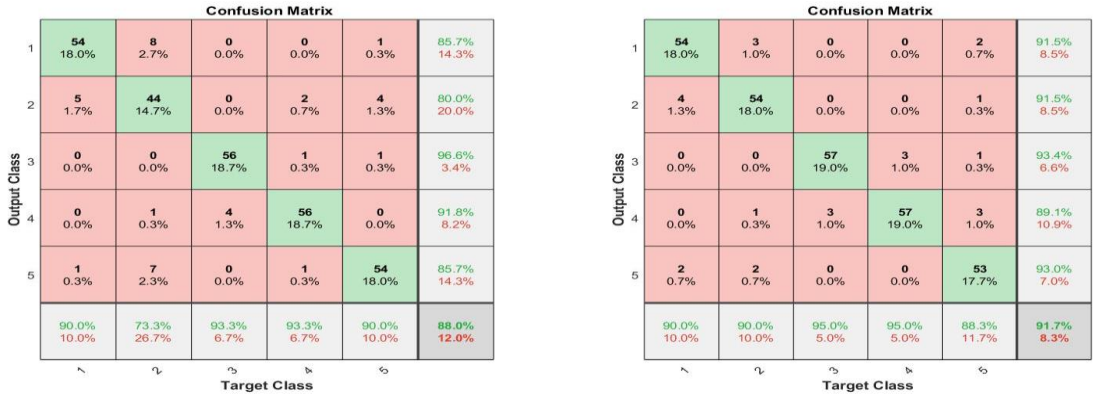
3. RESULTS AND ANALYSIS

The SER is performed using MATLAB R2018b on an Intel (R) Core i5-7200U CPU @ 2.50GHz. The number of MFCC is varied from 1-30. Each investigation is repeated 3-5 times then the average is taken when necessary to ensure accuracy.

3.1. Experiments of VAD in EMO-DB

From the methodology, we firstly experiment the SER accuracy for 5 emotions. The chosen emotions are happy, angry, sad, fear, and neutral. 300 emotional voice audio signals from the clean dataset (EMO-DB) are passed to the system for the MFCC feature extraction without pre-processing. The obtained recognition rate of emotion is 88%, as shown in Figure 7(a).

The experimental setup is then repeated but now with VAD pre-processing. All 300 files are passed through the vadsohn VAD, removing any segments without detected voice. Although the dataset is considered with minimum noise, the VAD typically trims the beginning and the end of the emotion audio. The processed data is then passed to the feature extraction and finally the classifier. The obtained recognition rate is 91.7% as shown in Figure 7(b), an improvement of 3.7%. The best results were obtained when MFCC is set to 30 coefficients.



(a) (b)

Figure 7. Emotion recognition results for clean dataset, (a) Without VAD, and (b) with VAD

Benchmarking with the other papers using similar methodology and VAD, our results were not as accurate as that obtained in [13] or [14], however considering that we have also included more emotion to detect, this recognition rate is acceptable. For the paper by [13], the amount of emotions detected is only limited to angry neutral sad, [14] detects happiness anger sadness fear, while our study recognizes up to 5 emotions, happiness angry sadness neutral and fear. As the number of emotions detected increases the complexity of the system also increases.

Another aspect to consider is the ratio of training/validation/testing. The research in [13], used a ratio of 80/10/10 compared to our 70/15/15. Theoretically when the training size is increased, so does the accuracy rate.

One last factor to consider is that [13] has split every file into 20 millisecond chunks with no overlap vectors of length 320 (16 kHz * 20ms) while our methodology takes the whole audio file and average the obtained MFCC, which limits the amount of training and testing data. Ideally for training a deep neural network, more samples are better, hence in the future study we plan to add more sample files for training.

3.2. VAD Experiment on Both EMO-DB and LQ Emo Dataset

From the second part of the investigation, we investigated the effects of utilizing VAD when mixing datasets from clean and noisy data. From our previous study [16], we found the accuracy rate was poor when the noisy and clean are mixed, yielding an accuracy rate of 23.3%, as shown in Figure 8(a).

240 lines of 4 emotions (happy, angry, sad, and neutral) from the EMO-DB are mixed with 120 noisy lines for network training and testing. A sample of the VAD process using LQ Audio Dataset can viewed in Figure 5(b). The number of MFCC is fixed to be 30, following the best results obtained in 3.1. By applying VAD to both, we achieved a great improved recognition rate of 73.6%, which not only implies that VAD can be used to process to noisy signals, but also can be used when dealing with a mix of low quality audio and clean audio, as displayed in Figure 8(b). This result is consistent with the results obtained by [15] which shows that VAD is particularly useful when dealing with noisy signals, or when mixing between both clean and noisy.

(a) (b)

Figure 8. Emotion recognition results from lq audio database, (a) Results without VAD [16], (b) Results with VAD

Another factor to consider is that the dataset in EMO-DB is in German while the LQ Audio Dataset is in English. The high improved results using VAD may be attributed to the fact that certain languages convey more information at different rates. According to [22] among 7 languages, German is one of the slowest in terms of syllables per second, while English hovers around the middle. There is also the consideration of different lingos or dialects for each speaker. By applying VAD, the unnecessary information such as gaps between words can be filtered out and the SER system can focus on the important signals only.

4. CONCLUSION

To achieve a more accurate speech emotion recognition, this study has employed Sohn's VAD algorithm in the pre-processing step. We have tested the algorithm against two datasets–a clean dataset from EMO-DB and a noisy dataset created from our previous study. From the results obtained, we have shown that by using VAD, the accuracy rate of SER has improved by 3.7% for 5 emotions when testing against a clean dataset. The second part of the study tested the usage of VAD when mixing a dataset of clean and noisy dataset. From the results obtained, we have found that the recognition rate greatly improved from our previous study, up to 50% recognition rate. Our results encourage future studies to include VAD in the pre-processing step especially when dealing with mixed datasets.

There are a few shortcoming in this research. The first is the limited amount of data for training and testing. Our methodology takes the average MFCC of a single speech sample rather than taking multiple

frames per sample, hence limiting the total number of samples. The next is the usage of the LQ Emo Dataset. The dataset's number of speech samples are less than that of the EMO-DB samples used. Ideally to train the network an equal amount of each dataset is used. Nonetheless, both limitation are addressed by repeating the experiment 3-5 times to ensure that the errors are kept at the minimum. Future studies can further expand the number of emotions detected, as the EMO-DB has 7 types of emotion recorded. For the LQ Emo Dataset, additional audio is planned to be recorded in different languages.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the Malaysian Ministry of Education (MOE), through the Fundamental Research Grant Scheme, FRGS19-076-0684 and Universiti Teknologi MARA (UiTM) Shah Alam which have provided funding and facilities for the research.

REFERENCES

[1] A. Rajasekhar and M. K. Hota, "A Study of Speech, Speaker and Emotion Recognition Using Mel Frequency Cepstrum Coefficients and Support Vector Machines," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0114-0118. 2018.

[2] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 1080-1084. 2016.

[3] A. Sonawane, M. U. Inamdar, and K. B. Bhangale, "Sound based human emotion recognition using MFCC & multiple SVM," in *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, pp. 1-4. 2017.

[4] J. Watada and Hanayuki, "Speech Recognition in a Multi-speaker Environment by Using Hidden Markov Model and Mel-frequency Approach," in *2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN)*, pp. 80-83. 2016.

[5] Chandni, G. Vyas, M. K. Dutta, K. Riha, and J. Prinosil, "An automatic emotion recognizer using MFCCs and Hidden Markov Models," in *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 320-324. 2015.

[6] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, pp. 333-336. 2017.

[7] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 26-29. 2017.

[8] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 2227-2231.

[9] W. Q. Ong and A. W. C. Tan, "Robust voice activity detection using gammatone filtering and entropy," in *2016 International Conference on Robotics, Automation and Sciences (ICORAS)*, 2016, pp. 1-5. 2016.

[10] A. Sehgal and N. Kehtarnavaz, "A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection," in *IEEE Access*, vol. 6, pp. 9017-9026, 2018.

[11] J. Song et al., "Research on Digital Hearing Aid Speech Enhancement Algorithm," in *2018 37th Chinese Control Conference (CCC)*, pp. 4316-4320. 2018.

[12] K. Gokul and S. Lalitha, "Personality Identification Using Auditory Nerve Modelling of Human Speech," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1731-1737. 2018.

[13] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, 2017, pp. 137-140.

[14] P. A. Bustamante, N. M. Lopez Celani, M. E. Perez and O. L. Quintero Montoya, "Recognition and regionalization of emotions in the arousal-valence plane," *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, 2015, pp. 6042-6045.

[15] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An Unsupervised frame Selection Technique for Robust Emotion Recognition in Noisy Speech," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2055-2059. 2018.

[16] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, "Speech Emotion Recognition Using Deep Feedforward Neural Network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, 2018.

[17] Jongseo Sohn, Nam Soo Kim and Wonyong Sung, "A statistical model-based voice activity detection," in *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[18] D. Brookes. (2010, 14/2/2019). VOICEBOX: A speech processing toolbox for MATLAB. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[19] T. S. Gunawan, M. F. Alghifari, M. A. Morshidi, and M. Kartiwi, "A Review on Emotion Recognition Algorithms using Speech Analysis," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 6, no. 1, pp. 12-20, 2018.
[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.
[21] F. Karpisek, I. Baggili, and F. Breitinger, "WhatsApp network forensics: Decrypting and understanding the WhatsApp call signaling messages," Digital Investigation, vol. 15, pp. 110-118, 2015.
[22] F. Pellegrino, C. Coupé, and E. Marsico, "Across-language perspective on speech information rate," Language, vol. 87, no. 3, pp. 539-558, 2011.

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 and is currently pursuing his Masters in Computer Engineering while working as a research assistant. His research interests are in signal processing, artificial intelligence and affective computing. He received a best FYP award from IEEE Signal Processing–Malaysia chapter and achieved recognition in several national level competitions such as Alliance Bank EcoBiz Competition and IMDC2018.

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 (2017-2018), International Islamic University Malaysia. He is Chartered Engineer (IET, UK) and Insinyur Profesional Madya (PII, Indonesia) since 2016, and registered ASEAN engineer since 2018.

Mimi Aminah binti Wan Nordin received her B.Eng and Masters at International Islamic University Malaysia and completed her PhD at National University of Malaysia in 2014. She also obtained her MBA at Asia School of Business in 2018. Mimi has started several startups and has experience working on sensor-based solutions. Currently, she is the deputy director for innovation and commercialization unit at research management centre, IIUM, since 2019.

Syed Asif Ahmad Qadri completed his Bachelor of Technology in Computer Science and Engineering, from Baba Ghulam Shah Badshah University, Kashmir. Currently, he is working as a Research Assistant with the Department of Elec. & Comp. Engineering in International Islamic University Malaysia IIUM. His research interests include Speech processing, Artificial Intelligence, Information security, IoT etc. Currently, he is working on the application of speech emotion recognition using Deep Neural Networks.

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, Kulliyyah of Information and Communication Technology, and Deputy Director of e-learning at Centre for Professional Development, International Islamic University Malaysia. Her research interests include electronic commerce, data mining, e-health and mobile applications development.

Zuriati Janin received her B.Eng in Electrical Engineering from the Universiti Teknologi Mara, Malaysia in 1996 and MSc. in Remote Sensing & GIS from the Universiti Putra Malaysia (UPM) in 2001. In 2007, she began her study towards a Ph.D in Instrumentation and Control System at the Universiti Teknologi Mara, Malaysia. She has served as a lecturer at Universiti Teknologi Mara for more than 20 years and currently she is a Senior Lecturer at Faculty of Electrical Engineering, UiTM, Shah Alam. She has been involved with IEEE since 2012 and been mainly working with IEEE Instrumentation & Measurement Chapter (IM09), Malaysia Section since 2013. The IM09 acknowledged her role as a founder Treasurer in initiating and promoting ICSIMA as a series of annual chapter's flagship conferences since its inception in 2013. She also has more than 10 years experiences in organizing the International Conferences, Workshops and Seminars. Her role as a conference treasurer started since 2005.