Data Analytics on Eco-Conditional Factors Affecting Speech Recognition Rate of Modern Interaction Systems

A. C. Kaladevi[1,*], R. Saravanakumar[2], K. Veena[3], V. Muthukumaran[4], N. Thillaiarasu[5] and S. Satheesh Kumar[6]

[1]*Department of Computer Science and Engineering, Sona College of Technology, Salem, India*
[2]*Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore 560082, India*
[3]*Department of Computer Science, J.K.K.Nataraja College of Arts & Science, Komarapalayam, Namakkal Dt., India*
[4]*Department of Mathematics, School of Applied Sciences, REVA University, Bengaluru, Karnataka*
[5]*School of Computing and Information Technology, REVA University, Bengaluru, India*
[6]*Department of Computer Science, School of Applied Sciences, REVA University, Bengaluru*
E-mail: kaladeviac@sonatech.ac.in; saravanakumar.rsk28@gmail.com; veenabharathi44@gmail.com; muthu.v2404@gmail.com; thillai888@gmail.com; satheesh.sampath@gmail.com
[*]*Corresponding Author*

Received 25 October 2021; Accepted 02 January 2022;
Publication 05 March 2022

Abstract

Speech-based Interaction systems contribute to the growing class of contemporary interactive techniques (Human-Computer Interactive system), which have emerged quickly in the last few years. Versatility, multi-channel synchronization, sensitivity, and timing are all notable characteristics of speech

Journal of Mobile Multimedia, Vol. 18_4, 1153–1176.
doi: 10.13052/jmm1550-4646.1849

recognition. In addition, several variables influence the precision of voice interaction recognition. However, few researchers have done a significant study on the five eco-condition variables that tend to affect speech recognition rate (SRR): ambient noise, human noise, utterance speed, and frequency. The principal strategic goal of this research is to analyze the influence of the four variables mentioned earlier on SRR, and it includes many stages of experimentation on mixed noise speech data. The sparse representation-based analyzing technique is utilized to analyze the effects. Speech recognition is not noticeably affected by a person's usual speaking pace. As a result, high-frequency voice signals are more easily recognized ($\sim$98.12%) than low-frequency speech signals in noisy environments. By performing the experiments, the test results may help design the distributive controlling and commanding systems.

Keywords: Interaction system, eco-conditional factors, recognition rate, ambient noise, human noise, utterance speed, frequency.

1 Introduction

Speech recognition (SR) is a multidisciplinary field that mainly includes digital signal processing as well as phonetics [35]. Speech perception assessments were originally developed to measure speech comprehension ability with regulated and manipulated external sounds while skilled conversationalists deliver specifically chosen word lists, such that interaction devices (transponders, microphones, handsets) can be evaluated on the premise of the intelligible of the conveyed message. Sound masking and the resultant concealed cutoff rely on the structural properties of the noise/sound (spectra, levels, period etc.), the listener's sensitivity for hearing loss (that unless existent), the attenuate properties of the listening defender (if employed), as well as how noise is interpreted in the eardrum and central nervous system. However, not many of such variables can be regulated, but it's essential to know the way they interact to apprehend why a variation among them impacts sound/noise audibility. Neuroscience, linguistics, cognitive science, computer science, pattern classification, and machine learning are intimately linked. Human-Computer Interaction (HCI) includes speech engagement, which seems to be the practice of utilizing regular human discourse to provide directions to a machine to gain dynamic control [42]. Thus, it became one among the newest interactive techniques that have exploded in popularity over the years.

In the mid-1950s, scientists started working on improving voice recognition technology. Intuitively, it ought to be harder to identify an objective signal in noisy environments when the ambient noise is stronger or louder. Various manipulations resulted in a lower signal-to-noise ratio (SNR), thus, making it complicated to recognize an objective signal among the varying noise levels.

In speech and noise assessment, the communicator and ambient noise types, as well as the linguistic that has been used, are examined wherein the language and ambient noise are separated to improve the quality of the speech and to interpret its content (recognition). A variety of applications, such as speech and noise categorization [7, 9] source segregation [39], speech synthesis and recognition [26, 31, 32, 44], acoustic coding [19, 28], prosody adjustment [24], and speaker normalization [26], have been investigated in the scientific sector for addressing noisy signal data. Many such applications fall within the scope called "machine listening", which seeks to decipher the content of noisy acoustic features and comprehends in a way that is similar to how humans do it themselves [34]. For example, voice alteration may be accomplished by adjusting speech characteristics like prosody (duration and pitch period), while Text-To-Speech (TTS) synthesizing systems [41] produce speech for automatic assistive technologies.

HCI uses voice-based conversation via auditory channels because it is both remarkably effective and natural. The following features of voice communication are flexibility, cross-channel synergy, transience and sensitivity coexistence, and prone to interference.

Flexibility, where speech has a greater degree of adaptability than other forms, requires less space and is not as constrained by lighting factors. It's critical when visual channels aren't the best way to get a message across.

In Cross-channel synergy, conventional interactive devices like mouse and keyboard don't impact voice interaction and may be utilized with conventional input tools to accomplish jobs faster and better. In transience and sensitivity coexistence, humans are highly receptive to audio data (sound) and have become a popular method for the reminder and alert method, but the phenomenon of speech signals also utilizes the short cognitive capacity of the operator, raising the pressure of remembering. In prone to interference, interacting via sound waves using speech which usually follows non-contact consequences of actions for transmitting command information. As a result, noise may readily disrupt voice engagement, which severely restricts the scope of applications for such a non-contact auditory communication.

M-(Speech recognition (SR) is the foundation for voice command, and such voice management is the base deployment to the SR technology in the area of control. Technology for voice recognition influences software for controlling one's speech. SR technology has gone through various stages of development from its inception in 1952, which was made possible by computer technology advancements in the early 1960s. And also, the associated software has reached a sophisticated stage by mid of the 1990s. Due to the emergence of interactive media, a realistic SR system must be developed from the beginning. Google's Speech-based Action System, launched in 2010, supports users to perform and manage voice-oriented command operations. Similarly, at the beginning of 2011, Microsoft's new deep neural network (DNN) framework gained success in SR recognition activities [19, 24, 26].

Before considering the impact on speech perception, eco-conditional factors like ambient noise, utterance speed, human noise, and frequency affecting SR are taken into account. In other words, listeners' progress will be influenced by the impacts of the affecting factors. Thus, in this research, some basic data science paradigms are utilized to analyze such impactful factors effectiveness on SR rates via test data.

In this article, we examined the issue of utterance recognition using sparse representations and the effect of environmental variables on noisy speech. Our work further assists in the research work of classifying speakers and noise types. Moreover, the findings of the experiments may be used to assist the development of a commanding and controlling system.

2 Background

The study does a literature review on noisy SR methods, examining over 100 research articles. A speech-specific system will admit data signals with speech characteristics and avoid unstructured perturbing data (noise), making it noise-resilient. To ensure the efficiency of the voice/speech approximation, the Augmented Reality (AR) all-pole framework has been utilized under the assumption that the sound signal has a specific spectral-based structure. From [22, 30] the iterative method implements optimum probability estimation to get parameters. Using Wiener filtering, this method calculates the parameters of the model from the assessment of the clear speech. In LPC-based parameterization method enhanced the prediction degree to increase the recognition rate [16].

In constrained MAP estimation of speech [36], some speech specific constraints such as the stability of an all-pole speech model, the relative position

of poles, and the inertia of the vocal tract characteristics are exploited during the iterative estimation. The resulting speech has more reliable formant positions and reduced frame-to-frame pole jitter in noisy environments.

As an improvement of the iterative procedure in [36], HMMs classifiers are used to partition speech frames into broad phonetic classes and the classification result controls the termination of the iterative estimation [27]. Suitable termination reduces peak distortion and improves objective speech quality. Using the all-pole clean speech spectral model and short-time noise level, Itakura distortion [17] between noisy spectrum and the sum of the all-pole and noisy spectra is iteratively minimized. The estimated clean speech model parameters are then fed to a speech recognizer [43]. Within these iteration procedures, each iteration produces filtered speech with lower residual noise but larger spectral distortion [21], which limits the number of iteration. The formant narrowing and shifting introduced by the iteration [21, 36] may be harmful to recognizers [33]. Another problem is that most all-pole estimates degrade rapidly as the SNR decreases [11].

For the purpose of improving noisy speech intelligibility, it was proposed in [11] to resynthesize speech from VQ templates. Formant distance was used as the similarity measure. The output of such processing is noise-free speech, with the remaining degradation appearing as a spectrum mismatch. Sparsity based speaker identification using discriminative dictionary learning was done by [10] while non-negative matrix factorization for feature extraction was explored by [8]. Representation of audio signals as a sparse, linear combination of non-negative vectors called as dictionary atoms has been used for audio source separation [2, 13, 39], recognition [5, 15], classification [37, 45] and coding [19, 28].

As a preprocessing step for recognition, speech enhancement techniques are intended to recover either the waveform or the parameter vectors of the clean speech embedded in noise [22, 38, 43] These techniques make different uses of a priori information about the speech and the noise. The criteria used in speech enhancement techniques can be based either on the probability of clean speech or the distortion between clean and recovered speech, and usually not directly related to speech recognition accuracy. Table 1 depicts the existing noisy speech recognition techniques.

3 Selection of Speech Segment

This section focuses on the dictionaries built from the data gathered by analyzing the noise and clear speech recordings. Table 2 displays around

Table 1 Noisy speech recognition techniques

Existing Techniques	Speakers	Accuracy (%)
LC-MBCE-HMM	Multi	96
Constrained MAP	Single	88
Base-transform	Single	95
Model combination STM-HMM	Single	95.2
Matched condition	Single	95
HMM vector equalization	Multi	94.3

Table 2 Experimental dataset

	Clean Speech Source (TIMIT)	
Noisy Data Source (NOISEX)	Female	Male
'volvo' (interior vehicle noise)	'fskp0'	'mhmg0'
'factory1' (industry noise)	'fsmm0'	'mclm0'
'babble' (babble voice)	'ftbw0'	'mdwh0'
'leopard' (heavy military vehicle noise)	'fsms1'	'mges0'
'machinegun' (gun sound)	'fjxm0'	'mwac0'
'f16' (f16 cockpit noise)		
'hfchannel' (high frequency radio channel noise)		
'm109' (tank noise)		
'destroyerengine' (Destroyer engine room noise)		
'buccaneer1' (Jet cockpit noise)		

ten sources of ambient noises selected from database 'NOISEX' [1], that primarily comprises military-oriented sounds as well as other frequently recurring noises. All the concerned noise ranges from 3 to 4 seconds, out of which around 60% are utilized for training, 40% is shared for testing and validation purpose. In addition, there is a test set used for modeling and separating sources of overlapping and noisy voice signals. Similarly, ten sources of precise speech data from TIMIT database [20] was considered which further comprises five dialects of both female and male speakers each for training purpose. Six utterances are utilized in learning for every speaker, with the other four being utilized for testing and validation purpose. Each speech lasts between 2 to 4 seconds. Validation sets are used mainly for tuning the parameters. For our research, MPS Speech recognizer library is utilized for processing, and analyzing the speech data.

It is also necessary to examine and classify the noise signals to improve and separate them from noisy speech more effectively. A sample of an undistorted speech utterance and its matching spectrogram is shown in Figures 1(a)

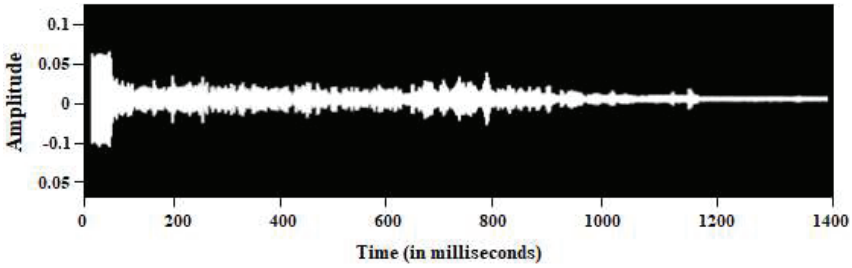

Figure 1(a) Time domain of clear speech segment *[Sample: 'mclm0']*.

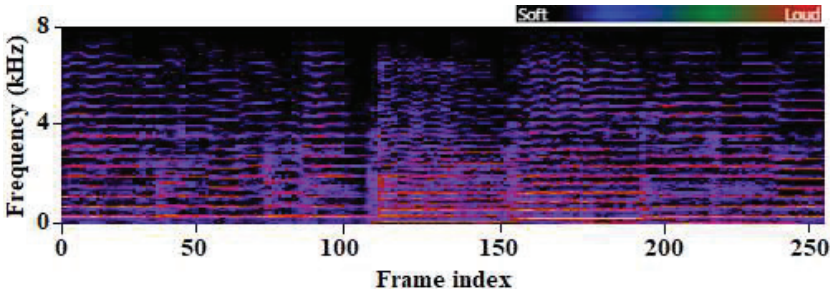

Figure 1(b) Spectrogram of clear speech segment *[Sample: 'mclm0']*.

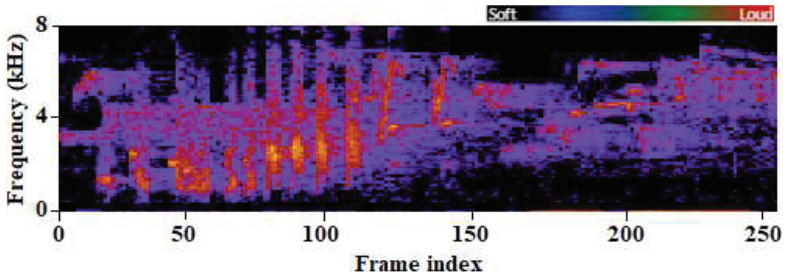

Figure 2(a) Time domain of noisy signal segment *[Sample: 'babble']*.

and 1(b) respectively, whereas a sample of babbling noise and its matching spectrogram is shown in Figures 2(a) and 2(b), respectively. For the noise signal, we utilized NOISEX and the clear spoken utterance from TIMIT [16]. The spectrogram reveals that the majority of the energy intensity is focused in the lower frequency range. As illustrated in Figure 3, on utilizing SNMF source recovery strategy, at 0 dB SNR background noisy signal is introduced to produce the noise appended speech signal.

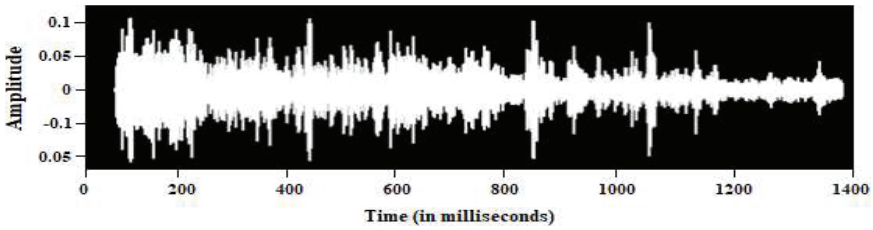

Figure 2(b) Spectrogram of noisy signal segment *[Sample: 'babble']*.

3.1 Source Recovery

To effectively generate sparse reconstructions of the test audio signals under non-negative aspects, we utilized the Supervised Non-negative Matrix Factorization (SNMF) [18] methods. First, a non-negative nonlinear atom combination and a reported amplitude spectra are compared to see which one has the least generalized divergence. The following Equation (1) is a possible solution to the minimization issue for SNMF [18],

$$\min_{a} D(m_b \| dm_a + \xi \| m_a \|_1; \quad m_a \geq 0 \tag{1}$$

Here, 'm_b' represents the matrix with test characteristics (b) as its columns. 'm_a' denotes weighted matrix, 'd' indicates the dictionary, and 'D' is the difference among 'm_Y' and dm_a (divergence). Multiplicative revisions [80] are applied to determine 'm_a' using ξ, which is set to 5 and depicted in Equation (2) as

$$m_a = \left[\frac{\frac{b}{Dm_a} \cdot (D^T)}{D^T + \xi} \right] \bigoplus m_a \tag{2}$$

In the Equation (2), $\bigoplus$ indicates attribute-wise multiplicative operation, whereas the fraction bar represents component-wise reduction. Thus, whenever the absolute deviation throughout the divergence drops under 0.001ε or even the total iterations count reaches >100, the above multiplicative updatable mechanism stops.

4 Classification of Clean and Noisy Speech Source

For an effective study, we opted for a recognizer based on *DTW* (Dynamic Time Warping) [4]. It's a feature reference recognition method that makes the modeling recognize isolated lexical items/words appropriately and effectively. Moreover, it uses dictionary learning relied on Cosine similarities [25].

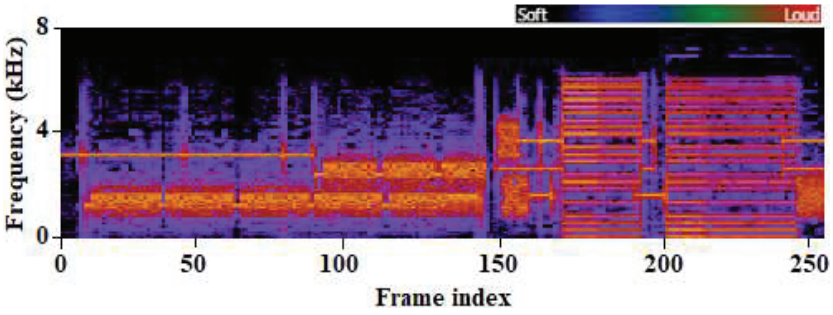

Figure 3 Spectrogram of noisy speech segment.

Dictionary learning (*DL*) is one way of machine learning (*ML*) approach to compute *'d'* matrix of the sparse dictionary, where the learning (trained) set is a linear mixture of items (columns) and the respective sparse linear weighted vector, $\widehat{\mathbb{V}}$ of *d*. The *DL* method was chosen because of its core characteristics like discriminative, non-negative and in-coherent are required for improved categorization and isolation. There are *'n'* column units/vectors known to be "atoms," labeled as 'd_i' $(1 \leq i \leq n)$. Here, *d* with a perfect matrix which is specified as $d \in r^{s\,x\,n}$ where '*n*' column units/vectors known to be "atoms", and depicted as d_i. The L_2 norm has been applied to all of the atoms in *'d'* to ensure consistency. The goal is to determine the weight vector $x \in r^n$, whose units match every dictionary atom, d_i, for a provided feature vector $y \in r^s$ and *'d'*. If *y* is not inside the range of *d's* columns, the calculation $y = dx$ seems to have no result or has an unlimited number of possibilities. It is possible to get around the issue of nil resolution by making *'d'* a complete rank matrix that covers all of r^s. Then, according to [19, 20], enforcing sparseness upon this weight matrix m_a yields a distinct resolution. Furthermore, a distinct resolution for m_a can be produced by limiting the range of non-zero items in m_a to a value depending on the dictionary's mutual spark or cohesiveness [19, 20]. The lowest count of sequentially relied columns in *'d'* is termed as spark, whereas the mutual coherence, $\mathcal{G}_d$ is the maximal for the bilateral actual intrinsic product among distinct columns in *'d'* which is depicted in Equation (3) as,

$$\mathcal{G}_d = \max_{[j \leq n,\ 1 \leq i,\ i \neq j]} |d_j d_i^T| \tag{3}$$

The equivalency between *y* and *dx* must be relaxed (i.e. $y \approx dx$) to obtain a suitable resolution for real-world usage. A constraint optimizing issue is used to exert an upper limit on the relative sparsity to obtain a resolution that

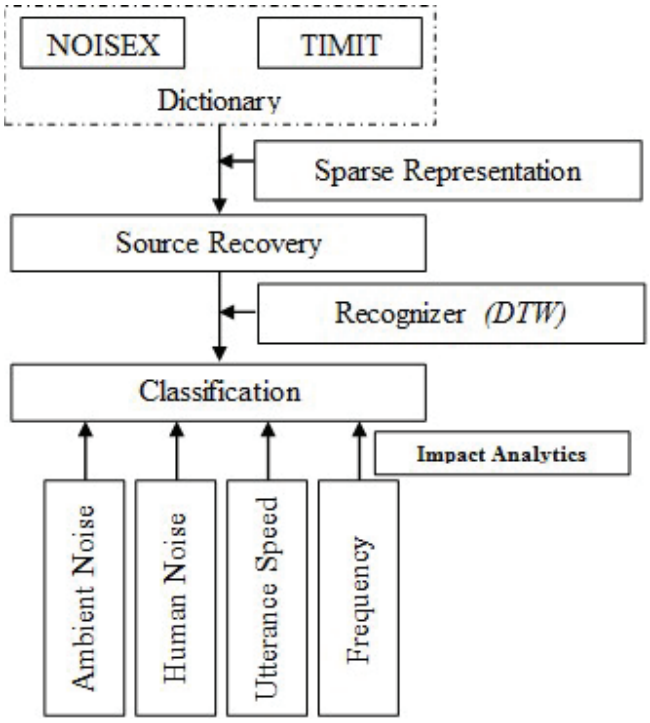

Figure 4 Procedure of principal strategy.

is significant and unique, which is represented in Equation (4) as,

$$\min_{x} D[dx, y] \quad w.r.to \quad \tau \geq \|m_a\|_0 \tag{4}$$

wherein, D estimates the range between dx and y, and τ is the maximum count of non-entries in the matrix m_a. Thus, this technique of calculating weights is known as sparse recovery, sometimes also referred to as sparse scripting or coding. Figure 4 represents the entire strategy process.

5 Speech Data Analytics

5.1 Experimental Parameters

A wide range of variables influences speech interaction. In this study, ambient/external noise, utterance speed, human interference/noise, and frequency were considered independent variables. Several trial ranges were chosen for

Table 3 Parameters for experimental trails

Impact Factors	Study Setup
Ambient Noise	*Decibel ranges:* {20, 30, 40, 50, 60, 70, 80}
Utterance Speed	*Speech pace:* slower (200 wpm), regular (300 wpm), and rapid (450 wpm)
Human Noise	*Interference decibel range:* {55, 60, 65}
Frequency	*Male in TIMIT:* {Low, High}
	Female in TIMIT: {Low, High}

Table 4 SR accuracy analysis at varying ambient noise level

Noise Level	Speech Recognition Accuracy (%)
20	98.12
30	94.36
40	89.42
50	77.24
60	68.28
70	61.23
80	54.25

each. For considered datasets (NOISEX and TIMIT), one of the popular speech recognition modules, namely "mihup" is utilized for experimental assessments with additional functionalities like customizable linguistic database, voice-oriented controlling functionality, etc. The vital parameters for experimental trails are depicted in Table 3.

5.2 Influence of Ambient Noise

Considering typical speech has a sound intensity ranging from 50 to 60 dB, the lowest ambient noise level is 60 dB (ranging from 20 to 80 dB). Researchers discovered that when the noise level exceeds 80 dB, noise interference on SR becomes exceptionally substantial, and linguistic communication among employees is severely hampered due to this distortion. This meant that the trial's peak loudness range was fixed at 80 decibels. The ambient noisy signal's spectral element distribution is mostly defined between 0 and 20 kHz.

As can be seen from the Table 4, SR accuracy varies under the different ambient noise levels ranging from 20 dB to 80 dB. SR in a low-noise setting has a high recognition rate between 20 and 30 decibels (dB). However, in accordance with the research's initial expectations, SR accuracy falls as the intensity of ambient noise increases. For instance, as shown in Figure 5,

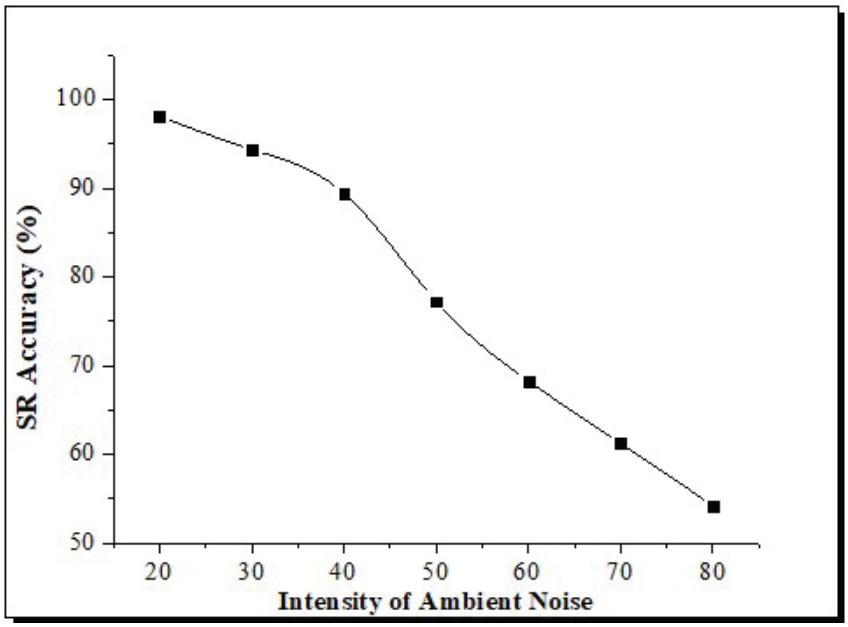

Figure 5 Evaluation of SR accuracy under varying ambient noise conditions.

when ambient noise rises, the percentage of noisy speech signals increases; in contrast, the recognition rate decreases.

This is because ambient noise destroys acoustic signals, leaving a significant portion/subset of the signals unrecognized by the voice unit. The accuracy at decibel levels ranging from 70 dB to 80 dB is substantially different from other decibel levels. Thus, increased intensity in noise seems to have a significant impact on strengthening the SR accuracy. However, the final SR accuracy result at decibel levels of 80 dB is only 54.25 percent, suggesting that SR may not be a suitable interactive tool for any commanding/controlling station without the inclusion of proactive noise-reducing systems.

5.3 Influence of Utterance Speed

Most individuals' conversational speech ranges between 250 and 350 terms a minute (wpm-word per minute) in everyday conversation, and even when they deliberately decelerate their speech pace, they still exceed 200 wpm. The speech pace of most individuals is around 450+ wpm. As a result of these considerations, the research's speech pace was categorized into three segments: slower (200 wpm), regular (300 wpm), and rapid (450 wpm). It

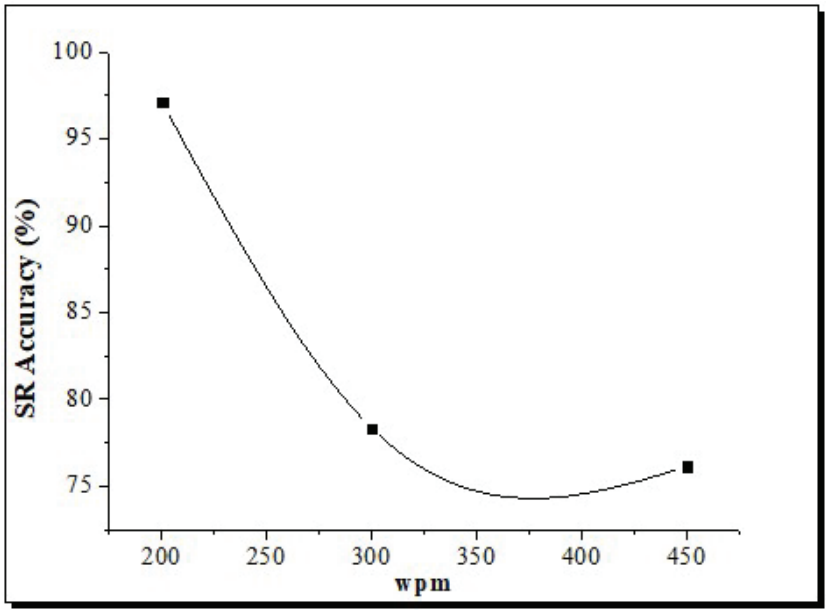

Figure 6 Evaluation of SR accuracy under varying wpm.

is shown in Table 4 indicates the speed variation of typical speech starts from 200 wpm to 450 wpm, with a recognition accuracy rate of over 80% in between these speed paces. SR precision is extremely high, even at a relatively modest pace of 200 wpm. Speech characteristics (which including liaison, swallowing sound, linguistic euphemism, etc.) are mostly intact. The moderate speech pace is at 300 wpm, and the detection rate has dropped significantly. Speech at 450 wpm is rapid, and accuracy drops slightly from 300 to 450 wpm.

Similarly, speaking faster than the maximum speed has a minimal effect on SR. However, speaking faster can unavoidably cause issues, including distortion of speech and voice intonation shift. These issues may compromise SR accuracy and make it more difficult for users to understand what is being said.

5.4 Influence of Human Noise

Due to the possibility of mutual interference amongst communicative people during actual activities, other persons' interference noises are considered an audio signal with almost the same loudness level that ranges from 55 to 65

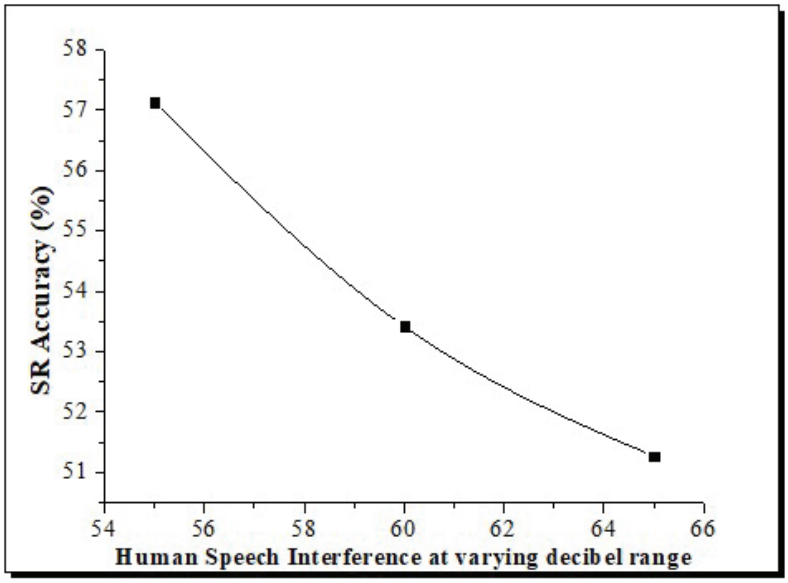

Figure 7 Evaluation of SR accuracy under varying human speech interference.

dB. As a result, in this research, we used human noisy intervention signals with dB values of 55, 60, and 65.

5.5 Influence of Frequency

Given the sound's ability to obscure the listener's perception (mask effect), the study used two audio signals: low and high, which were achieved by choosing both male and female vocal speech datasets. According to the results from Figure 8, the study findings indicate that audio frequencies and noisy intensity have a particular interaction impact. According to the results of the research, audio frequency and noisy intensity have an interaction impact. The recognition levels from high and low voice signals decline dramatically as the level of background noise increases. Furthermore, when evidenced by the higher frequency range, the accuracy of low-frequency indicators decreases significantly, and the rate of unrecognized changes especially. When ambient noises are increased, male vocal speech recognition's accuracy degrades more noticeably, making ambient sound a more critical factor. In addition, as a consequence of the increased presence of low-frequency elements in the male vocal speech, the amount of unrecognized speech has increased significantly.

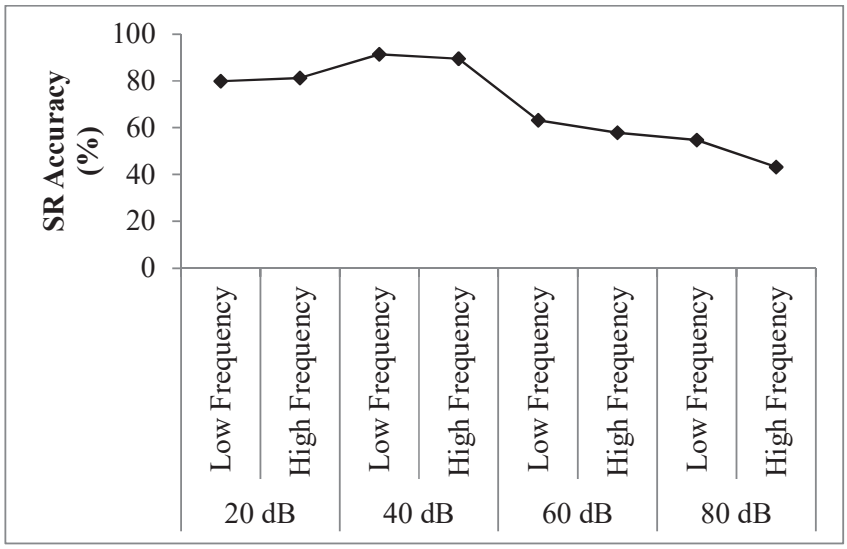

Figure 8 Recognition rate at varying frequency.

6 Conclusion

The concluding statements can be inferred from studying the effects of ambient noise, human noise, frequency, and utterance speed on SR effectiveness. SR is greatly impacted by noise, both from humans as well as from the surroundings. SR suffers considerable interference even at a reasonable noise level of 40 dB, resulting in a high number of false positives. Other ambient noise will substantially impact recognition exactly at 80 dB, which is the acceptable level of noise, leading to many incorrect assessments. Another important element affecting any voice command system is noisy interference from humans. SR is unaffected by speech pace up to a certain point in the usual range. Nevertheless, slowing down the pace of speech is essential to guarantee better recognition accuracy. The aforementioned experimental findings must be used as a guide for HCI system design, especially in the future control and command service sectors if voice interaction is included.

As we've shown, our work dealt with interference and speech categorization in a variety of different situations. Consequently, high-frequency voice signals are more readily identified ($\sim$98.12 per cent) than low-frequency speech signals. Mobile communications and auditory devices highly benefit from the suggested sparse representations classification approach for noisy speech. The audio knowledge-based technique for recognition has been

handled independently from the well-known speech dataset. Ultimately, in the future, we plan to utilize the sparse representation-based categorization for additional classes such as dialect, ethnicity, intonation, accent, and even emotions.

References

[1] "NOISEX-92." http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html. [Online] Accessed: 2017-03-30. S18

[2] Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 550–563, 2010. S41

[3] Aron, J. (2011). How innovative is Apple's new voice assistant, Siri? In: Elsevier. M10

[4] B. Laperre, J. Amaya, and G. Lapenta, "Dynamic Time Warping as a New Evaluation for Dst Forecast With Machine Learning," Frontiers in Astronomy and Space Sciences, vol. 7, Jul. 2020. DWt

[5] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in INTERSPEECH, pp. 717–720, 2010. S77

[6] Bellegarda, J. R. (2014). Spoken language understanding for natural interaction: The Siri experience. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 3–14): Springer. M11

[7] C. Couvreur, V. Fontaine, P. Gaunard, and C. G. Mubikangiey, "Automatic classification of environmental noise events by hidden Markov models," Applied Acoustics, vol. 54, no. 3, pp. 187–206, 1998. S2

[8] C. Joder and B. Schuller, "Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition," in Speech Communication; 10. ITG Symposium; Proceedings of, pp. 1–4, VDE, 2012. S74

[9] C. Müller, Speaker Classification II. Springer, 2007. S1

[10] C. Tzagkarakis and A. Mouchtaris, "Sparsity based robust speaker identification using a discriminative dictionary learning approach," in Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European, pp. 1–5, IEEE, 2013. S73

[11] D. O'Shaughnessy (1989), "Enhancing speech degraded by additive noise or interfering speakers". IEEE Commun. Mag., February 1989, pp. 46–52.

[12] Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Williams, J. (2013). Recent advances in deep learning for speech research at Microsoft. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. M9

[13] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in International Conference on Latent Variable Analysis and Signal Separation, pp. 140–148, Springer, 2010. S75

[14] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," IEEE Transactions on Information theory, vol. 50, no. 10, pp. 2231–2242, 2004. S20

[15] -J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2067–2080, 2011. S76

[16] J. Hernando and C. Nadeu (19941, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques", Proc. IEEE Internat. Con& Acoust. Speech Signal Process., Adelaide, Australia, April 1994, Vol. II, pp. 69-72.

[17] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in Proc. of the 6th Int. Conference on Digital Audio Effects, Citeseer, 2003. S13

[18] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF–half-baked or well done?," Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 2015. S80

[19] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in Audio Engineering Society Convention 128, Audio Engineering Society, 2010. S9

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon technical report, vol. 93, 1993. S43

[21] J.M. Salavedra, E. Masgrau, A. Moreno and X. Jove (1993), "A speech enhancement system using higher order AR estimation in real environments", Proc. European Con& Speech Technology, Berlin, 1993, Vol. 1, pp. 223–226.

[22] J.S. Lim and A.V. Oppenheim (1978), "All pole modeling of degraded speech", IEEE Trans. Acoust. Speech Signal Process., Vol. 26, pp. 197–210.
[23] J.S. Lim and A.V. Oppenheim (1983), "Ah pole modeling of degraded speech", in Speech Enhancement, ed. by J. Lim (Prentice-Hall, Englewood Cliffs, NJ). pp. 101–114.
[24] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 972–980, 2006. S10
[25] K. V. V. Girish, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources," in India Conference (INDICON), 2016 IEEE Annual, IEEE, 2016. S44
[26] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in Acoustics, Speech, and Signal Processing (ICASSP), 1996 IEEE International Conference on, vol. 1, pp. 353–356, IEEE, 1996. S11
[27] L.M. Arslan and J.H.L. Hansen (19941, "Minimum cost based phoneme class detection for improved iterative speech enhancement", Proc. IEEE Internat. Conf Acoust. Speech Signal Process., Adelaide, Australia, April 1994, Vol. II, pp. 45–48
[28] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," Proceedings of the IEEE, vol. 98, no. 6, pp. 995–1005, 2010. S8
[29] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer Publishing Com'pany, Incorporated, 1st ed., 2010. S19
[30] M. Feder, A.V. Oppenheim and E. Weinstein (19891, "Maximum likelihood noise cancellation using the EM algorithm", IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-37, No. 2, pp. 204–216.
[31] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2140–2151, 2013. S6
[32] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 857–869, 2005. S4

[33] Pandiyan, Sanjeevi, Ashwin M., Manikandan R., Karthick Raghunath K.M., and Anantha Raman G.R. "Heterogeneous Internet of Things Organization Predictive Analysis Platform for Apple Leaf Diseases Recognition." Computer Communications 154 (March 2020): 99–110.

[34] R. G. Malkin, Machine listening for context-aware computing. PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2006. S12

[35] Rabiner, L. R., Juang, B.-H., and Rutledge, J. C. (1993). Fundamentals of speech recognition (Vol. 14): PTR Prentice Hall Englewood Cliffs. M1

[36] S. Nandkumar and J.H.L. Hansen (1994), "Speech enhancement based on a new set of auditory constrained parameters", Proc. IEEE Internal. Conf. Acoust. Speech Signal Process.. Adelaide, Australia, April 1994. Vol. I, pp. 1–4.

[37] S. Zubair, F. Yan, and W. Wang, "Dictionary learning based sparse coefficients for audio classification with max and average pooling," Digital Signal Processing, vol. 23, no. 3, pp. 960–970, 2013. S79

[38] S.F. Boll (19791, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust. Speech Signal Process., April 1979, Vol. ASSP-27, No. 2, pp. 113–120.

[39] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE transactions on audio, speech, and language processing, vol. 15, no. 3, pp. 1066–1074, 2007. S3

[40] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Communication, vol. 9, no. 4, pp. 351–356, 1990. S16

[41] W. B. Kleijn and K. K. Paliwal, eds., Speech Coding and Synthesis. New York, NY, USA: Elsevier Science Inc., 1995. S14

[42] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview. In: Elsevier. M2

[43] Y. Ephraim and D. Malah (19841, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", ZEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-32, pp. 1109–1112.

[44] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech communication, vol. 49, no. 7, pp. 588–601, 2007. S5

[45] Y.-C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," Pattern Recognition Letters, vol. 26, no. 9, pp. 1327–1336, 2005. S78

Biographies

A. C. Kaladevi working as Professor in the Department of Computer Science and Engineering at Sona College of Technology, Salem, India has more than 25 years of teaching experience. She obtained her B.Sc degree in Computer Science from Cauvery College for Women, Tiruchirapalli followed by MCA at PSG College of Technology, Coimbatore. She completed M.Phil Computer Science from Manonmaniam Sundaranar University, Tirunelveli and M.E Computer Science and Engineering from V.M. K.V. Engineering College, Salem which was then affiliated to Anna University, Chennai. She was awarded Ph.D degree in Information and Communication Engineering during 2014 by Anna University, Chennai. Her research interest includes Data Analytics, Cloud Computing and Image Processing. She has published 21 papers in various International Journals and presented 32 papers in both national and international conferences. She has co-authored 3 books in computer science discipline. She has conducted 2 national workshops one on "Big data and Cloud for bigger transformations" funded by Department of Science and Technology (DST), New Delhi under BDI Scheme and the other one on "Empowering the Tribal Women in and around Yercaud Hills, Salem by inculcating self-employment opportunities using innovative ICT based skill development techniques" funded by Tamil Nadu State Council for Science and Technology (TNSCST), Chennai under Dissemination of Innovative Technology Scheme. She has guided more than 25 PG and 40 UG projects out of which 3 UG projects were funded by TNSCST under Students Project Scheme. As an enthusiastic student counselor she has given a great moral support to students who are now placed in a much renowned positions in their career.

R. Saravanakumar currently working as an Associate Professor in the Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore 560082. He has also served as an Assistant Professor at Jayam College of Engineering and Technology, Dharmapuri from June 2006 to December 2014. Obtained his B.E., in Computer Science and Engineering from Bharathiyar University in 2003 and received his M.E., in Computer Science and Engineering from Anna University, Chennai in 2007. He received his Ph.D., Degree from Anna University, Chennai in 2015, he has published more than 20 research papers in refereed, Springer, and IEEE Xplore conferences. he has organized several workshops, summer internships, and expert lectures for students. He has worked as a session chair, conference steering committee member, editorial board member, and reviewer in Springer Journal and IEEE Conferences. His area of interest includes Machine learning, and Deep Learning.

K. Veena currently working as an Assistant Professor in the Department of Computer Science, J.K.K.Nataraja College of Arts & Science, Namakkal(Dt). She has completed M.Phil. – Computer Science in 2006 and Pursuing Ph.D. in Computer Science at Periyar University, Salem. She has

been working as Assistant Professor at J.K.K.Nataraja College of Arts & Science, Namakkal Dt. since 2005. She has completed one Minor Research Project funded by UGC. She has published 2 research papers in International Journals. Her area of interest includes Data Analytics, Neural Networks, Machine Learning and Deep Learning.

V. Muthukumaran was born in Vellore, Tamilnadu, India, in 1988. He received the B.Sc. degree in Mathematics from the Thiruvalluvar University Serkkadu, Vellore, India, in 2009, and the M. Sc. degrees in Mathematics from the Thiruvalluvar University Serkkadu, Vellore, India, in 2012. The M. Phil. Mathematics from the Thiruvalluvar University Serkkadu, Vellore, India, in 2014 and Ph.D. degrees in Mathematics from the School of Advanced Sciences, Vellore Institute of Technology, Vellore in 2019. He has 4 years of teaching experience and 8 years of research experience, and he has published various research papers in high-quality journals Springer, Elsevier, IGI Global, Emerald, River etc. At present, he has a working Assistant Professor in the Department of Mathematics, REVA University Bangalore, India. His current research interests include Algebraic cryptography, Fuzzy Image Processing, Machine learning, and Data mining. His current research interests include Fuzzy Algebra, Fuzzy Image Processing, Data Mining, and Cryptography. Dr. V. Muthukumaran is a Fellow of the International Association for Cryptologic Research (IACR), India; He is a Life Member of the IEEE. He has published more than 40 research articles and 4 book chapters in peer-reviewed international journals. He has published 6 IPR patents in algebraic with IoT applications. He also presented 25 papers presented at national and international conferences. He has also been a guest editor of several international journals including, Journal of Intelligent Manufacturing (Springer), International Journal of Intelligent Computing and Cybernetics, International Journal of e-Collaboration (IJeC), International Journal of Pervasive Computing and Communications (IJPCC), International Journal

of System of Assurance Engineering(IJSA), International Journal Speech Technology (IJST)-Springer, Journal of Reliable Intelligent Environments (JRIE).

N. Thillaiarasu currently working as an Associate Professor in the School of Computing and Information Technology, REVA University, Bengaluru, He has also served as an Assistant Professor at Galgotias University, Greater Noida from July 2019 to December 2020. He worked 7.3 Years as an Assistant Professor in the Department of Computer Science and Engineering, SNS College of Engineering, Coimbatore. Obtained his B.E., in Computer Science and Engineering from Selvam College of Technology in 2010 and received his M.E., in Software Engineering from Anna University Regional Centre, Coimbatore in 2012. He received his Ph.D., Degree from Anna University, Chennai in 2019, he has published more than 22 research papers in refereed, Springer, and IEEE Xplore conferences. he has organized several workshops, summer internships, and expert lectures for students. He has worked as a session chair, conference steering committee member, editorial board member, and reviewer in Springer Journal and IEEE Conferences. He is an Editor board Member of editing books titled "Machine Learning Methods for Engineering Application Development" Bentham Science. He is also working as editor for the title, "Cyber Security for Modern Engineering Operations Management: Towards Intelligent Industry", Design Principle, Modernization and Techniques in Artificial Intelligence for IoT: Advance Technologies, Developments, and Challenges" CRC Press Tylor and Francis, His area of interest includes Cloud Computing, Security, IoT, and Machine Learning.

S. Satheesh Kumar currently working as an Assistant Professor & Coordinator in the Department of Computer Science, School of Applied Sciences, REVA University, Bangalore. Currently Pursuing Ph.D. in Computer Applications at Visvesvaraya Technological University, Karnataka. He has also served as an Assistant Professor at Acharya Bangalore B-School Bangalore, from Jan 2018 to July 2019. He worked 3.5 Years as an Assistant Professor in the Department of MCA, Dayananda Sagar Academy of Technology and Management Bangalore from August 2014 to Jan 2018. He worked 4 Years as an Assistant Professor in the Department of MCA, Sri Nandhanam College of Engineering & Technology Tirupattur, Tamilnadu from August 2009 to Aug 2013. Obtained his B.Sc. in Electronics from MGR Arts & Science College in 2005 and received his MCA from Priyadarshini Engineering College, Anna University Chennai in 2008. He has published more than 7 research papers and 3 Book Chapters in refereed, Springer, Elsevier, IGI Global and IEEE Explore conferences. He is also a reviewer in Springer, Elsevier, Emerald Group publishers, IGI global. He has organized several workshops, student development program, and expert lectures for students. His area of interest includes Data Security, Cloud Computing, IoT, and Machine Learning, Network Security.