RESEARCH ARTICLE OPEN ACCESS

Intelligent System for Improving Educational Videos Performance

B.M Elamir [1], A.F. Elgamal [2], M.Hussein [3]

[1], [2], [3] Department of Computer Science, Faculty of Specific Education, Mansoura University - Egypt

ABSTRACT

Artificial intelligence (AI) is revolutionizing many aspects of daily life including education. Educational videos are essential to video-based learning (VBL) and e-learning. However, many educational videos have image and sound quality issues during recording. The challenges include problems with video frames such as low light and blur, and audio challenges like noise. The present study proposes an intelligent system to address some of these challenges. This system involves several techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE), adjustment of image intensity values or color map, atmospheric haze reduction, Gaussian filter, and Lucy-Richardson method used to enhance video frames and remove blur. The Fast Fourier Transform (FFT) technique is employed to remove noise from the audio. Wav2vec 2.0 extracts text from audio files using a self-supervised learning approach. Natural Language Processing (NLP) techniques such as N-gram and TextRank are used to generate a title and summary from the extracted text. To evaluate the proposed system, this study employs metrics such as MSE, PSNR, and SSIM to assess video frame enhancement, SNR for audio denoising, WER for extracted text, and ROUGE score for text summarization. Results demonstrate high performance ratios achieved by our system.

Keywords: Intelligent System, Image Processing, Audio Processing, Automatic Speech Recognition, Natural Language Processing, Text Summarization, Educational Videos

I. INTRODUCTION

Artificial intelligence (AI) is gaining popularity in a variety of research fields, including computer vision (CV), natural language processing (NLP), and speech recognition [1]. AI is an area of computer science that may be described as the efficient use of computer technology via enhanced programming approaches [2]. One of the computer vision technologies is image and video processing [3]. Deep Learning (DL) technology is now regarded as one of the crucial subjects in the fields of Machine Learning (ML) and AI [4]. Natural language processing (NLP) is a rapidly advancing field in computer science that has seen significant progress due to the advancements in deep and machine learning technologies [5] [6]. Educational video has become a crucial tool in enhancing learning outcomes and satisfaction, particularly in the context of e-learning, it provides both theoretical and practical skills in various fields [7, 8]. Interactive videos and simulations are used in a variety of educational settings including Massive Open Online Courses (MOOCs) [9]. The impact of educational videos on the learner is significant because they provide an individual learning environment with unlimited and simple access, increase the learner's ability to social interaction, and improve the learner's motivation and concentration level. Educational video content requires careful consideration of methodological, psychological, didactic, ergonomic, technical, and legal aspects. Methodological aspects involve scenario planning, dynamism, audio and image synchronization, and video duration ranging from 2 to 10 minutes.

Psychological aspects involve reducing cognitive overload and focusing on visual and auditory communication. Didactic aspects focus on learning objectives, explanation methods, title selection, and summary development. Ergonomic aspects involve careful background care, animations, subtitle placement, and interactive elements. Technical aspects include aspect ratio (4:3), frame rate (24:30 fps), image blur prevention, sound quality, and lighting. Legal considerations include copyright and background music [10]. The rapid development of multimedia applications has resulted in a significant increase in video data [11]. Digital videos are everywhere, and related applications are gaining traction [12]. Video improvement in low light is a popular area of computer vision research [13]. However, it is difficult due to excessive noise, detail loss, non-uniform exposure, and other factors. These issues are exacerbated in videos captured from dynamic scenes [14]. Several issues pertaining to the speed, clarity, and quality of educational videos may emerge following their production with technical aspects comprising the majority of problems [15]. Intelligent systems are employed across various domains, such as healthcare, education, and other sectors to address challenges and find solutions. It has been employed in decision support systems, disease diagnostics [16], elearning, video processing [17], image recognition, speech recognition [18], and text summarization.

The present study addresses the technical aspects of the challenges by controlling and adjusting the frame rate, removing blur if it occurs, overcoming lighting challenges, and removing audio noise. It also addresses automatic speech recognition using deep learning to convert audio to written text. Furthermore, by using NLP techniques, a title and summary for the text obtained from the audio file are established, which serves as a general description of the video content. Because it may be included with the video when it is published on the Internet; it is regarded as a solution to the educational aspects and attracts the learner. This study addresses many challenges with the acquisition of high-quality videos for use in closed classrooms, online publishing, or e-learning.

II. RELATED WORKS

A number of studies have focused on making videos better, because of their significance in various fields. Numerous studies have attempted to enhance the contrast of an image or video. Hong et al. suggested a fusion-based deconvolution algorithm. The current study eliminates the need for direct transmission map estimation. Results demonstrated that the suggested system outperformed alternative haze removal techniques in terms of effectiveness in removing haze [19]. Thanh et al. proposed a single image elimination method that combined an adaptive histogram equation, an HSV color model, and linear gamma correction. The results show that the proposed method fades well and computer with other modern blurry removal methods [20]. Numerous studies dealt with removing noise from the sound. The study Senthamizh Selvi concentrated on the Kalman filter, which is used to remove weak speech signals from signals. Fast Fourier transform (FFT) and discrete cosine transform (DCT) are the transformations that are employed. There will be an improvement in the output signal and some degree of noise reduction [21]. Khan and Chouksey supposed that deep learning is combined with a highly optimized adaptive filter, or DLAF filter, and showed that the aim is to develop high-resolution filters capable of processing any input signal with a low bit error rate (BER), a high signal-to-noise ratio (SNR), and an exceptionally low mean square error (MSE) [22]. Other studies deal with extracting text from speech. Baevski et al. demonstrated how learning robust representations from speech sounds alone and then tuning into written speech can outperform the most effective semi-supervised techniques while being more straightforward conceptually. Wav2vec 2.0 quantifies co-learned latent representations to mask speech input into latent space and solve a particular contrast task [23]. Fan et al. expand the self-supervised framework to include language identification and speaker verification through some initial experiments that wav2vec 2.0 is capable of capturing speaker and language information and its performance on each of the two tasks [24]. Schneider et al., through the acquisition of raw audio representations, investigated unsupervised pre-training for speech recognition. The representations produced by wav2vec, which is trained on a lot of unlabeled audio data, are then used to enhance the training of acoustic models [25]. Jose outlines a simple and efficient technique for speech recognition and he confirms the effectiveness of the suggested strategyby carruing out many experiments [26]. Other studies dealt with the field of extracting a summary from the text. Alrumiah and Shargabi declared a different approach and used Latent Dirichlet Allocation (LDA), which has been shown to be effective in summarizing documents. The authors created useful summaries of the current 'EDUVSUM' education videos dataset for evaluation [27]. Abhilash et al. used subtitles to summarize lecture videos. NPTEL (National Program for Technological Reinforcement Learning) video evaluations were compared to man-made summaries. Punctuation in subtitles appears to be important in summarizing lecture videos [28]. Aswin et al. clarified employing speech recognition to generate subtitles for videos with and without subtitles, followed by applying NLPbased text summarizing algorithms to the subtitles. The experimental findings show the importance of the proposed approach [29]. Dilawari and Khan clarified abstract video summarization employs a deep neural network to generate natural language descriptions and abstract text summarization of the input video. Furthermore, the experiments show that the combined paradigm produces better results [30].

According to the previous review, most existing studies focus on a single aspect of video enhancement, such as enhancing video frames, reducing noise in audio, detecting speech, or summarizing the text of the subtitle file accompanying the video. This makes the system of current study more distinctive and unique, as it allows combining all aspects improvement to an intelligent system.it also allows for automatic or manual video frame rate adjustments to control the speed or slowness of the video for laboratory experiments in physics, chemistry, and other subjects, and to improve video frames by enhancing lighting, adjusting intensity, removing haze and blur, and combining these techniques to improve video quality. Unlike current studies,

which focus on summarizing text from the video's subtitle file, the proposed system summarises text extracted from speech and allows for a wide range of methodologies, techniques, and design tools to be used.

III. THE PROPOSED SYSTEM

Figure 1, shows the framework for proposed system used in this research paper. The proposed system for improving educational videos is divided into three subsystems. First, Video Enhancment by splitting video frames and audio, then video frame enhancement and audio enhancement. Finally, merge enhanced video frames and enhanced audio to produce improving video. Second, Speech Recognition by converting speech to text via audio file. Third, Extractive text summarization by suggesting titles for the video, and drawing a summary of the video content based on the extracted text from speech in video.

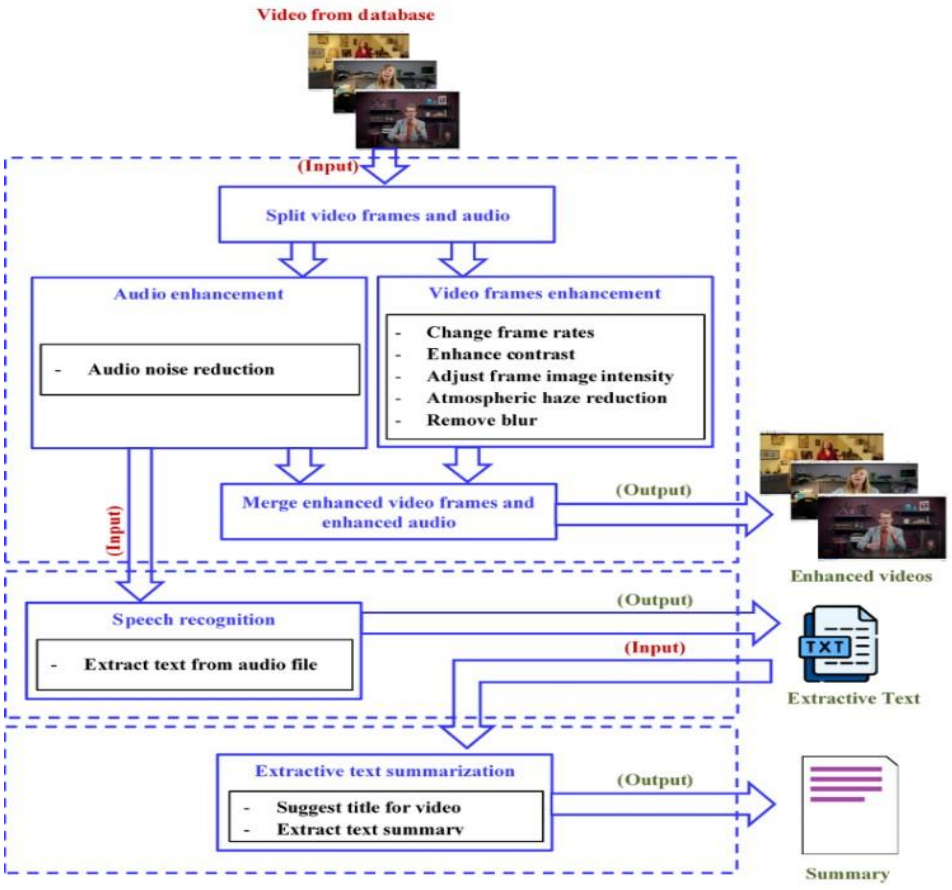

Figure 1. Schema for the proposed system

IV. VIDEO ENHANCEMENT

Improving the video begins with improving the image and sound. The image is a representation of the video frames. A video shot can include many frames. The technique of enhancing frames is critical because it is one of the video components. The technique of sound enhancement is also carried out. In this section, a separation will be made between the video and audio frames. The subsequent phase involves enhancing video frames through an adjustment of frame rate, enhancement of contrast, adjustment of frame image intensity values, reduction of haze, and elimination of blur. The subsequent phase entails enhancing audio by the elimination of extraneous noise on the audio. Ultimately, the integration of enhanced video frames and increased audio culminates in the creation of video content of superior quality.

Figure 2. shows the flowchart for the four steps, which consist of separating the frames and audio form video, enhancing the frames, enhancing the audio, and merging the enhanced frames and the enhanced sound.

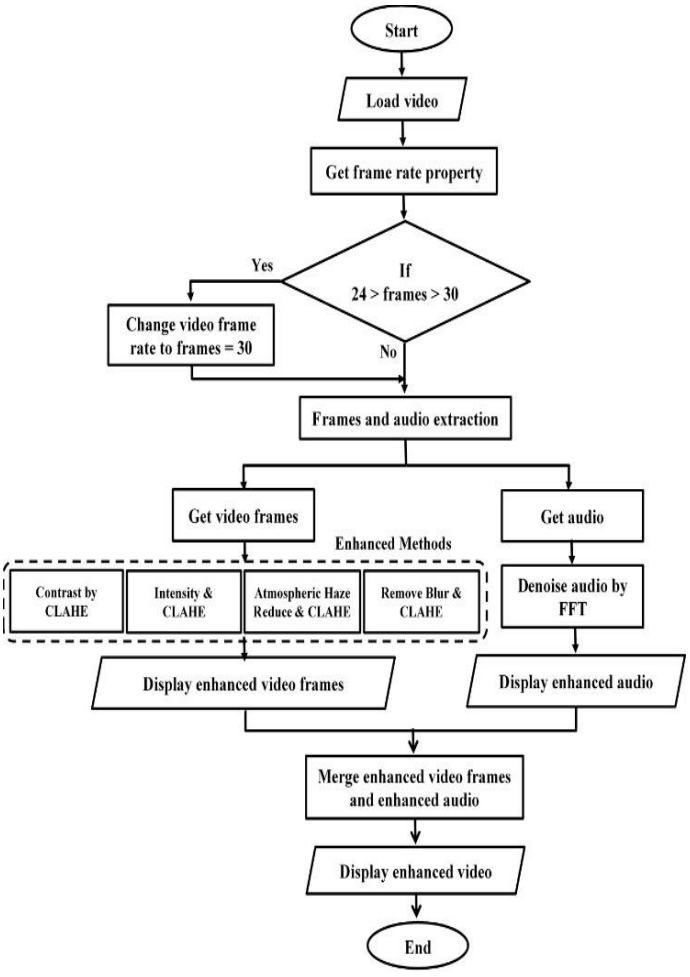

Figure 2. Flowchart for the video enhancement

A. Split Video Frames and Audio

After loading the video file into the system, the process of separating the audio and frames and saving each individually in a file begins the preparation for the efficiency procedures that will be performed on both files.

B. Video Frames Enhancement

The frames within the video are processed by separating the sound from the image and then optimizing the images of the frames. In addition, adjust the frame rate per second and set it to 30 frames per second, in case the video is fast or slow. Figure 3, shows the pseudo-code for the change in frame rate

Pseudo-code for the change in frame rate as 30

Input:
- Vid_input = Input Video will change frame rate.

Steps:
1- Read input video (Vid_input).
2- Get (Nframe_input) frame rate property from (Vid_input).
3- Get (Nduration_input) duration property from (Vid_Input).
4- Calculate total frames in (Vid_input) by
   a. tot_frame = round (Vid_input.Nframe_input * Vid_input.Nduration_input).
5- Assign the (new_frame) as 30.
6- Set output video for write (Vid_output)
7- Assign frame rate property for output video as
   (Vid_output.Nframe_output = new_frame)
8- Check if (Nframe_input) more than 30 and less than 24.
   a. For every frame in input video to (tot_frame)
   b. Read frame.
   c. Write in output video (Vid_output).
   d. End for.
9- Else
10- Vid_output = Vid_input.
11- Write output video (Vid_output). 12- End if.

Output:
- Vid_output = Output video after change frame rate.

Figure 3. Pseudo-code for the change in frame rate

Image restoration is a crucial issue in the field of image processing [31]. The enhancement of frame image contrast was addressed by employing Contrast Limited Adaptive Histogram Equalization (CLAHE) adjusting of frame image intensity values or color map. CLAHE is a more advanced variant of adoptive histogram equalization (AHE) that increases local pixel areas to improve image visibility [32]. It is applied to each pixel to control image quality [33]. the atmospheric haze reduction was applied to rectify instances of video darkness and mitigate its impact [19]. The blur seen in the image can be effectively eliminated with the application of the Gaussian filter in conjunction with the Lucy-Richardson method, which is an iterative non-linear method for image deburring that uses the blur kernel [34].

CLAHE was used to improve the contrast in the frame image, where each frame image was passed through the technique applied to the components of red, green, and blue, and finally combined the components of the three colors, as shown in the figure 4.

In the proposed system, adjust frame image intensity and CLAHE were adjusted simultaneously to boost color brightness and clarity. Both the atmospheric haze removal in the frame image and CLAHE were combined to remove the haze and raise the degree of contrast. Both were combined.

Gaussian filter and Lucy-Richardson method for removing blur in the frame image and CLAHE were combined to remove blur and boost contrast at the same time. Figure 5, the flowchart depicts the progression of these steps.

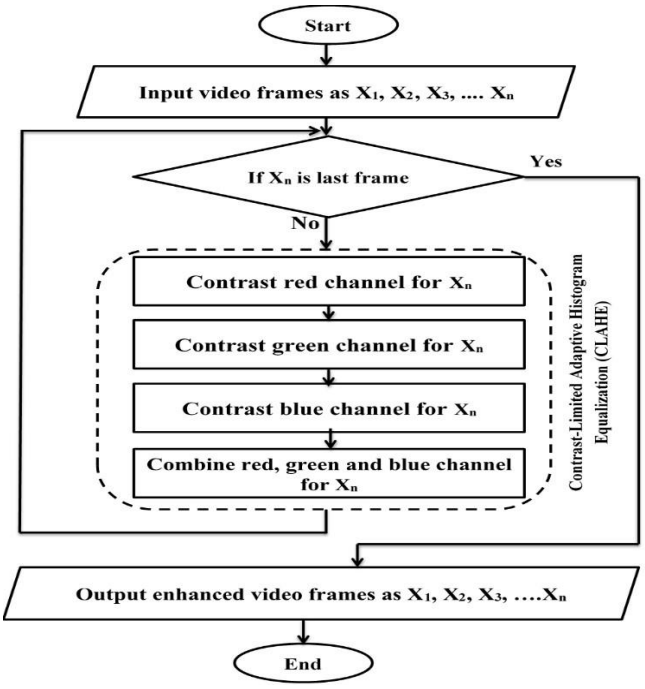

Figure 4. Flowchart for CLAHE Steps

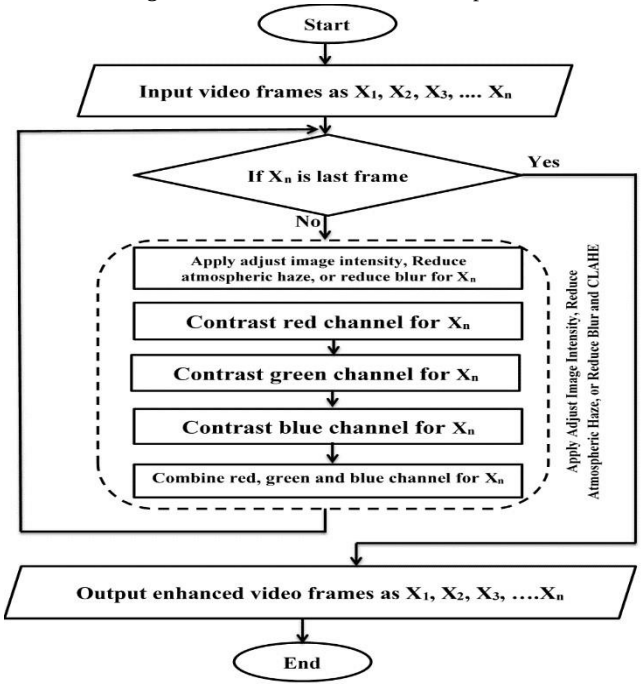

Figure 5. Flowchart for combining adjusting image intensity, atmospheric haze reduction, or removing blur with CLAHE processes

V. AUDIO ENHANCEMENT

In the proposed system was applied the reduction of noise in audio is using a Fast Fourier Transform (FFT). FFT is a mathematical algorithm that computes the Discrete Fourier Transform (DFT) of a given sequence breaking it down into its frequency parts. It is distinct from the Fourier Transform (FT) in that takes a discrete signal as input [21].

The process of removing noise from sound is done using a

FFT technique with a set of steps as follows:

a)Load the audio file.

b) Extract audio samples from the audio file.

c)Using FFT to reduce noise in an audio file.

d) Reassemble and reformat an audio file.

VI. MERGE ENHANCED VIDEO FRAMES AND ENHANCED AUDIO

Following the completion of the enhancement process in the preceding steps, the final stage is the merging process, in which both the enhanced video frames and the enhanced audio are joined to form an improved video. Finally, merge audio with frames. Figure 6, shows the pseudo-code for merge frames and audio.

Pseudo-code for merge enhanced video frames and enhanced audio Input:

- Vid_frames = Input Video frames.
- Audio_file = Input Audio file **Steps:**

1- Read Vid_frames and Audio_file.
2- while has frame in (Vid_frames)
a. read frame (Vid_frames) and store in variable (mov).
b. take another frame.
3- end
4- Get number of frames from (Vid_frames) property and assign as (nFrames).
5- Write output merge video as Vid_merge.
6- Assign (Vid_merge) as equal frame rate property form (Vid_frames)
7- Calculate variable (x) as next equation x = floor(size(Audio_file)/nFrames);
8- For every frames in total frames (nFrames) as z
a. Get every frame as (fr_var) from (mov)
b. Calculate the duration (data_audio) in audio file for every frame by the next equation data_audio =
Audio_file (x*(z-1)+1 : x*z,:)
c. Write (fr_var, data_audio) in output video merge (Vid_merge),
9- end **Output:**

- Vid_merge = Output video after merge frames and audio.

Figure 6. Pseudo-code for merge enhanced video frames and enhanced audio

A. Speech Recognition (convert speech to text)

This proposed system's model corresponds to the inferenceonly wav2vec 2.0 base model with a fine-tuning split of 960 hours. The LibriSpeech dataset was used to train it. A feature encoder, a positional encoder, a context network, and a decoder comprise the wav2vec 2.0 inference path. Figure 7 depicts the procedures involved in speech-to-text detection.

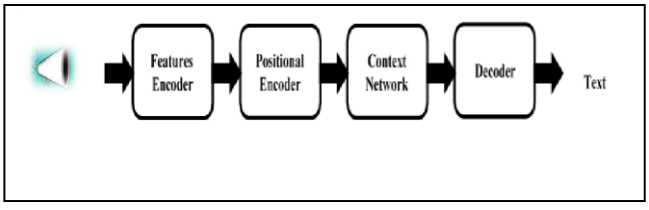

Figure 7. Extractive text from speech

The raw audio input is routed via seven 1-D convolutional blocks by the feature encoder. Between the convolution and GELU activation layers in the first block is an instance (channelwise) normalization layer. Layer normalization is applied to the output of the convolutional blocks. The total context of the encoder receptive field is 400 samples, which translates to 25 ms at a sample rate of 16 kHz.

The positional encoder runs the latent features through grouped 1-D convolution to generate a relative positional vector, which is then added to the latent features to encode their relative location to one another.

The context network employs twelve encoder blocks sequentially. Each block employs multi-head attention and sequentially feeds forward blocks. The feedforward block is made up of two fully connected layers that are separated by a GELU layer.

Each context network block contains three linear (fullyconnected) layers that output the Q (query), K (key), and V (value) vectors. The Q, K, and V vectors are then chunked into twelve non-overlapping pieces (the number of heads in the system). Scaled dot-product attention is applied to each piece individually, and the results are concatenated. The output is routed through a linear layer to create multi-head attention.

A greedy decoding technique is used to sample the most likely tokens at each time step during text decoding. Because the model was trained using connectionist temporal classification (CTC), it requires post-processing to remove repeated blank tokens. This is the most basic decoding approach, with a minimum language model [23].

B. Extractive Text Summarization

The proposed system uses n-grams, a series of n adjacent letters or words, to extract suitable titles for educational videos. Unigrams are used if numerical prefixes are used size one, while bigrams are used if prefixes are used size two. The TextRank was used in the proposed system for summarizing text extracted from audio files. The TextRank algorithm is a graph-based algorithm that is commonly used for text summarization. The algorithm uses a graph representation of the input text, where nodes represent sentences and edges represent the similarity between sentences [29, 35]. Figure 8, shows text summarization steps.

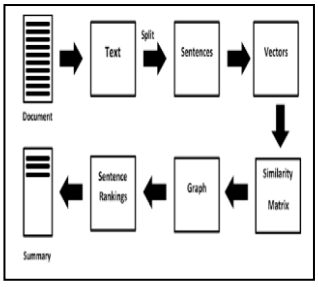

Figure 8. Extractive text summarization process

VII. RESULTS AND DISCUSSION

Experiments were performed using MATLAB version R2022a for the improving educational video system [36]. The experiment is done using a computer system having the following specifications of hardware and software respectively, Intel® Core™ i5-2430M CPU, 4GB RAM, and the operating system is Windows 10 Home Edition 64 bits.

This study seeks to improve educational films in order to save learners' time and resources providing quick and easy searching and indexing processes. It also uses video and subtitle files from the "EDUVSUM" educational videos. The proposed system's phases are implemented on a database that can be downloaded from next URL(http://doi.org/10.5281/zenodo.4002958) [37]. Table 1 shows more details for the used dataset. Figure 9 shows the samples for videos in dataset in these proposed systems.

Table 1.
Show details for dataset

Dataset Name	Number of videos	Videos Size	Videos Duration
EDUVSUM	98 videos	Between 1.8 MB and 114.1 MB	Between 28 sec to 14:52 min

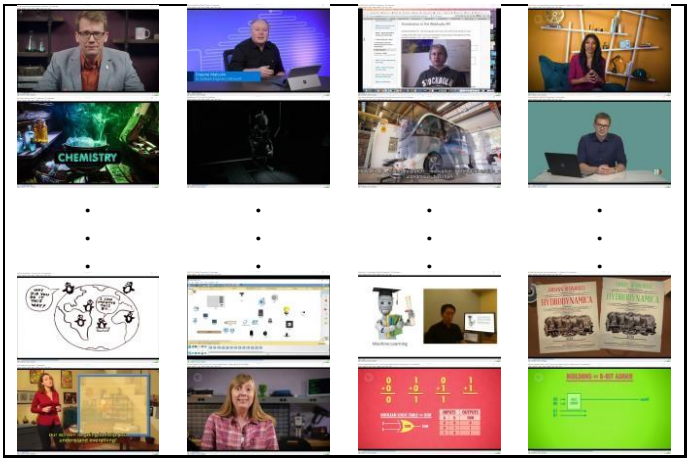

Figure 9. Dataset Samples

In the proposed system, Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Metric (SSIM) [38, 39] are used for enhanced video frames [17]. The PSNR value affects the MSE value with the value of the PSNR increase indicating a clearer image. A peak signal- to -noise ratio of ∞ results in zero mean square error and the resulting value of structural similarity becomes highest [40]. Signal-to-noise ratio (SNR) for de-noising audio, Word Error Rate (WER) for extracted text, and ROUGE score for text summarization.

- Mean Square Error (MSE): is a popular image metric estimator with values closer to zero being better. MSE between two video frames such as $X_{i,j}$ $Y_{i,j}$ is defined as

$$MSE \triangleq \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{i,j} - X_{i,j}\right) \quad (1)$$

where $X_{i,j}$ in the original video frame and $Y_{i,j}$ the enhanced video frame [39, 41].

- Peak Signal-to-Noise Ratio (PSNR): is the ratio of signal to noise power represented using a logarithmic decibel scale. It is calculated using the following process.

$$PSNR = 10.\log\left(\frac{MAX_i^2}{MSE}\right) \quad (2)$$

where MAX_i is the maximum value a pixel can take (e.g. 255 for 8-bit images) and the MSE [39, 40].

- Structural Similarity Index Metric (SSIM): assesses image and video quality by evaluating brightness, contrast, and structure similarity between the original and recovered frame images. The SSIM index is defined as:

$$SSIM = \left(\frac{(\mu_2^2\mu_x\mu_y^2{}_y + C_1^1)((\sigma_2\sigma_x^2\sigma_x^2 + \sigma\sigma_y^2{}_y + C_2^2))}{}\right)_x \quad (3)$$

where μ and σ represent the average and the standard deviation of the original video frame X and the improved video frame Y, respectively. The covariance between X and Y is σ_{XY} that is the covariance. Constants C_1 and C_2 prevent numerical instabilities [39, 40].

- Signal-to-noise ratio (SNR): is a scientific and engineering metric that compares the intensity of a desired signal to the amount of background noise. SNR is the signal-to-noise power ratio, which is often expressed in decibels. A ratio greater than 1:1 (more than 0 dB) indicates that the signal outnumbers the noise. The SNR is defined as follows:

$$SNR = 10 \times \frac{Power^{signal}}{Power_{noise}} \quad (4)$$

where the signal-to-noise ratio in decibels (dB) is defined as 10 times the logarithm of the ratio of the power of a standard signal to the noise power of the audio made by its self-noise [42].

- Word Error Rate (WER): is a commonly used metric for assessing the performance of a speech. It is defined as follows and is based on the Levenshtein distance:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (5)$$

$$WER_{Accuracy} = 1 - WER \quad (6)$$

where S represents the number of substitutions, D represents the number of deletions, I represents the number of insertions, C represents the number of correct words, and N is the total number of words in the reference (N=S+D+C) [43].

- ROUGE score: use the ROUGE similarity score is used to evaluate summarization. ROUGE-N counts the number of Ngrams that match between the model-generated text and a human-produced reference. The ROUGE score is defined as follows:

$$ROUGE - N = \sum_{S \in refsum} \sum_{gramn \in S} count_{match}(gram^n) \quad (7)$$

$$\sum_{S \in refsum} \sum_{gramn \in S} count\,(gramn)$$

where the values of N are 1 and 2. ROUGE_1 and ROUGE_2 represent the overlap of 1-gram and bi-gram summaries between the system and sample summaries [35].

The results, as shown in Table 2, indicate the percentage of MSE, PSNR, and SSIM, which shows the effect of the improvement on video frames, increased lighting quality, and reduced noise. The results, as shown in Table 3, refer to the ratio of SNR to illustrate the reduction of noise associated with the sound. The results of Table 4, indicate the accuracy of the text recognized and extracted from the audio. The Results, as shown in Table 5, refer to summarize the text extracted from the audio

Table 2.
Show PSNR, MSE, and SSIM for video frames enhancement

Modal	Method	Avg / PSNR	Avg / MSE	Avg / SSIM
Proposed System	**CLAHE**	19.738	6.96	80.950
	Remove Haze + CLAHE	21.584	4.55	81.965
	Adjust Intensity + CLAHE	26.581	1.44	93.403
	Remove blur + CLAHE	27.634	1.13	94.344

Table 3.
Show SNR for noise reduction

Modal	Method	Avg / SNR
[21]	FFT	44.2
Proposed System	FFT	45.8

Table 4. Shows WER for Convert speech to text

Modal	Labeled Data	Method	Avg / WER

[23]	Librispeech	Wev2vec	4.8/8.2
[25]	Audio data (without labels) of WSJ + baseline 960 hours modal	Wev2vec	16.24
Proposed System	EUDVSUM + baseline 960 hours modal	Wev2vec	4.27

Table 5.
Shows ROUGE score text summarization

Modal	Dataset	Method	Avg / ROUGE score
[27]	EUDVSUM	TF-IDF LSA LDA	0.2846 0.7262 0.8666
[28]	NPTEL lectures	TF-IDF	0.822
Proposed System	EUDVSUM	TextRank	0.93

In the proposed system, the following figures depict how an intelligent system would be represented in the next graphical user interface (GUI). The following screenshot depicts the suggested system in action. Figure 10 depicts the Video tab's contents for processing video frames. Figure 11 depicts the Audio tab's contents for processing audio and eliminating noise accompanying sound. Figure 12 depicts the Audio to Text tab's contents for extracting text from audio. Figure 13 depicts the Text Summarization tab's contents for summarizing the text collected from the audio.

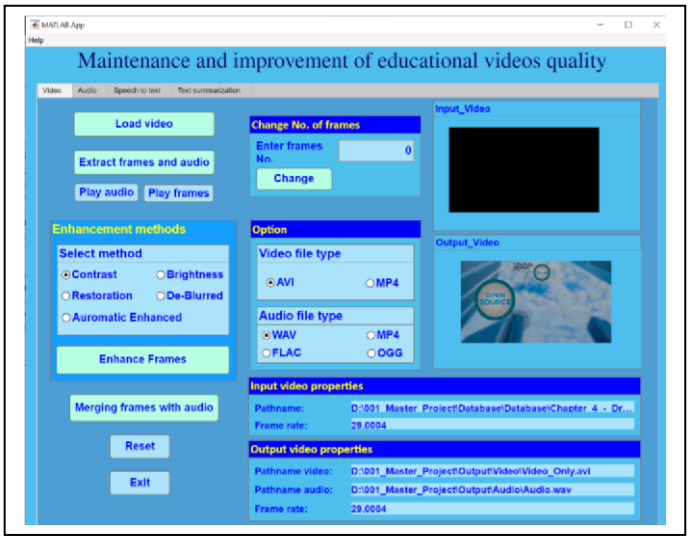

Figure 10. GUI for the proposed system: Video Tab

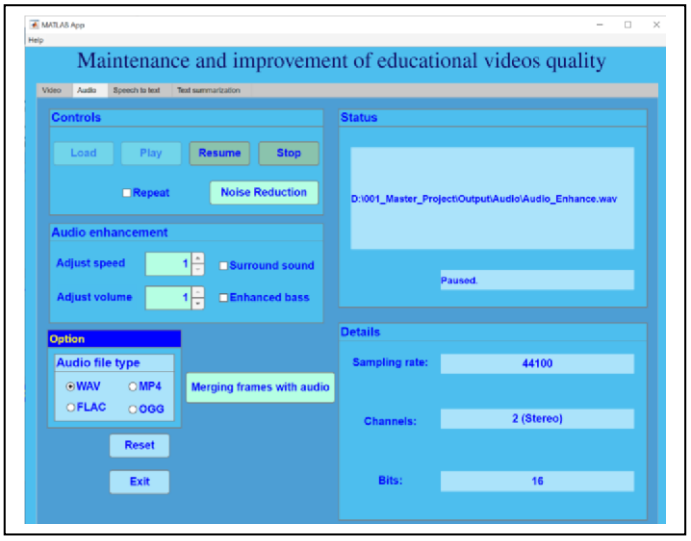

Figure 11. GUI for the proposed system: Audio Tab

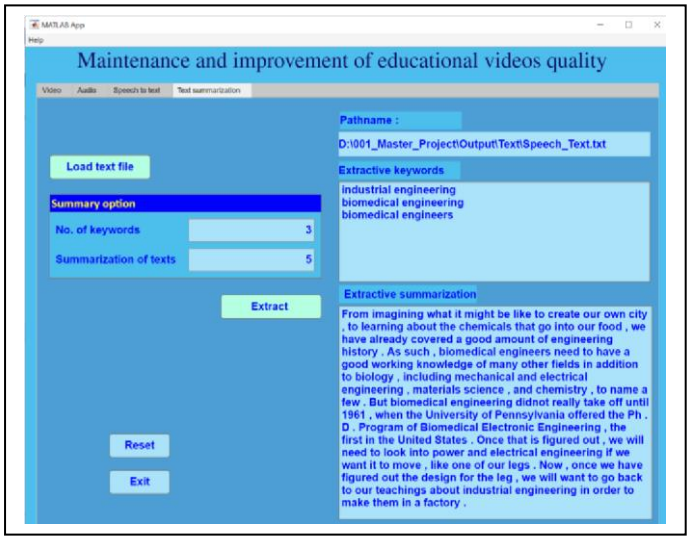

Figure 12. GUI for the proposed system: Speech to Text Tab

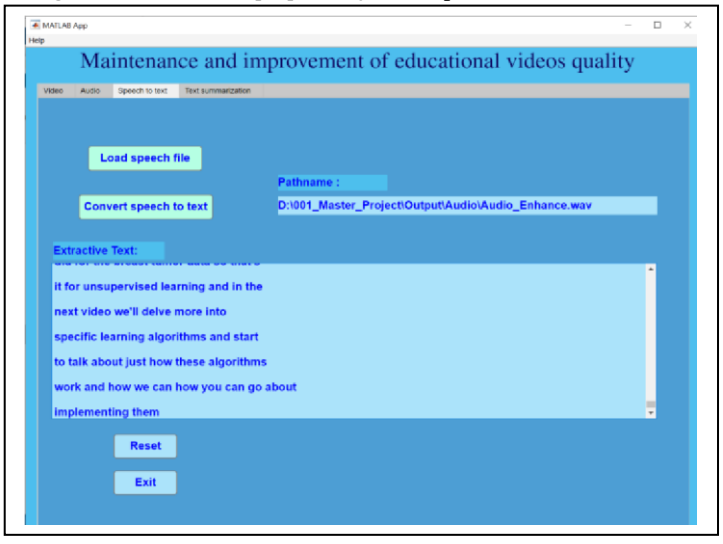

Figure 13. GUI for the proposed system: Text summarization Tab

[3]

[4]

[5]

[6]

[7]

IX.CONCLUSIONS AND FUTURE WORK

Educational videos are exposed to many aspects of problems during recording, which results in poor-quality videos. In the [8] proposed system, an intelligent system for improving educational videos has been presented based on processing four main components: frames, audio, text extracted from audio, and summarizing the text. The proposed system was based on an easy and simple graphical interaction interface. The [9] proposed system with a solution to some problems and an attempt to overcome them. These problems, for which we have provided solutions in processing, such as increasing the 2134-2152, 2023. https://doi.org/10.1080/10494820.2021.1875001.

R. Szeliski, Computer vision: algorithms and applications. Springer Nature, 2022. https://doi.org/10.1007/978-3-030-34372-9.

I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," SN Computer Science, Vol. 2, No. 6, p. 420, 2021. https://doi.org/10.1007/s42979-021-00815-1.

M. Kim, "Document Summarization via Convex-Concave Programming.," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 16, No. 4, pp. 293-298, 2016. https://doi.org/10.5391/IJFIS.2016.16.4.293.

D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," Multimedia tools and applications, Vol. 82, No. 3, pp. 3713-3744, 2023. https://doi.org/10.1007/s11042-022-13428-4.

N. L. Andriyani and N. W. Suniasih, "Development of learning videos based on problem-solving characteristics of animals and their habitats contain in IPA subjects on 6thgrade," Journal of Education Technology, Vol. 5, No. 1, pp. 37-47, 2021. https://doi.org/10.23887/jet.v5i1.32314.

S. Tuncer, B. Brown, and O. Lindwall, "On pause: How online instructional videos are used to achieve practical tasks," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, pp. 1-12, 2020. https://doi.org/10.1145/3313831.3376759.

B. Mbouzao, M. C. Desmarais, and I. Shrier, "Early prediction of success in MOOC from video interaction features," in 21st International Conference, International Conference on Artificial Intelligence in Education, AIED 2020. , Ifrane, pp. 191-196, 2020.

of Educational Technology in Higher Education, Vol. 17, No. 1, pp. 1-18, 2020.

lighting within the video, removing blur, as well as reducing

the noise associated with the sound in the educational video.

This proposed system also presents a proposal to extract the text from the audio and summarize this text to make a summary of the video that can be published with the video in cases of e-learning. The performance results of the proposed system were

The prospect of developing the current system for use during live broadcasts in e-learning processes is part of future work.

In addition, different techniques for live broadcasting have been added. constant bit rate video quality enhancement,

VIII. **CONFLICT OF INTEREST**

No potential conflict of interest relevant to this article was reported.

REFERENCES

https://doi.org/10.1007/978-3-030-52240-7.

[10] O. Voronkin, "Educational video in the university: Instruments, technologies, opportunities and restrictions," in Proc. 15th Int. Conf. ICT in Education, Research, and Industrial Applications. ICTERI 2019, Kherson, pp. 302-317, 2019. http://ceur-ws.org/Vol-2387/.

[11] Q. Ding, L. Shen, L. Yu, H. Yang, and M. Xu, "Patch-wise high. spatial-temporal quality enhancement for HEVC compressed video," IEEE Transactions on Image Processing, Vol. 30, pp. 6459-6472, 2021. http://doi.org/10.1109/TIP.2021.3092949.

[12] H. Zhao et al., "CBREN: Convolutional neural networks for Transactions on Circuits and Systems for Video Technology, Vol. 32, No. 7, pp. 4138-4149, 2021. http://doi.org/10.1109/TCSVT.2021.3123621.

[13] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, pp. 9700-9709, 2021. http://doi.org/10.1109/ICCV48922.2021.00956.

[1] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," Artificial Intelligence Review, Vol. 56, No. 3, pp. 2193-2240, 2023. https://doi.org/10.1007/s10462-022-10224-2.

[2] K.-Y. Tang, C.-Y. Chang, and G.-J. Hwang, "Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019)," Interactive Learning Environments, Vol. 31, No. 4, pp. https://doi.org/10.1186/s41239-020-00190-6.

[14] L. Liu et al., "Low-light video enhancement with synthetic event guidance," in Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, pp. 1692-1700, 2023. https://doi.org/10.1609/aaai.v37i2.25257.

[15] C. Lange and J. Costley, "Improving online video lectures: learning challenges created by media," International Journal

[16] A. Abd El-badie Abd Allah and F. Abd El-Sattar Zahran, "A Fuzzy Decision Support System for Diagnosis of Some Liver Diseases in Educational Medical Institutions," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 20, No. 4, pp. 358-368, 2020. https://doi.org/10.5391/IJFIS.2020.20.4.358.

[17] B. Veerasamy and C. M. Sangeetha, "MB-FL: MacroBlock Fuzzy Logic for Video Compression in Multimedia Applications," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 22, No. 4, pp. 366-372, 2022. https://doi.org/10.5391/IJFIS.2022.22.4.366.

[18] J. A. Qadir, A. K. Al-Talabani, and H. A. Aziz, "Isolated spoken word recognition using one-dimensional convolutional neural network," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 20, No. 4, pp. 272-277, 2020. https://doi.org/10.5391/IJFIS.2020.20.4.272.

[19] S. Hong, M. Kim, and M. G. Kang, "Single image dehazing via atmospheric scattering model-based image fusion," Signal Processing, Vol. 178, No. 107798, pp. 1-14, 2021. https://doi.org/10.1016/j.sigpro.2020.107798.

[20] D. N. H. Thanh, N. M. Hue, and V. B. S. Prasath, "Single image dehazing based on adaptive histogram equalization and linearization of gamma correction," in 2019 25th AsiaPacific Conference on Communications (APCC), Ho Chi Minh City, pp. 36-40, 2019. http://doi.org/10.1109/APCC47188.2019.9026457.

[21] R. Senthamizh Selvi, "Speech enhancement using adaptive filtering with different window functions and overlapping sizes," Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol. 12, No. 13, pp. 1886-1894, 2021. https://doi.org/10.17762/turcomat.v12i13.8841.

[22] F. Khan and A. Chouksey, "Design of audio enhancement using deep learning for adaptive filter (DLAF)," International Research Journal of Modernization in Engineering Technology and Science, Vol. 2, No. 8, pp. 1474-1482, 2020, Available. https://www.irjmets.com/uploadedfiles/paper/volume2/iss ue_8_august_2020/3096/1628083119.pdf.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing System, Vancouver, pp. 12449-12460, 2020. https://dl.acm.org/doi/10.5555/3495724.3496768.

[24] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in Proc. Interspeech 2021: International Speech Communication Association (ISCA 2021), Brno, pp. 15091513, 2021. http://dx.doi.org/10.21437/Interspeech.2021-1280.

[25] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in Proc. Interspeech 2019: International Speech Communication Association (ISCA 2019), Graz, pp. 3465-3469, 2019. http://dx.doi.org/10.21437/Interspeech.2019-1873.

[26] D. V. Jose, "Speech to text conversion and summarization for effective understanding and documentation," International Journal of Electrical and Computer Engineering (IJECE), Vol. 9, No. 5, pp. 3642-3648, 2019. http://doi.org/10.11591/ijece.v9i5.pp3642-3648

[27] S. S. Alrumiah and A. A. Al-Shargabi, "Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement," Computers, Materials & Continua, Vol. 70, No. 3, pp. 6205-6221, 2022. http://doi.org/10.32604/cmc.2022.021780.

[28] R. K. Abhilash, C. Anurag, V. Avinash, and D. Uma, "Lecture video summarization using subtitles," in 2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing, Springer, pp. 83-92, 2021. https://doi.org/10.1007/978-3-030-47560-4_7.

[29] V. B. Aswin et al., "NLP-driven ensemble-based automatic subtitle generation and semantic video summarization technique," in Proceedings of AIDE 2019: Advances in Artificial Intelligence and Data Engineering. Advances in Intelligent Systems and Computing, Singapore, pp. 3-13, 2021. https://doi.org/10.1007/978-981-15-3514-7_1.

[30] A. Dilawari and M. U. G. Khan, "ASoVS: abstractive summarization of video sequences," IEEE Access, Vol. 7, pp. 29253-29263, 2019. http://doi.org/10.1109/ACCESS.2019.2902507.

[31] B. S. Omarov, A. B. Altayeva, and Y. Im Cho, "Exploring image processing and image restoration techniques," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 15, No. 3, pp. 172-179, 2015. https://doi.org/10.5391/IJFIS.2015.15.3.172.

[32] J. Lee, S. R. Pant, and H.-S. Lee, "An adaptive histogram equalization based local technique for contrast preserving image enhancement," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 15, No. 1, pp. 35-44, 2015. https://doi.org/10.5391/IJFIS.2015.15.1.35.

[33] P. K. Verma, N. P. Singh, and D. Yadav, "Image enhancement: a review," Ambient Communications and Computer Systems: RACCCS 2019, Vol. 1097, pp. 347-355,

2020 https://doi.org/10.1007/978-981-15-1518-7_29.
[34] R. Singh and S. Bansal, "A comparative study of image deblurring techniques," Journal of Computational and Theoretical Nanoscience, Vol. 17, No. 9-10, pp. 4571-4579, 2020. https://doi.org/10.1166/jctn.2020.9282.
[35] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-based text summarization using modified TextRank," in Proceedings of International Conference on SCDA 2018: Soft Computing in Data Analytics . Advances in Intelligent Systems and Computing, Springer,Singapore, pp. 137-146, 2019. https://doi.org/10.1007/978-981-13-0514-6_14.
[36] L. Z. Sansyzbay, B. B. Orazbayev, and W. Wójcik, "Development and Analysis of Models for Assessing Predicted Mean Vote Using Intelligent Technologies," International Journal of Fuzzy Logic and Intelligent Systems, Vol. 20, No. 4, pp. 324-335, 2020. https://doi.org/10.5391/IJFIS.2020.20.4.324.
[37] J. A. Ghauri, "Annotated educational videos and subtitles (EDUVSUM)," ed, 2020. http://doi.org/10.5281/zenodo.4002958.
[38] J. Søgaard, L. Krasula, M. Shahid, D. Temel, K. Brunnström, and M. Razaak, "Applicability of existing objective metrics of perceptual quality for adaptive video streaming," in Electronic Imaging, Image Quality and System Performance XIII, San Francisco, pp. 1-7,2016. https://nantes-universite.hal.science/hal-01395510.
[39] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study," Journal of Computer and Communications, Vol. 7, No. 3, pp. 8-18, 2019. http://doi.org/10.4236/jcc.2019.73002.
[40] D. Fuoli et al., "AIM 2020 challenge on video extreme super-resolution: methods and results," in Proceedings of the 16th European Conference on Computer Vision, Glasgow, pp. 57-81, 2020. https://doi.org/10.1007/978-3-030-66823-5_4.
[41] S. Nah et al., "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, pp. 19962005, 2019. http://doi.org/10.1109/CVPRW.2019.00251.
[42] K. F. A. Darras et al., "High microphone signal-to-noise ratio enhances acoustic sampling of wildlife," PeerJ, Vol. 8, No. 9955, pp. 1-21, 2020. http://doi.org/10.7717/peerj.9955.
[43] B. Schuetz and N. Aschenbruck, "SPQER: Speech Quality Evaluation Using Word Recognition for VoIP Communication in Lossy and Mobile Networks," IEEE Open Journal of the Computer Society, Vol. 1, pp. 145154, 2020. http://doi.org/10.1109/OJCS.2020.3011392.

Bassant Mohamed Elamir (B. M. Elamir) received the Bachelor's degree computer science, Faculty of Specific Education, Mansoura University, Egypt, in 2017. She is a master's student in Artificial Intelligence Techniques in Education at Faculty of Specific Education, Mansoura University, Egypt. She works teaching assistant at Mansoura University, Egypt. Her research interests educational videos include intelligent system, image and sound processing, speech recognition, nature language processing.
E-mail: basantelamir@mans.edu.eg

Amany Fawzy Elgamal (A. F. Elgamal) , professor of computer science at Faculty of Specific education, Mansoura University, Egypt. She got her master and PHD from Faculty of Engineering, Mansoura University. Scientific areas of her interest include: programming languages, Artificial Intelligent and its application in education. She has published many research papers and books in Arabic and English in areas related to teaching programming and computer programs software, as well as applications of artificial intelligence in education.
Email: amany_elgamal@mans.edu.eg

Marwa Hussein Abdelfattah (M. Hussein) is lecturer in department of computer science, Faculty of Specific Education, Mansoura University, Egypt. Scientific areas of her interest include: Artificial Intelligent, machine learning, natural language processing and image processing.
E-mail: marwahussien@mans.edu.eg