Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 2, May 2020, pp. 835~842
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i2.pp835-842

An ensemble technique for speech recognition in noisy environments

Imad Qasim Habeeb[1], Tamara Z. Fadhil[2], Yaseen Naser Jurn[3], Zeyad Qasim Habeeb[4], Hanan Najm Abdulkhudhur[5]
[1,2,3]College of Engineering, University of Information Technology and Communications, Baghdad, Iraq
[4]Biomedical Engineering, University of Technology, Baghdad, Iraq
[5]Ministry of Higher Education and Scientific Research, Baghdad, Iraq

Article Info

Article history:

Received Aug 7, 2019
Revised Nov 8, 2019
Accepted Nov 22, 2019

Keywords:

An Ensemble technique
Automatic speech recognition
Noisy speech
Speech enhancement

ABSTRACT

Automatic speech recognition (ASR) is a technology that allows a computer and mobile device to recognize and translate spoken language into text. ASR systems often produce poor accuracy for the noisy speech signal. Therefore, this research proposed an ensemble technique that does not rely on a single filter for perfect noise reduction but incorporates information from multiple noise reduction filters to improve the final ASR accuracy. The main factor of this technique is the generation of K-copies of the speech signal using three noise reduction filters. The speech features of these copies differ slightly in order to extract different texts from them when processed by the ASR system. Thus, the best among these texts can be elected as final ASR output. The ensemble technique was compared with three related current noise reduction techniques in terms of CER and WER. The test results were encouraging and showed a relatively decreased by 16.61% and 11.54% on CER and WER compared with the best current technique. ASR field will benefit from the contribution of this research to increase the recognition accuracy of a human speech in the presence of background noise.

Corresponding Author:

Imad Qasim Habeeb,
Department of Mobile Communications and Computing Engineering (UOITC),
College of Engineering, University of Information Technology and Communications, Baghdad, Iraq.
Email: emadkassam@uoitc.edu.iq

1. INTRODUCTION

The main objective of ASR research is to build a system to convert speech signals to text [1]. In recent years, speech to text technology began to change the manner in which we live and became one of the basic means for humans to communicate with certain devices. Hence, many applications have been created in which speech to text technology plays an essential role [2-3]. These applications provide services, such as voice search, speech translation, personal assistant, and gaming [4-5]. The ASR systems comprise of four conceptually distinct stages: signal processing, feature extraction, acoustic model, and N-gram language model [6-7]. The signal processing enhances the speech signal by eliminating noise and making it more suitable for recognition. The feature extraction stage identifies important features in the speech signal and extracts them. The acoustic model measures a score for all characters in order to classify them using standard features. The N-gram language model measures the probability of a sequence of words to validate the resulted sentences.

Most ASR applications perform acceptably in clean environments [8]. However, they do not work well in the presence of noise [9-11]. Noise as a term refers to the unwanted elements present in speech signals. The noise of any type makes the process of ASR harder. For instance, identifying the speech of a person in a silent room is much easier than identifying the speech in a noisy environment. Thus, several researchers reported that ASR accuracy is still low for a degraded speech signal [12-14]. The effect of

different noise types on speech recognition varies significantly according to the source of the noise [15]. However, the environment of the audio signals is the main cause of noise and contrast in the speech signal [16]. The noise types may result from hundreds of sources, such as microphone quality, speaker characteristics, background sounds, and dialect differences [4]. Furthermore, various types of noise give different levels of errors, making it difficult to implement a filter technique for each type of noise or training the ASR on them [14]. Thus, the more efficient is the design of a general technique that can be more accurate for speech recognition in the presence of noise. Figure 1 presents four types of speech signals.

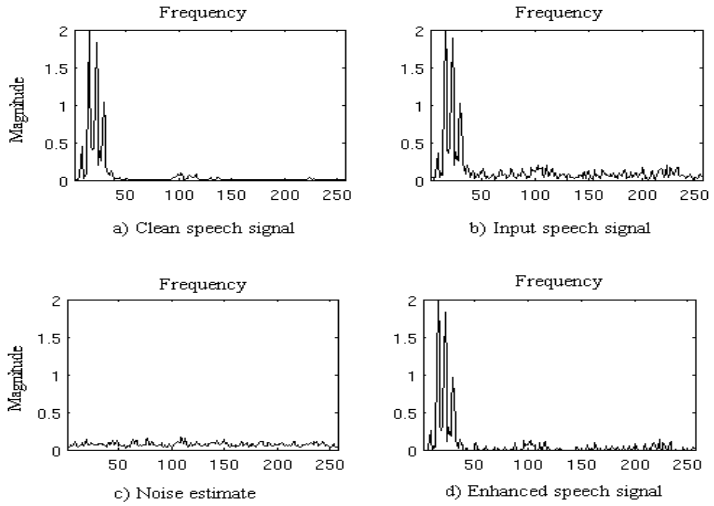

Figure 1. Four types of speech signals

In Figure 1, the four types of spectrograms show that the enhanced speech signal (d) is not exactly like the clean speech signal (a). This is because noise estimation (c) presented by one of the speech enhancement techniques does not exactly match the noise signal due to the noise is unpredictable [17]. For example, when riding in a car and listening to the friend's speech, the noise from the car increases and decreases disproportionately as the car changes its speed while the speech from the friend remains constant. Therefore, it is difficult to categorize the input signal as either noise or speech. Ensemble techniques are used effectively in a variety of domains such as optical character recognition to produce high accuracy when using noisy testing datasets [18-19]. Therefore, this research modified this technique to be suitable for the automatic speech recognition domain. The main factor of this technique is the generation of K-copies of the speech signal using three noise reduction filters. The best speech features for these copies can then be selected as final ASR features [19].

The remainder of this research was prepared in four sections: the current and older studies were formulated in Section 2 while the implementation of the proposed technique was explained in Section 3. In Section 4, the results of the proposed technique evaluation were reported. Finally, the conclusion with a brief discussion and future work of this research were described in Section 5.

2. RELATED WORK

Researchers have proposed several techniques that provide a variety of solutions to deal with the noisy speech signal. In [15], the ASR system was designed to recognize the speech of a speaker in a noisy environment. The ASR system consists of three sequentially connected components to remove the noise of a speech signal. The first component was used to maintain the speech signal coming from the speaker's direction and to ignore any noise coming from another direction. The second component was used to deal with any noise associated with the speech signal. The last component improved ASR recognition by mapping of spectral features to standard features. Experimental results of different settings of the ASR system show that integrating the three components could increase the ASR accuracy.

In [12], the ASR system was created and trained using electroencephalography features to increase the recognition accuracy in the absence and presence of the noisy environment. The electroencephalography features can be measured by recording the electrical signals that occur in the human brain. Based on this

technology, a set of speech signals features were identified that could be used for better representation of audio signals. These electroencephalography features have been determined using a deep learning model. The ASR system has been tested with a combination of acoustic and electroencephalography features. The experimental results show that using electroencephalography features could help the accuracy of speech recognition systems. However, the testing dataset was small, which consist of the four English words 'no', 'yes', 'right', 'left' and five English vowels.

In [13], the ASR model was designed to include an additional subsystem to correct the resulting errors in speech recognition. The subsystem measured the context of the phrase using a neural network that trained from a large corpus for better choosing between different possibilities as well as re-introduces unseen phrases in the corpus. Hence, it could provide corrections for ASR errors resulted from noisy environments. Experimental results showed that the ASR model could improve the recognition accuracy (1) by scoring the lattices, (2) by correcting words pruned from the lattices, and (3) by generating candidates for any word not shown in the dictionary.

In [20], the authors designed the ASR system based on neural networks and deep learning of unsupervised data to improve recognition accuracy in noisy environments. This ASR system integrates residual learning and batch normalization, showing more robustness than other existing works. It also focused on training using large and different types of noises. Furthermore, the system processes the speech signal several times in which each stage corrects the errors of the previous stage. In this way, the recognition accuracy for each stage is increased. The evaluation process of the system was achieved using clean and noisy speech signals. Experimental results showed that using neural networks and deep learning could reduce the word error rate by 5.67%.

In [9], the authors claimed that traditional neural networks that use speech classification are sensitive to various noisy conditions. Therefore, they proposed a new model for the neural network to handle uncertainty data. They suggested that speech signals were considered as input signal and their noise was modeled as uncertainty data. Uncertainty data was calculated for specific frequency points of speech spectrogram to produce the uncertainty matrix. Then, two parallel paths based classification model is suggested. The first path used a speech spectrogram as input while the second path used uncertainty matrix. The two paths outputs were joined to calculate the final output of the ASR classifier. The proposed technique has been compared with traditional neural networks using isolated words. The experimental results showed that the proposed technique achieves recognition accuracy of 85% in noisy environments.

In [17], the authors used a variant of several deep neural networks (SDNN) based speech recognition techniques. This technique estimates the desired speech spectrum as an average of multiple SDNN outputs. The weights were measured by an additional network. The multiple SDNNs and the additional network are trained together. Experiments have been conducted using two and four SDNNs that trained on the large corpus with various noise types. The proposed technique has been compared with a single DNN based ASR system. The evaluation metrics were non-standard, which are Short-Term Objective Intelligibility and Perceptual Evaluation of Speech Quality. The test results indicated that the proposed technique was better than the baseline scheme in both clean and noisy environments. The improvement was 0.07 and 0.04 in Perceptual Evaluation of Speech Quality compared to single DNN for clean and noisy speech signals respectively.

Related work of this research shows that various efforts and techniques were achieved for recognizing noisy speech signal. However, most of them did not involve ensemble techniques as a mechanism to correct ASR errors. Hence, the contribution of this research is to design and evaluate whether ensemble techniques can make a difference in improving ASR systems. Furthermore, any improvements in the ASR field can increase the overall performance of speech recognition technology.

3. PROPOSED TECHNIQUE

As mentioned previously, since the noise sources can vary widely and the conditions of speech surrounding are variable and change over time, it is not possible to design a filter technique for each type of noise. Hence, this research proposes the ensemble technique that can be more accurate for speech recognition. The main idea of this technique is that instead of relying on anyone imperfect noise reduction filter, the proposed technique incorporates information from multiple noise reduction filters of the same speech signal to improve ASR output [21]. The proposed technique suggested that different noise reduction filters potentially offered complementary information about the phonemes to be classified which could be harnessed to increase the performance of the ASR system. Figure 2 presents a diagram to illustrate the proposed ensemble technique.

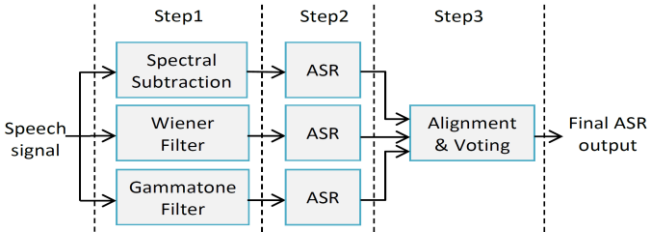

Figure 2. The Diagram of the proposed ensemble technique

In Figure 2, three main steps are included in the proposed ensemble technique. In Step1, the ensemble technique is used to create 3-copies of the input speech signal by using three different noise reduction filters: Spectral Subtraction, Wiener filter, and Gammatone filter. These three filters were chosen because they are considered the best in removing noise from the signal [14]. The generated copies are similar but non-identical. Hence, the small difference between generated copies can produce different ASR outputs and then choose the best among them. The details of noise reduction filters are in the following paragraphs.

(Filter 1) Spectral subtraction is one of the basic algorithms proposed for removing the noise in an audio signal [22]. In this algorithm, a clean signal of noisy speech can be resulted from removing the estimation of noise from the input audio signal. The noise is estimated during non-speech areas, which are gaps in the signal that contains only noise. The spectral subtraction model in the time domain is expressed by:

$$d(m, k) = s(m, k) - n(m, k) \quad (1)$$

where $s(m, k)$, $d(m, k)$, and $n(m, k)$ represent the signals of the input speech, the desired speech, and the noise estimation respectively. The variables m and k refer to the discrete-time and the frame number respectively [22]. Assuming the speech signal and noise are produced by independent sources for most real-world cases. In the frequency domain, (1) can be represented as:

$$S(w, k) = D(w, k) + N(w, k) \quad (2)$$

where the variables $S(w, k)$, $D(w, k)$ and $N(w, k)$ are the short time discrete Fourier transforms of noisy speech, desired signal, and noise respectively, and the variable w represents the discrete frequency index of the frames. The first step in spectral subtraction is to divide the noisy speech signal $s(m, k)$ into overlapping frames. The frame length is usually equal to 0.020s when an audio file is sampled to 16 kHz. Hence, each frame has samples of 400 per second. Since the overlap between frames is 50%, then the frame-0, frame-1, and frame-3 start at samples 0, 200, and 400 respectively and so on. Since the first few frames of an input signal consisting of silence, they should be good samples of the noise spectrum. Consequently, the mean of these first few frames can be taken to estimate the noise signal. Finally, the clean signal of noisy speech can be produced from subtracting the noise estimation from the input speech signal [22].

(Filter 2) The Wiener filter is the most important technique for noise reduction and has been used in various signal enhancement applications [23]. The basic idea of this technique is to measure the estimation of the desired signal from that degraded by noise signal. This could be achieved by calculating the Mean Square Error using (3) assuming known the desired signal $P_d(w,k)$ and the input signal $P_n(w,k)$ and then trying to minimize it. The filter transfer function of the frequency field is expressed by:

$$H(w,k) = \frac{P_d(w,k)}{P_d(w,k)+P_n(w,k)} \quad (3)$$

where $H(\omega, k)$ is the Wiener filter transfer function, $P_d(\omega, k)$ is the spectrum of the desired signal, $P_n(\omega, k)$ is the spectrum of the noise, k is the frame number, and the variable w represents the discrete frequency index of the frames. The desired signal estimation in the frequency field using this filter is expressed by:

$$d\hat{}(w) = \sum_{k=-\infty}^{\infty} h_k s(w) \quad (4)$$

where *d^(w)* is the estimation of the desired speech signal, *s(w)* is the input speech signal, and h_k is the Wiener filter coefficients [23].

(Filter 3) The Gammatone filter was designed to express the performance of the human auditory system and to improve the automatic speech recognition system [14, 24]. It is a linear filter that uses logarithmically spaced defined in the impulse response time, which is measured by the product of a sinusoidal tone and gamma distribution. Hence, the mathematical expression of this filter is shown in (5).

$$r(t) = ct^{n-1}e^{-2\pi bt}\cos(2\pi ft + \theta) \quad (5)$$

where *r(t)* represents the Gammatone filter impulse response, the symbols *t, c, n, f, θ,* and *b* refer to the time, the amplitude, the filter's order, the frequency, the phase of the carrier, and the filter's bandwidth respectively [14].

Figure 2 also shows that in Step 2 of the proposed technique, the 3-copies of the input speech signal are processed by three ASR systems in parallel to create different 3-ASR outputs. Figure 3 shows an example of different 3-ASR outputs.

Reference Text =	"that	sounds	great"
Output of ASR 1=	"taat	soand	great"
Output of ASR 2=	"thit	sounds	create"
Output of ASR 3=	"that	saunds	greet"

Alignment of three different ASR outputs=

t	a	a	t		s	o	a	n	d			g	r	e	a	t	
t	h	i	t		s	o	u	n	d	s		c	r	e	a	t	e
t	h	a	t		s	a	u	n	d	s		g	r	e	e	t	

Figure 3. An Example of different 3-ASR outputs

From Figure 3, it can be seen that the number of characters of ASR output is different. This causes vertical overlap between words of the ASR resulting texts. Hence, In Step 3 of the proposed technique, there is a need to match each letter with an equivalent in other ASR outputs, which is called an alignment task. The alignment task has been performed in this research by using the Smith-Waterman algorithm. After the alignment task, a voting task will select the best character of each column to produce a final ASR output. For the voting task, this research uses the recognition confidence value returned by the ASR engine to select the best character of each column. The recognition confidence value is a metric calculated by evaluating how close the features detected in the phoneme signal are to standard phoneme signal and should be a single number from 0 to 100. Finally, Step 1 and Step 2 are executed in parallel processing to increase ASR performance.

4. RESULT AND DISCUSSION

This section highlights the results of the tests of this study. The evaluation of the proposed ensemble technique (ET) has been organized in four experiments. The first experiment was conducted to test and record the results of the ET technique. Besides, three experiments were conducted to test and record the results of three existing noise reduction techniques to be compared with the first experiment. They are Spectral Subtraction [22], Wiener filter [23], and Gammatone filter [24]. All these techniques have been implemented in MATLAB software. In addition to that, the experiments used the Kaldi toolkit as an ASR engine, which is a free, open-source library for ASR research [25-26]. The four experiments used Word Error Rate (WER) and Character Error Rate (CER) as comparative measures. WER and CER have been calculated using (6) and (7) respectively [18, 26-27].

$$WER = \frac{Incorrect\ words}{Total\ words} \quad (6)$$

$$CER = \frac{Incorrect\ characters}{Total\ characters} \quad (7)$$

Levenshtein algorithm has been used to count incorrect words and incorrect characters in (6) and (7) [27]. Besides, 5000 phrases contained 26531 words have been used as a testing dataset in the

experiments. These phrases were chosen randomly from Google Books Ngram Viewer, which is free to download and is commonly used in several domains [18-19]. The text of these phrases is useful as a reference in the experiments. The speech audio files of these phrases have been produced using Google Cloud Text-to-Speech to create the testing dataset. Noise signal has been generated using Gaussian distribution [28] and added to the speech audio files of the testing dataset. Table 1 presents the tests result of the four experiments using the WER.

Table 1. Tests Result of the WER

	Spectral Subtraction	Wiener filter	Gammatone filter	The proposed technique
Total words	26531	26531	26531	26531
Wrong words	13615	12411	11747	8686
WER	51.32%	46.78%	44.28%	32.74%

In Table 1, the test results of the experiments present various values of WER for each tested technique and the WER of all tested techniques is still high for noisy speech signals. They also present that the Spectral Subtraction technique achieved the lowest value of accuracy with the WER of 51.32%. The results confirm that the Wiener filter technique achieved better accuracy than the previous with the WER of 46.78%, followed by the Gammatone filter technique with WER of 44.28%. Moreover, the proposed technique was the one that achieved the most robust results with the WER of 32.74%. It obtained a 14.72% relative reduction compared to the average error rate for older noise reduction methods. In addition, it obtained an 11.54% relative reduction compared to the best error rate for these methods. This indicates that the proposed ensemble technique achieved the best accuracy against other current methods. Table 2 presents the tests result of the four experiments using the CER.

Table 2. Tests Result of the CER

	Spectral Subtraction	Wiener filter	Gammatone filter	The proposed technique
Total characters	108124	108124	108124	108124
Wrong characters	49477	41854	41022	23072
CER	45.76%	38.71%	37.94%	21.33%

As can be seen in Table 2, the worst performance in terms of CER value was achieved by the Spectral Subtraction technique, which has a rate of 45.76%. Moreover, the values of CER for the current techniques of Wiener filter and Gammatone filter differ slightly with values of 38.71% and 37.94% respectively. In contrast, the proposed technique achieved the best CER reduction with a value of 21.33%. It obtained a 19.47% relative reduction compared to the average error rate for older noise reduction methods. In addition, it obtained a 16.61% relative reduction compared to the best error rate for these methods. Hence, when comparing the results of the proposed techniques to those of older studies, they indicate that the ET technique achieved the best reduction in the incorrect character count against the older studies.

The experimental results prove how combining multiple noise reduction techniques can help in designing a better ASR system. Furthermore, they show that the WER of noisy speech signal will always be a problem since noise is, by classification, unpredictable. Moreover, the Spectral Subtraction technique has a low accuracy because it assumes that noise is a slowly varying process or stationary and that the spectrum of noise is not significantly different due to background sounds.

4. CONCLUSION

Due to the complexity of the human hearing system and the quality of speech signals, improving the accuracy of ASR is still a challenging task. Therefore, in a real environment, robust automatic speech recognition is a common interest in the speech recognition communities. The main reason is that the speech of the speaker is corrupted by different background noises. In this work, an ensemble technique has been proposed for speech recognition in noisy environments based on three different noise reduction filters. As mentioned previously, the main idea of this technique is that instead of relying on anyone imperfect noise reduction filter, the proposed technique incorporates information from multiple noise reduction filters of the same speech signal to improve ASR output.

The experimental results proved that different noise reduction filters potentially offered complementary information about the phonemes to be classified which could be harnessed to increase the

accuracy of the ASR system. Furthermore, the results also show that the proposed technique outperforms older studies for the comparative measures of WER and CER. Consequently, it is clear that the experiments of this research confirm that the goal of this study has been accomplished. Moreover, they found clear support for the evidence that it is hard for the ASR error rate to be 0% due to the different types of noise. In future work, farther research is required to develop methods that can improve the current studies' limitations of speech to text-domain for mobile applications. In addition, since a person can estimate words and phrases even if he does not hear them completely, then it can design a post-processing technique based on the n-gram language model to provide a mechanism similar to that of the human hearing system.

REFERENCES

[1] K. Ramli and A. Jarin, "A real-time application framework for speech recognition using HTTP/2 and SSE," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, pp. 1230-1238, 2018.

[2] S. Ajami, "Use of speech-to-text technology for documentation by healthcare providers," *The National medical journal of India*, vol. 29, p. 148, 2016.

[3] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, p. 49, 2018.

[4] J.-Y. Fourniols, N. Nasreddine, C. Escriba, P. Acco, J. Roux, and G. Soto-Romero, "An Overview of Basics Speech Recognition and Autonomous Approach for Smart Home IOT Low Power Devices," *Journal of Signal and Information Processing*, vol. 9, p. 239, 2018.

[5] R. Shadiev, W.-Y. Hwang, Y.-M. Huang, and C.-J. Liu, "Investigating applications of speech-to-text recognition technology for a face-to-face seminar to assist learning of non-native English-speaking participants," *Technology, Pedagogy and Education*, vol. 25, pp. 119-134, 2016.

[6] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*: Wiley Online Library, 2018.

[7] G. Bohouta and V. Këpuska, "Performance of WUW and general ASR speech recognition systems in different acoustic environments," *The Journal of the Acoustical Society of America*, vol. 143, pp. 1758-1758, 2018.

[8] A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, "*Robust speech recognition using generative adversarial networks*," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5639-5643.

[9] E. Rashno, A. Akbari, and B. Nasersharif, "A Convolutional Neural Network model based on Neutrosophy for Noisy Speech Recognition," *arXiv preprint arXiv:1901.10629*, 2019.

[10] K. Vermeire, A. Knoop, C. Boel, S. Auwers, L. Schenus, M. Talaveron-Rodriguez, *et al.*, "Speech recognition in noise by younger and older adults: Effects of age, hearing loss, and temporal resolution," *Annals of Otology, Rhinology & Laryngology*, vol. 125, pp. 297-302, 2016.

[11] V. Z. Këpuska and H. A. Elharati, "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden markov model classifier in noisy conditions," *Journal of Computer and Communications*, vol. 3, p. 1, 2015.

[12] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech Recognition with no speech or with noisy speech," *arXiv preprint arXiv:1903.00739*, 2019.

[13] P. G. Shivakumar, H. Li, K. Knight, and P. Georgiou, "Learning from past mistakes: improving automatic speech recognition output via noisy-clean phrase context modeling," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[14] K. Garg and G. Jain, "*A comparative study of noise reduction techniques for automatic speech recognition systems*," in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 2098-2103.

[15] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "*Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition*," in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 496-503.

[16] D. Jurafsky and J. H. Martin, "*Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*", 2nd ed.: Pearson Education India, 2009.

[17] P. Karjol, M. A. Kumar, and P. K. Ghosh, "*Speech enhancement using multiple deep neural networks*," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5049-5052.

[18] I. Q. Habeeb, "Hybrid model of post-processing techniques for Arabic optical character recognition," PhD thesis, Universiti Utara Malaysia, Kedah, Malaysia, 2016.

[19] I. Q. Habeeb, Z. Q. Al-Zaydi, and H. N. Abdulkhudhur, "*Enhanced Ensemble Technique for Optical Character Recognition*," in International Conference on New Trends in Information and Communications Technology Applications, 2018, pp. 213-225.

[20] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1393-1405, 2018.

[21] I. Q. Habeeb, S. A. Yusof, and F. B. Ahmad, "Improving Optical Character Recognition Process for Low Resolution Images," *International Journal of Advancements in Computing Technology*, vol. 6, pp. 13 - 21, May 30 2014.

[22] S. Puligilla and P. Mondal, "Co-existence of aluminosilicate and calcium silicate gel characterized through selective dissolution and FTIR spectral subtraction," *Cement and Concrete Research,* vol. 70, pp. 39-49, 2015.
[23] D. Wang and C. Bao, "*An Ideal Wiener Filter Correction-based cIRM Speech Enhancement Method Using Deep Neural Networks with Skip Connections*," in 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018, pp. 270-275.
[24] B. Marković, J. Galić, Đ. Grozdić, S. Jovičić, and M. Mijić, "Whispered speech recognition based on gammatone filterbank cepstral coefficients," *Journal of Communications Technology and Electronics,* vol. 62, pp. 1255-1261, 2017.
[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, *et al.*, "The Kaldi speech recognition toolkit," *IEEE Signal Processing Society*, 2011.
[26] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language,* vol. 49, pp. 37-51, 2018.
[27] H. N. Abdulkhudhur, I. Q. Habeeb, Y. Yusof, and S. A. M. Yusof, "Implementation of Improved Levenshtein Algorithm for Spelling Correction Word Candidate List Generation," *Journal of Theoretical and Applied Information Technology*, vol. 88, pp. 449-455, 2016.
[28] S. Chehrehsa and T. J. Moir, "Speech enhancement using Maximum A-Posteriori and Gaussian Mixture Models for speech and noise Periodogram estimation," *Computer Speech & Language,* vol. 36, pp. 58-71, 2016.