Smart Emotion Recognition Framework: A Secured IoVT Perspective

Pavan D. Paikrao
TPCT College of Engineering

Amrit Mukherjee
Anhui University

Deepak Kumar Jain
Chongqing University of Posts and Telecommunications

Pushpita Chatterjee
Ton Duc Thang University

Waleed Alnumay
King Saud University

***Abstract*—The promise of automated-driving cars causes the automotive and consumer electronics (CE) sector to rethink not only what it means to drive, but also the relationship between the car and the consumer. Recent trend in Internet of Vehicle Things (IoVT) promotes robust interactions between humans and vehicles, which ultimately points to enhance human abilities, such as hearing, visual surveillance, or emotion awareness, as a part of safety concern. The voice-based interactions (speech recognition and stress monitoring) will improve in-time awareness of the vehicle status. Unfortunately, the existing modulation domain speech enhancement techniques achieve low satisfactory performance in detecting humans' stress emotions where the environmental noise is inevitable and varies with the location of every passing vehicle. Furthermore, the computational load introduces challenges in their implementation in automated vehicles. In this direction, we propose a**

Digital Object Identifier 10.1109/MCE.2021.3062802
Date of publication 1 March 2021; date of current version 6 December 2022.

2162-2248 © 2021 IEEE

Published by the IEEE Consumer Technology Society

front-end processing framework, in particular to stress emotion detection cases (such as anger, sad, fear, and happy) in different nonstationary noisy environments, such as car, airport, traffic, and train. This article encompasses three interrelated issues: 1) analysis, modification, and synthesis of noisy speech emotion in modulation domain in real-time background noise, 2) extracting set of Mel-frequency cepstral coefficients features from noisy speech stimuli for speech emotion recognition, and 3) evaluation of overall system performance by means of objective parameters, and confusion matrix in adverse environments using speech emotion database Interactive Emotional Dyadic Motion Capture. The experimental results show that favorable performance in state-of-the-art stress monitoring yields high levels of consumer satisfaction for security in vehicle comparison to traditional frameworks.

The automatic speech emotion recognition is a very recent topic of research in the human–machine interaction field.

It has a wide range of applications such as in-car devices where mental state information of the car driver is used to initiate their safety.[1,2] In automatic remote call center, it is used to timely detect customers dissatisfaction. With the rapid increase in the use of smart speech technologies, the human beings can communicate with the machines, which is used in automatic cars, education, military, and medical applications .

Recent trend in Internet of Vehicle Things (IoVT) promotes robust interactions in between humans and vehicles, which ultimately points to enhance human abilities, such as hearing, visual surveillance, or emotion awareness, as a part of safety concern.[3]

Figure 1 shows the functional block architecture of the IoV system and the speech server. The devices within client environment connect with speech servers, IoT servers, and IoV servers with various network layers and protocols.

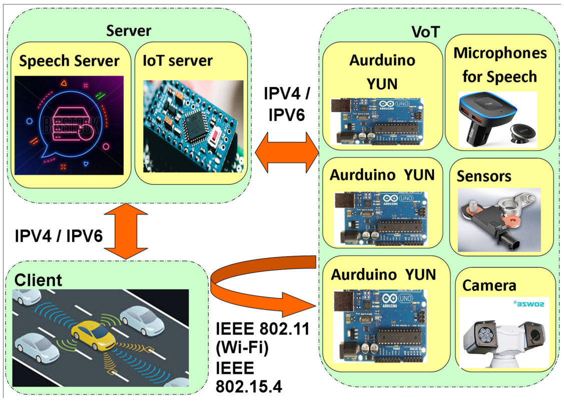

Figure 1. Functional block architecture of IOV system and speech servers.

The main contribution of this article is to highlight the vital potential of speech methodologies for improving the state of the art in IoVT, which are listed as follows.

1) *Preprocessing Methodology*: To monitor the human's emotion, preprocessing of real-time voice captured by speech sensors is essential. However, the noisy environmental conditions hinder the IoVT system from this task. Therefore, we propose a method that incorporates various noise estimation techniques.
2) *To reduce complexity*: The computational load introduces challenges in the implementation of preprocessing methods; thus, we need quick response in automated vehicles. Therefore, the solution to conventional methods is provided to reduce the computational load along with a set of parameters.
3) *To ensure security and quality of service (QoS)*: In this article, we study the benefits of the proposed optimized modulation spectral subtraction (OMSS) speech enhancement for a speech recognition system, in particular, the influence of different nonstationary noises (background environment such as airport, car, train, and traffic) on typical speech emotion recognition.
4) *Validation*: In order to find the performance of speech emotion recognition on the proposed OMSS method, results were compared to the traditional ModSpecSub method.[4] The speech emotion database Interactive Emotional Dyadic Motion Capture (IEMOCAP) with various SNRs is applied for the experimental

verification. The maximum accuracy of speech emotion recognition is found to be 98.5% observed for two case studies of anger emotion in airport noise and traffic noise.

The results convey that the modulation domain enhancement algorithm, when employed as preprocessing stage, dramatically enhances recognition of an emotion recognition system in the presence of varying noise environmental conditions and inputs SNRs for driver's safety in IoVT applications.

PREPROCESSING METHODOLOGY: ANALYSIS–MODIFICATION–SYNTHESIS (AMS)

To present the application of the modulation and phase spectra in automatic speech recognition (ASR) system, we employ the AMS method for speech enhancement as preprocessing. Consider additive noise scenario as in

$$x(n) = s(n) + N(n) \tag{1}$$

where $x(n)$ is the discrete-time noisy voice signal, $s(n)$ is the discrete-time clean voice, and discrete-time background noise is indicated by $N(n)$. The discrete time index is represented as "*n*." In our whole discussion, when referring to magnitude or phase spectrum in an acoustic AMS framework, the discrete-time short-time Fourier transform (STFT) with small window typically 20–40 ms is employed until and unless stated

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)W(n-Ql) \times e^{\frac{-j\times 2\times \pi \times k\times l}{M}} \tag{2}$$

where *l* refers to an acoustic frame number, $x(l)$ is a discrete acoustic sample, $W(n)$ is an acoustic analysis window function, *k* refers to discrete frequency, and *M* refers to frame duration in samples. *Q* denotes frame shift. The maximum frame shift of 40% (about 8-ms duration) can be applied. Modified Hamming window is applied at both acoustic processing and modulation processing, since it has reported improved performance as to other windows. Typically, an *M*is chosen between 20 and 40 ms and is applied in speech processing. We apply spectral subtraction on a modulation domain spectrum, as shown in Figure 2. Thus, by applying STFT in (1), $X(n,k)$, $S(n,k)$, and $N(n,k)$ are referred as a discrete-time STFT spectrum of noisy speech, pure voice, and background noise, respectively.

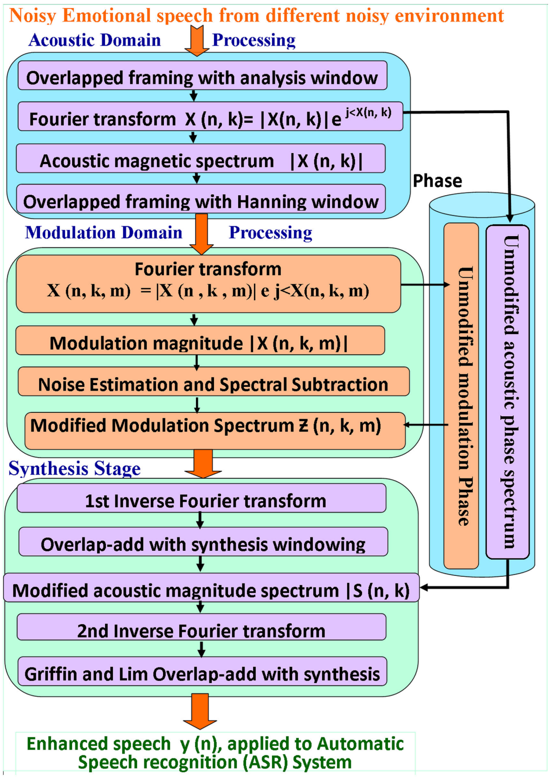
Noisy Emotional speech from different noisy environment
Acoustic Domain
Processing
Overlapped framing with analysis window
Fourier transform X (n, k)= |X(n, k)| e j<X(n, k)
Acoustic magnetic spectrum |X (n, k)|
Overlapped framing with Hanning window
Phase
Modulation Domain
Processing
Fourier transform
X (n, k, m) = |X (n , k , m)| e j<X(n, k, m)
Modulation magnitude |X (n, k, m)|
Noise Estimation and Spectral Subtraction
Modified Modulation Spectrum Ƶ (n, k, m)
Unmodified modulation Phase
Unmodified acoustic phase spectrum
Synthesis Stage
1st Inverse Fourier transform
Overlap-add with synthesis windowing
Modified acoustic magnitude spectrum |S (n, k)
2nd Inverse Fourier transform
Griffin and Lim Overlap-add with synthesis
Enhanced speech y (n), applied to Automatic Speech recognition (ASR) System

Figure 2. Preprocessing flowchart of a proposed optimized modulation spectral subtraction for IoVT.

AMS-BASED MODULATION ANALYSIS FOR IoVT

The modulation domain enhancement is achieved by extending acoustic AMS procedure discussed in the section "Preprocessing Methodology: Analysis–Modification–Synthesis (AMS)." In this processing, the discrete-time STFT is computed with the respective acoustic frequency using time series of magnitudes $|X_R(n,k)|$ at that frequency. In modulation domain processing, the Hamming window with 128- and 16-ms frame shift is used.

Algorithm 1 shows the flowchart of the proposed OMSS method. *m* is termed as an index of the discrete modulation frequency. In this section, we discuss the concurrent effect of various noise estimation methods with the proposed method. One way is to employ a Voice Activity Detector (VAD), but this will result in increase in the computational complexity. It is observed that at a large frame duration of modulation frame shift, no considerable effect of noise updating is found. Therefore, we avoid VAD for noise estimation and use minimum statistic noise estimation in order to reduce the computational load on the conventional ModSpecSub.[4]

Algorithm 1. Preprocessing of Noisy Speech Emotion

```
Input:X= $(X_1, X_2, \ldots, X_n,)$
Output:Y= $(Y_1, Y_2, \ldots, Y_n,)$
Initialization:
 1: Collect noisy speech from environment
    LOOP Process
 2: Initialize framing noisy emotion speech Xi (i=1, 2,...., n) in discrete frames using window $W(n)$
 3: Compute the STFT of overlapped frames of 32-ms duration each with 40% overlapping
 4: Extract magnitude spectrum $|X(n,k)|$
 5: Discard the noisy Phase $< X(n,k)$
 6: for $m <$ number of modulation frames with frame duration of 180 ms in modulation domain do
 7:     Repeat steps 2 to 5 to calculate $|X(n,k,m)|$
 8: end for
 9: Compute the noise estimate with noise floor $\beta$
10: Calculate the spectral subtraction estimate of clean spectrum to reduce the real-time noise
11: Combine the unmodified phase in synthesis stage
12: Compute the discrete time inverse Fourier transform
13: Apply overlap–add with synthesis windowing
14: STOP
```

Following steps are involved in the proposed OMSS approach, as shown in Figure 2.

Modification step 1: We use noisy speech signal (no mean subtracted) and, then, segmented into overlapping frames by a window of duration (speech frame length) of 32 ms and 8-kHz sampling rate with an acoustic frame increment of 8 ms. By applying the discrete STFT to every frame in a sequence results in a complex acoustic spectrum *X* (*n, k*) represented as given in

$$X(n,k) = X_R(n,k) + iX_I(n,k) \tag{3}$$

where the real part of a complex STFT acoustic spectrum is $X_R(n,k)$. $X_I(n,k)$ is referred as an imaginary part of discrete-time complex STFT $X(n,k)$. In this step, only real part of a discrete-time complex STFT acoustic spectrum $(X_R(n,k))$ is taken into consideration (by rejecting an imaginary spectrum). We named this $|X_R(n,k)|$ as real acoustic magnitude spectrum (RAMS). Mod $|.|$ represents absolute magnitude value and, simultaneously, phase is extracted from this RAMS. The extracted unmodified phase will be reused for synthesis of signals, as shown in Figure 2.

Modification step 2: The RAMS is applied for further processing. The noisy RAMS $|X_R(n,k)|$ is subdivided, which results in overlapped modulation frames using the modulation frame duration of 128 ms. After that, second Fourier transform is employed on time axis (on frequency), which results in a complex STFT spectrum $X(n,k,m)$. In our further discussion, we will define it as real modulation magnitude spectrum (RMMS) $|X_R(n,k,m)|$ by leaving an imaginary spectrum. Modulation domain noisy phase $< X(n,k,m)$ is estimated from the RMMS. This unmodified phase will be used later for synthesis.

In modulation domain, long frame duration results in speech musical noise or temporal slurring. Simultaneously, the longer frame increases the computational load on the system. Optimal modulation frame duration was decided to 128 ms, and the frame shift of 16 ms is employed so as to minimize the temporal speech slurring and the computational load.

Modification step 3: In the conventional speech enhancement, the input noisy speech signal is used to find noise estimate. But in contrast to the conventional step, we applied RMMS to find noise estimation. It means that, in the proposed approach, noise estimation from RMMS for the spectral subtraction in modulation domain is employed. Figure 3 shows how the different noise estimation methods, such as minimum statistics noise estimation[5] and unbiased MMSE noise estimation, can be combined with the proposed OMSS method.

Parameters used: α is referred as an oversubtraction factor, which is conventionally applied between 0 and 6. γ is a spectral subtraction domain. When γ = 1 and γ = 2, it is referred as magnitude and power spectral subtraction, respectively. In case of the minimum statistics method for estimation of noise, α may be 0–3. For $\alpha = 1$, improved performance of a system is reported. In experimentation for various noise estimation methods, unbiased minimum mean-square noise estimator, when $\alpha = 1$ is used, it reported reduced objective scores.

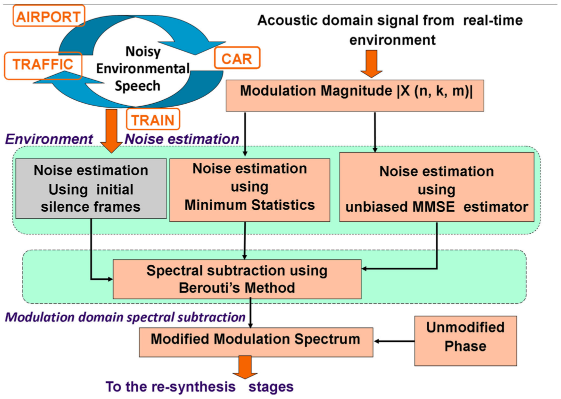
AIRPORT
TRAFFIC
CAR
TRAIN
Noisy Environmental Speech
Acoustic domain signal from real-time environment
Modulation Magnitude |X (n, k, m)|
Environment
Noise estimation
Noise estimation Using initial silence frames
Noise estimation using Minimum Statistics
Noise estimation using unbiased MMSE estimator
Spectral subtraction using Berouti's Method
Modulation domain spectral subtraction
Modified Modulation Spectrum
Unmodified Phase
To the re-synthesis stages

Figure 3. Real-time noise environment in modulation domain for IoVT.

We use $0.1 \leq \alpha \leq 3$. In modulation domain enhancement, objective performance scores are found to be relatively improved at $\gamma = 2, \alpha = 1$.

Noise Estimation:

Conventionally using input noisy speech and initial silence frames: The noise estimate is computed by initial silence frames and is updated by recursive averaging.[4] Traditional ModSpecSub method[4] uses the VAD method to update estimate of noise. But VAD noise estimate for spectral subtraction results in reduced perceptual evaluation of speech quality (PESQ) scores for nonstationary environment with increased computational load due to continuous noise updating.

Minimum statistics noise estimate: The power spectral density (PSD) of additive white Gaussian noise (AWGN) nonstationary environment is approximated by a minimum statistics [5] noise estimation method from noisy speech.

Reason for using minimum statistic noise PSD estimate: The conventional ModSpecSub method[4] employs VAD for noise update in-between nonvoice frames, whereas PSD of noise is derived by tracking spectral minima over each frame independently from both voice and nonvoice frames, avoiding VAD in case of the method by Martin.[5] Thus, the computational load is reduced in the proposed method.

HUMAN EMOTION RECOGNITION METHODOLOGY FROM SPEECH

Feature Extraction for Speech Emotional: Conventionally, the hidden Markov model (HMM) for training purpose employs the temporal features. Similarly, the Mel frequency cepstral coefficient (MFCC) uses statistical and delta-deltas features. This method uses initial 12 MFCCs, which are derived using time shift of 9 ms from the modified prepossessed and enhanced speech signal.

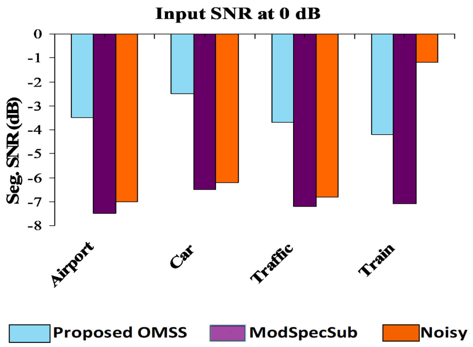

Figure 4. Segmental SNR improvement for anger emotion at 0 dB in different noise environments.

The Gaussian mixture model (GMM) has been implemented successfully in voice, speaker, and various regional communication language discriminations.

EXPERIMENTAL EVALUATION

Database Used, Stimuli Generation, and Performance

In our experimental evaluations, IEMOCAP speech corpus[6] is employed. The corpus includes audio stimuli emotions, such as sad, anger, happy, and neutral.

The voice stimuli with various states of emotions, such as sad, anger, happy, and natural, are created. Now, every voice stimulus is added with the environmental noise, such as car, airport, traffic, and train background noise along with various SNR ranging between 0 and 15 dB. As per discussion in the section "AMS-Based Modulation Analysis for IoVT," stimuli are used for the proposed OMSS method and Paliwal's ModSpecSub.[4]

Improved SNR seg. as a function of input SNR at different noise environments is shown in Figure 4.

Table 1 shows PESQ an objective score for happy emotion in different noise environments. PESQ score is improved from 0.26 to 0.68 on mean compared to ModSpecSub and noisy speech stimuli. But we found significant improvement for traffic noise of 0.84, i.e., approximately 33.13%.

Table 1. PESQ Happy Emotion in Different Environments.

		PESQ score (emotion happy)		
Noise Type	Input SNR (dB)	OMSS	Paliwal[4]	Noisy
Airport	0	2.237	1.975	1.122
	5	2.543	2.305	1.905
Car	0	3.318	3.732	2.732
	5	3.381	4.070	2.994
Traffic	0	2.179	2.043	2.012
	5	2.501	2.342	2.543
Train	0	2.334	2.302	2.543
	5	2.613	2.580	2.118

Speech Emotion Recognition Classification Results

The speech stimuli generated in the section “Database Used, Stimuli Generation, and performance” are used for a speech emotion recognition classification.

Classifier training and testing: In the experiment presented in this article, GMM classifier is first trained and tested five times. The results of human stress of various emotion types with the noise environment type are shown in Table 2. The performance of the proposed OMSS is compared with conventional Paliwal’s[4] method and noisy unprocessed speech at input SNR of 0 and 5 dB for various noise types (car, airport, traffic, and train) and speech emotion (neutral, anger, joy, sad, and fear).

Case 1—*Speech corrupted by airport noise:* The confusion matrix for airport noise case is presented in Table 2. The best recognized emotion for the proposed OMSS in this case is anger (98.5%), which is the mostly confused with happy speech. The worst recognized emotion is neutral with rate of 26.5%. Fear is the most often confused with sad and joy emotions of noisy speech stimuli.

Case 2—*Speech corrupted by car noise:* It is clear that speech recognition rates were degraded in noisy environment for different SNR levels. We have found that speech recognition rates have been improved in noisy conditions under anger

Table 2. Confusion Matrix Results for Different Methods in Case 1 Airport Noise.

		Emotion Recognized in %				
Type of stimuli		Neutral	Anger	Joy	Sad	Fear
Neutral	Noisy	11.5	39.0	49.5	0	0
	OMSS	26.5	23.5	25	20	5
	Paliwal’s	10.5	14.2	75.3	0	0
Anger	Noisy	4.5	50.5	22.5	11.5	11
	OMSS	0	98.5	1.5	0	0
	Paliwal’s	0	24.5	75.5	0	0
Joy	Noisy	16.5	25.4	43.6	14.5	0
	OMSS	0	26.5	73.5	0	0
	Paliwal’s	0	5.5	94.5	0	0
Sad	Noisy	6.3	0	42.5	7.5	43.7
	OMSS	26.5	0	0	42.5	31.0
	Paliwal’s	0	10.5	71.5	18.5	0
Fear	Noisy	0	0	56.5	37.0	6.5
	OMSS	6.0	0	0	45.5	48.5
	Paliwal’s	8.5	0	0	49	42.5

emotion. In Cases 3 and 4, we have analyzed the speech corrupted by traffic noise and train noise, respectively, which also provides enhanced recognition rates compared to the conventional one. We found that highest performance of ASR is reported in anger noise about 98.5% followed by conventional Paliwal's[4] method 88.5%.

CONCLUSIONS

Considering the various aspects of human stress detection and monitoring methodology in our daily life, this article describes the significance into CE aspects of secure technology with various key points of real-time environment. This article will be an essential step toward the development of a ubiquitous human stress monitoring system in the edge of secure IoV cloud computing. In real-world applications, the influence of noise is a major constrain in many of the automatic speech recognition. The performance of the proposed OMSS method is evaluated using objective evaluation parameters such as SNR seg. and PESQ. For the speech emotion type, anger and happy (with different noise type and input SNR) constructed by treatment type of the proposed scheme greatly improved SNR seg. as compared with the traditional ModSpecSub method. The best recognized emotion for the proposed OMSS is anger (98.5%), which is the mostly confused with speech emotion happy in conventional methods.

ACKNOWLEDGMENTS

This work was supported by King Saud University, Riyadh, Saudi Arabia, under Researchers' Supporting Project RSP-2020/250.

REFERENCES

1. A. A. Penilla and A. S. Penilla, "Methods and vehicles for capturing emotion of a human driver and customizing vehicle response," U.S. Patent App. 16/732,069, May 14, 2020.
2. T. Alladi, V. Chamola, B. Sikdar, and K.-K. R. Choo, "Consumer IoT: Security vulnerability case studies and solutions," *IEEE Consum. Electron. Mag.*, vol. 9, no. 2, pp. 17–25, Mar. 2020, doi: 10.1109/MCE.2019.2953740.
3. A. Kouris, S. I. Venieris, M. Rizakis, and C.-S. Bouganis, "Approximate LSTMs for time-constrained inference: Enabling fast reaction in self-driving cars," *IEEE Consum. Electron. Mag.*, vol. 9, no. 4, pp. 11–26, Jul. 2020, doi: 10.1109/MCE.2020.2969195.
4. K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, 2010, doi: 10.1016/j.specom.2010.02.004.
5. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001, doi: 10.1109/89.928915.
6. C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

Pavan D. Paikrao is currently an Associate Professor with the Department of Electronics Telecommunication, TPCT's College of Engineering, Osmanabad, India. He received the M.Tech. and Ph.D. degrees from Dr. Babasaheb Ambedkar Technological University, Lonere, India, in 2011 and 2019, respectively. Contact him at pavanpaikrao@coeosmanabad.ac.in.

Amrit Mukherjee is currently with the School of Electronics and Information Engineering, Anhui University, Hefei, China. He received the Ph.D. degree from KIIT University, Bhubaneswar, India, in 2017. He is a Member of IEEE. He is the corresponding author of this article. Contact him at amrit1460@ieee.org.

Deepak Kumar Jain received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China. He is currently an Assistant Professor with the University of Posts Telecommunications, Chongqing, China. Contact him at dkj@ieee.org.

Pushpita Chatterjee received the Ph.D. degree from IIT Kharagpur, Kharagpur, India, in 2013. She is currently with Future Networking Research Group, and with the Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. She was a Research Consultant with Old Dominion University, Norfolk, VA, USA. Contact her at puspitachatterjee@tdtu.edu.vn.

Waleed Alnumay is currently an Associate Professor with King Saud University, Riyadh, Saudi Arabia. Contact him at wnumay@ksu.edu.sa.