EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 12, Issue 3 (2024) 51 – 60

SPEECH RECOGNITION IN VIDEOS USING FEATURES OF DIFFERENT ENHANCEMENT FILTERS

Habeeb I. Q., Al-Zaydi Z. Q., Abdulkhudhur H. N.

Abstract Video Speech Recognition (VSR) is the ability of software to extract spoken text from input sources, such as offline videos and YouTube. It achieves a good recognition accuracy rate for videos with clean environments. However, with unwanted noisy elements in speech, this presents a major challenge. Depending on where the noise originates from, there are considerable differences in the impact of various noise kinds on video speech recognition, including background music, multiple speakers, microphone quality, and dialects. It is not possible to create a filter for every kind of noise because the sources of noise might differ greatly and the speech environment's circumstances can change over time. Hence, instead of using a single filter for each type of noise, this study proposed to combine features from a small number of noise removal filters applied to the same video speech signal, resulting in a better VSR output. According to experimental results, using the proposed framework increases the accuracy of video speech recognition when evaluated on YouTube videos with different noise levels. In test experiments, the proposed framework was evaluated against state-of-the-art approaches. The accuracy mean of the relevant current approaches was improved by 16.67% and against the best accuracy of them by 7.78%. This research contributes to the video speech recognition topic as the proposed framework can facilitate the extraction of spoken phrases in millions of hours of video available online, indexing them, and then searching for them through their index.

Key words: video recognition, spoken text, YouTube videos, speech noise, Archiving speech.

AMS Mathematics Subject Classification: 68W01, 68W32, 68T50.

DOI: 10.32523/2306-6172-2024-12-3-51-60

1 Introduction

The proliferation of user-generated video, video-sharing, free digital storage, and affordable internet [1] has led to an increase in the number of videos on social networks like Facebook, YouTube, entertainment websites, and news channels. Hence, an efficient scalable method is required to search through the hundreds of millions of hours of video content that will be available online [2]. But in this growing sea of video, search engines struggle to find relevant results. This is because they don't search the videos themselves but rather terms related to them, such as keywords, tags, and subtitles [3]. However, many Internet videos have unclear text, and clips frequently have inaccurate or no metadata[4]. The growth of Internet video is hampered by the challenges of knowing which videos are related to user search. The workaround is to extract the spoken phrases from videos, index them, and then use their index to search for them in addition to the current video searches.

Phrases uttered in videos are translated into written text by computer software using Video Speech Recognition (VSR) [5]. VSR-using applications perform poorly when music or background noise is present [6]. Therefore, this research aims to enhance video speech recognition when there's noise from outside. The term "noise" describes the unwanted components found in video speech transmissions. Any type of noise complicates the VSR process. [7]. For instance, it is much simpler to recognize someone's speech in a clear video than it is in one with background noise. Therefore, according to many researchers [8], the accuracy of VSR is still poor for speech signals in videos with noise.

Different types of noise have significantly various effects on video speech recognition [9], but the background environment is the primary source of errors in a video speech signal [10]. Noise sources, including microphone quality, background sounds, speaker characteristics, and dialect differences, can contribute to different types of VSR errors. It is therefore difficult to train a VSR on all of the noise types or to apply a filter approach to eliminate each one [11]. It is consequently more efficient to provide a general framework for video speech recognition in the presence of noise that is more effective [12]. Figure 1 shows four types of audio signals that are examples of speech and noise.

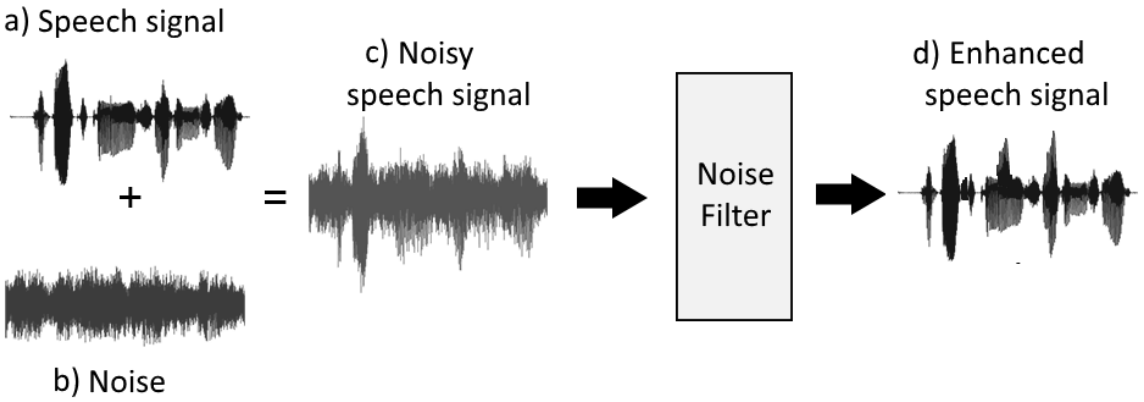

Figure 1: Four categories of audio signals that represent noise and speech.

The input signal (a) and the enhanced speech signal (d) in Figure 1 are not identical. This is because noise is unpredictable and its calculation (b) does not accurately reflect the actual noise signal [13]. For instance, when a person is driving in a video clip and his wife is talking, the noise the car makes varies depending on how the speed of the car changes, while the wifeB's speech remains constant. It is therefore difficult to identify the incoming signal as noise or speech. The multi-input framework is successfully applied in other domains using noisy test datasets, such as optical character recognition [14]. In order to make this framework appropriate for the video speech recognition domain, this study modified it. The creation of a few copies of the video voice signal using various noise removal filters is the main component of the multi-inputs framework. The final VSR output can then be selected from the best speech features of these copies.

2 Related work

This section explores the literature review and provides a critical review of existing work related to this topic. The goal is to identify limitations and reveal where this study fits into the research problem. In [15], a model has been demonstrated for visual speech recognition that is independent of the video signal and is based on lip movements. Moreover, they showed that employing bigger datasets can enhance state-of-the-art performance. Furthermore, they presented a new architecture based on auxiliary tasks to improve speech recognition in videos. English, Spanish, and Mandarin were among the languages in which all of these components were evaluated. The authors also demonstrate how integrating different training datasets enhances performance and significantly lowers the word error rate.

In [16], ambient noise inside video has been investigated, which has the potential to significantly impair speech recognition performance. Therefore, they suggested merging conventional automated speech recognition (ASR) systems with visual characteristics describing lip activity as a way to solve this issue. They presented a new multitasking audio-visual learning system to achieve this. This approach uses two tasks: automatic speech recognition as the primary task and visual audio activity detection as the secondary task. Thus, audiovisual information can be used to leverage the function of speech activity if the speech signal is not recognized. The outcomes of the experiments indicate that the suggested system gives good performance in any noise condition.

A unique language model was designed in [17] using the Google n-gram corpus as a reference text. When the output contains errors made by the speech recognition system, the proposed model corrects them. It leverages context information from sentences as part of a multi-pass filtering process that shortens processing times and increases efficiency. The effectiveness of this strategy in dropping the percentage of speech errors in the recognition process has been demonstrated by experimental results. Compared with the best word error rate of the compared approaches, it achieved a relative reduction of 15.71

Without any labeled data, the authors of [18] provided an unsupervised technique for training speech recognition models. Through adversarial training, they learned to map from these representations to phonemes by using self-supervised speech representations to segment unknown phonemes. The strategy lowers the testing benchmark's error rate from 26.1 to 11.3 when compared to the most recent unsupervised work. In the other test, their approach compared to the large English database, matches some of the top reported systems trained on 960 hours of labeled data with a word error rate of 5.9. In addition, the authors attempted nine additional languages, including low-resource tongues like Tatar and Swahili.

In [19], a single statistical model was covered, which included a hybrid model and the phonetic unit. The recognition rate using the hybrid classifier approach yields better results. The overall performance of the hybrid system is also superior to the existing models. The model demonstrates how difficult it is to resolve the conflict between information sources and self-training. Numerous issues can arise from various levels of syntax, phonetics, pragmatics, and semantics, including compilation time, disquieting execution, and power spectrum disorder. Another flaw in this suggested architecture is that, in order to be more successful, it requires an improved classification strategy.

Using electroencephalography features, a speech recognition system was created and trained in [20], to increase recognition accuracy whether or not there is noise in the surrounding environment. By capturing the electrical characteristics generated within the speaker's brain, the properties of electroencephalography may be quantified. This approach allowed for the identification of a set of speech signal attributes that could be applied to improve audio signal representation. Using a deep learning algorithm, these electroencephalography traits have been identified. According to the experimental findings, speech recognition systems' accuracy may be improved by employing electroencephalography features. However, the testing dataset, which included five English vowels and four English words was little, though. The most recent work described in this study demonstrates the range of approaches and attempts that have been made to improve noisy audio signals. However, the majority of them did not include the proposed framework, as shown in the next section.

3 Proposed Framework

The main concept behind the proposed framework is that it improves VSR output by combining data from various noise removal filters applied to the same video speech signal, rather than relying solely on one imperfect noise removal filter. A diagram illustrating the suggested VSR framework is shown in Figure 2.

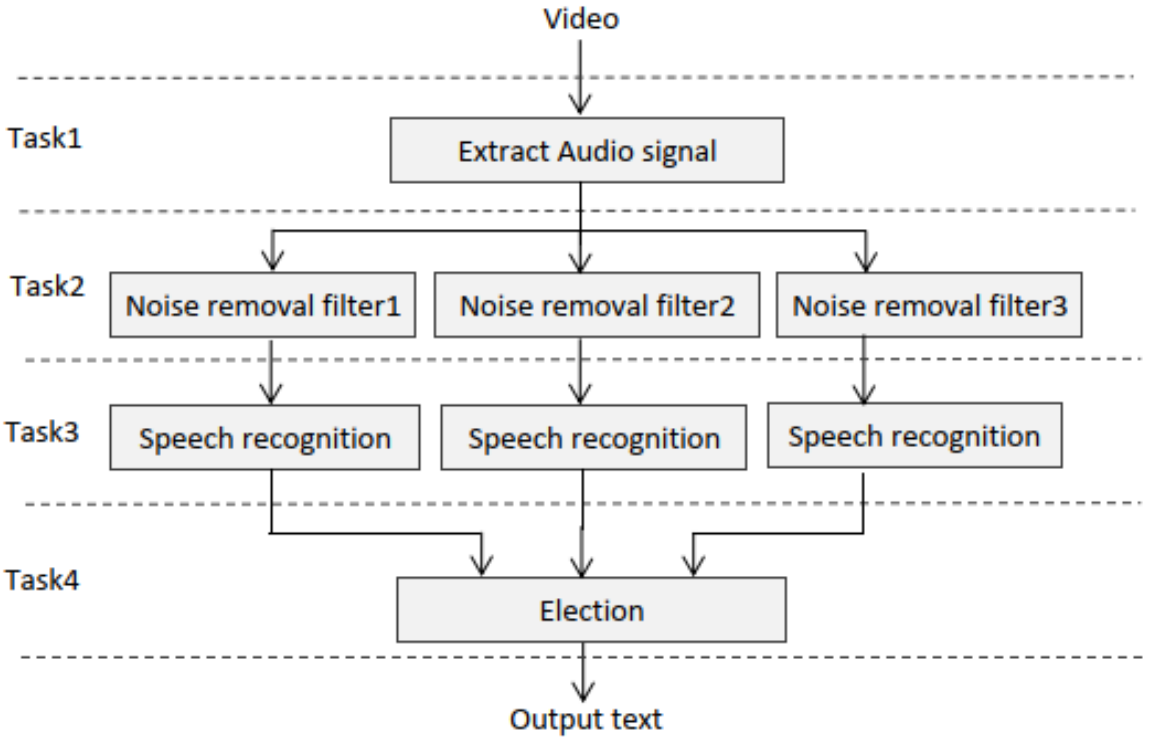

Figure 2: Video speech recognition framework.

According to the proposed framework, various noise removal filters might provide supplementary information about each character in the phrases that need to be rec-

ognized. Hence, this information could be used to improve the performance of the VSR final output. Figure 2 shows that the suggested VSR framework consists of four primary responsibilities. In Task 1, the audio signal from the input video is extracted, while in Task 2, the proposed framework creates three different versions of the audio signal. To achieve this, three different noise removal filters: the Gammatone filter, Wiener filter, and Spectral Subtraction are used. The three filters were chosen because, according to studies [21], they are the most effective in cleaning the video speech signal from noise. The versions that are created are not identical but similar. Because of this, even a slight variation in the versions that are made can produce a range of VSR outputs, from which the best one can subsequently be chosen. Further details regarding the three noise removal filters used in this investigation are provided in the paragraphs that follow. The first removal filter is based on the Gammatone filter [22], which is a result of the multiplication of the sinusoidal tone and gamma distribution. It is known as the linear filter that can be represented by the impulse response. The Gammatone method was initially applied to simulate the hearing system of humans. It is commonly used to enhance video speech signals when background noise or music is playing [23, 24]. The mathematical model for this filter is given using Equation 1 [22].

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

where *G(t)* is the impulse response function of the Gammatone filter. The variables *c, t, f, n, b,* and θ represent the amplitude, the time, the frequency, the order number, the bandwidth, and the carrier phase respectively [23]. The second removal filter depends on Spectral subtraction (SS) [25], which is a widely used method to improve the signal of audio speech when it contains additive noise. The spectrum of the additive noise is estimated when the speech spectrum is absent during non-speech periods. Therefore, the estimation of the additional noise is subtracted from the input signal to provide an enhanced speech signal. However, the disadvantage of this method is that the added noise may be very different in the non-speech periods causing difficulty in estimating the noise signal. The mathematical model for spectral subtraction in the time domain is given using Equation 2.

$$O(t, f) = I(t, f) - N(t, f) \tag{2}$$

where *O(t, f), I(t, f),* and *N(t, f)* refer to the signals of the output, the input, and the noise estimation respectively. The variable t is the discrete time while the variable f is the frame number [26]. In the frequency domain, the mathematical model for spectral subtraction is given using Equation 3.

$$O(w, f) = I(w, f) - N(w, f) \tag{3}$$

where the discrete Fourier transformations of the output, input, and noise signals are, respectively, *O(w, f), I(w, f),* and *N(w, f).* The character w is the discrete frequency index. The last removal filter is the Wiener filter [27], which is a statistical method used in several signal processing applications to improve a video speech signal that has been corrupted by noise. It can estimate the desired video speech signal from

that distorted by different types of noise. To achieve this task, the mean squared error is measured using the mathematical model given in Equation 4.

$$Y(w,f) = \frac{S_{\mathrm{d}}(w,f)}{S_{\mathrm{d}}(w,f) + S_{\mathrm{n}}(w,f)} \tag{4}$$

where *Y(w, f)* is the transfer function of the Wiener filter, $S_d(w, f)$ is the desired signal, and $S_n(w, f)$ refers to any type of noise. The character w refers to the discrete frequency index while the character k refers to the number of the frame in the signal [28].

Returning to Figure 2, it also shows that in Task 3 of the proposed VSR framework, the three versions of the input audio signal are processed by three similar Automatic Speech Recognition (ASR) algorithms in parallel to produce different 3-VSR outputs. Figure 3 shows an example of VSR outputs.

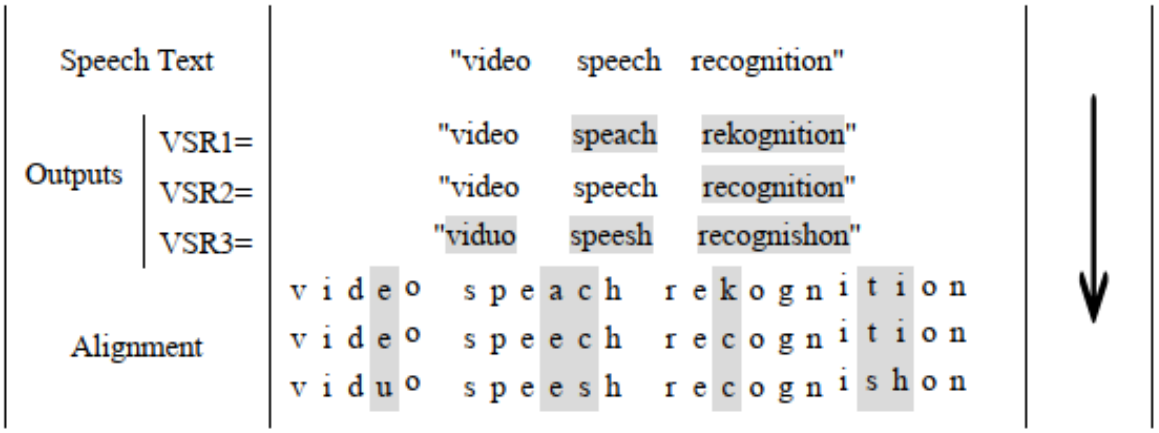

Figure 3: Different outputs of VSR.

It is evident from Figure 3 that the VSR output has a distinct character count. Words in the VSR-generated texts overlap vertically as a result of this. To handle the overlap, an alignment task [14], or matching each character to its counterpart in other VSR outputs, is required in Task 4 of the proposed VSR framework. The Smith-Waterman algorithm [29] was used in this study to solve the alignment problem. Following the alignment activity, the best character from each column will be chosen via an election task to create the final VSR output. In contrast, to choose the best word for each column in the election challenge, the research uses the dictionary to determine whether each word is included in its. If there is only one, it will be indicated as accurate; If there are multiple in the dictionary, the correct term with the highest frequency of shared letters is chosen.

4 Empirical Results

In this section, the experimental setting of the research is explained to implement the concept of the proposed framework. The test dataset was collected from YouTube

Table 1: Comparison results of the research experiments.

	Gammatone filter	Spectral Subtraction	Wiener filter	Proposed Framework (VSRF)
Correct words	11098	10916	12867	13955
Total words	15724	15724	15724	15724
Accuracy	70.58%	69.42%	81.83%	88.74%

videos with background music or annoying sounds. However, only YouTube videos with subtitles classified as not auto-generated were chosen as the base dataset for this evaluation. Hence, to solve this issue, each YouTube video was verified by the authors. After verification, these videos are added to the test dataset. These subtitles serve as a reference text to compare with the resulting texts generated by video speech recognition. The speech in the YouTube videos includes 3,421 phrases containing 15,724 words by different speakers.

Speech phrases may contain numbers, special characters, and punctuation to preserve their real case. The test video files were processed one by one as explained in Figure 2. Three other relevant techniques that are currently in use were compared to the suggested framework (VSRF): the Wiener filter, spectral subtraction, and Gammatone filter. The user interface was developed using C# in the Visual Studio.NET environment to implement and evaluate all these methods. Equation 5 shows how accuracy is calculated as a metric for evaluating the tested methods.

$$Acc. = \frac{correct\ words\ (output\ text)}{total\ words\ (reference\ text)} * 100 \tag{5}$$

In addition to that, Equation 6 was used to measure the relative improvement in video speech recognition rate:

$$Relative\ improvement\ =\ \frac{accuracy\ rate\ (A) -\ accuracy\ rate\ (B)}{accuracy\ rate(A)} * 100 \tag{6}$$

Where the term *"accuracy rate (A)"* represents the recognition rate of the proposed Framework. In contrast, the term *"accuracy rate (B)"* represents the recognition rate of the best existing method. Table 1 summarizes the comparison results of the research experiments.

Table 1 demonstrates that the Spectral Subtraction approach, with a rate of 69.42%, produced the worst accuracy. The Wiener filter method yields higher accuracy rates (81.83%) compared to the Gammatone filter method (70.58%). At 88.74%, the suggested framework VSRF yielded the best accuracy. The accuracy mean of the three related approaches that are now in use has been improved by 16.67% and against the best accuracy of them by 7.78%, using the proposed framework. This indicates that the proposed framework VSRF is the best compared to the three existing noise removal filters.

5 Conclusion

The goal of this research is to improve the accuracy of speech recognition of videos in the presence of background music or annoying sounds. This was done by proposing a solution to the restrictions of existing noise removal filters. As a result, this research presents and discusses the design details of the proposed framework. The concept of the proposed framework, flowchart, and contributions are presented in depth. Furthermore, four experiments were conducted in this study to evaluate this framework using the accuracy measure. A detailed presentation of the results of the evaluation process is explained. The results of the experiments are really promising. Compared with other existing denoising filters, the proposed framework performs better.

Accordingly, the practical results of the research indicate that the objectives of the study have been achieved. Finally, tests have also shown that disturbing videos have a high error rate due to the various types of noise. This highlights the fact that the accuracy of noisy video speech recognition is not always 100% accurate. Reducing the proposed framework's processing time is the research's next goal. Furthermore, additional research can be conducted to enhance the suggested framework to achieve higher accuracy.

References

[1] Bessarab A., Mitchuk O., Baranetska A., Kodatska N., Kvasnytsia O., Mykytiv G.,*Social networks as a phenomenon of the information society*, Journal of Optimization in Industrial Engineering, 14. 2 (2021), 17-24.

[2] Fyfield M., Henderson M., Phillips M.,*Navigating four billion videos: teacher search strategies and the YouTube algorithm*, Learning, Media and Technology, 46. 1 (2021), 47-59.

[3] Mager A., Norocel O. C., Rogers R.,*Advancing search engine studies: The evolution of Google critique and intervention*, Big Data & Society, 10. 2 (2023), 521-528.

[4] Habeeb Z. Q., Vuksanovic B., Al-Zaydi I. Q.,*Breast cancer detection using image processing and machine learning*, Journal of Image and Graphics, 11. 1 (2023), 1-8.

[5] Malik M., Malik M. K., Mehmood K., Makhdoom I.,*Automatic speech recognition: a survey*, Multimedia Tools and Applications, 80. 3 (2021), 9411-9457.

[6] Hamidi M., Satori H., Zealouk O., Satori K.,*Amazigh digits through interactive speech recognition system in noisy environment*, International Journal of Speech Technology, 23. 1 (2020), 101-109.

[7] Kawase T., Okamoto M., Fukutomi T., Takahashi Y.,*Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition*, IEEE Transactions on Consumer Electronics, 66. 2 (2020), 125-133.

[8] Alharbi S. et al.,*Automatic speech recognition: Systematic literature review*, IEEE Access, 9. 1 (2021), 131858-131876.

[9] Fahad M. S., Ranjan A., Yadav J., Deepak A.,*A survey of speech emotion recognition in natural environment*, Digital signal processing, 110. 4 (2021), 102951.

[10] Ng E. H. N., Ronnberg J.,*Hearing aid experience and background noise affect the robust relationship between working memory and speech recognition in noise*, International Journal of Audiology, 59. 3 (2020), 208-218.

[11] Fendji J. L. K. E., Tala D. C., Yenke B. O., Atemkeng M.,*Automatic speech recognition using limited vocabulary: A survey*, Applied Artificial Intelligence, 36. 1 (2022), 2095039.

[12] Abdulkhudhur H. N., Habeeb I. Q., Hussain A. B., Thamer A., Matcharan A.,*The UX of Banking Application on Mobile Phone for Novice Users*, Journal of Computational and Theoretical Nanoscience, 16. 5 (2019), 2218-2222.

[13] Yang Q. et al.,*Mixed-modality speech recognition and interaction using a wearable artificial throat*, Nature Machine Intelligence, 5. 2 (2023), 169-180.

[14] Habeeb I., Al-Zaydi Z., Abdulkhudhur H.,*Selection technique for multiple outputs of optical character recognition*, Eurasian Journal of Mathematical and Computer Applications, 8. 2 (2020), 41-51.

[15] Petridis P. Ma, S., Pantic M.,*Visual speech recognition for multiple languages in the wild*, Nature Machine Intelligence, 4. 11 (2022), 930-939.

[16] Tao F., Busso C.,*End-to-end audiovisual speech recognition system with multitask learning*, IEEE Transactions on Multimedia, 23. 2 (2020), 1-11.

[17] Habeeb I. Q., Abdulkhudhur H. N., Al-Zaydi Z. Q.,*Three N-grams Based Language Model for Auto-correction of Speech Recognition Errors*, Proceedings of the International Conference on New Trends in Information and Communications Technology Applications, Baghdad, Iraq (2021), 131-143.

[18] Baevski A., Hsu W.-N., Conneau A., Auli M.,*Unsupervised speech recognition*, Proceedings of the Advances in Neural Information Processing Systems, (2021), 27826-27839.

[19] Manoharan S., Ponraj N.,*Analysis of complex non-linear environment exploration in speech recognition by hybrid learning technique*, Journal of Innovative Image Processing, 2. 4 (2020), 202-209.

[20] Krishna G., Tran C., Yu J., Tewfik A. H.,*Speech recognition with no speech or with noisy speech*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019), 1090-1094.

[21] Park J.-S., Kim S.-H.,*Noise Cancellation Based on Voice Activity Detection Using Spectral Variation for Speech Recognition in Smart Home Devices*, Intelligent Automation & Soft Computing, 26. 1 (2020),

[22] Sivapatham S., Kar A., Christensen M. G.,*Gammatone Filter Bank-Deep Neural Network-based Monaural speech enhancement for unseen conditions*, Applied Acoustics, 194. 2 (2022), 108784.

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

[24] Markovic B., Galic J., Grozdic O., Jovicic S., and M. Mijic,*Whispered speech recognition based on gammatone filterbank cepstral coefficients*, Journal of Communications Technology and Electronics, 62. 11 (2017), 1255-1261.

[25] Gnanamanickam J., Natarajan Y., KR S. P.,*A hybrid speech enhancement algorithm for voice assistance application*, Sensors, 21. 21 (2021), 7025.

[26] Puligilla S., Mondal P.,*Co-existence of aluminosilicate and calcium silicate gel characterized through selective dissolution and FTIR spectral subtraction*, Cement and Concrete Research, 70. 2 (2015), 39-49.

[27] Kumar M. A., Chari K. M.,*Noise reduction using modified wiener filter in digital hearing aid for speech signal enhancement*, Journal of Intelligent Systems, 29. 1 (2020), 1360-1378.

[28] Wang D., Bao C.,*An Ideal Wiener Filter Correction-based cIRM Speech Enhancement Method Using Deep Neural Networks with Skip Connections*, Proceedings of the IEEE International Conference on Signal Processing (ICSP), (2018), 270-275.

[29] Xia Z. et al.,*A review of parallel implementations for the SmithB–Waterman algorithm*, Interdisciplinary Sciences: Computational Life Sciences, 14. 3 (2021), 1-14.

Habeeb I.Q.
College of Biomedical Informatics, University of Information Technology and Communications,
Baghdad, Iraq,
Email: emadkassam@uoitc.edu.iq,

Al-Zaydi Z.Q.
Biomedical Engineering, University of Technology,
Baghdad, Iraq,
Email: zeyad.q.habeeb@uotechnology.edu.iq,

Abdulkhudhur H.N.
Directorate of Second Karkh, Ministry of Education,
Baghdad, Iraq,
Email: hanan_nagem@yahoo.com,

Received 22.03.2024, Accepted 20.06.2024