Design of Voice Style Detection of Lecture Archives

Xiaoting Liu
Division of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
liuxiaoting@jaist.ac.jp

Wen Gu
Center for Innovative Distance Education and Research,
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
wgu@jaist.ac.jp

Koichi Ota
Center for Innovative Distance Education and Research,
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
ota@jaist.ac.jp

Shinobu Hasegawa
Center for Innovative Distance Education and Research,
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
hasegawa@jaist.ac.jp

***Abstract*—Due to the COVID-19 pandemic, most universities endeavored to adopt online education as an alternative to conventional face-to-face classroom instruction. However, capturing students' Temporal Region of Interest (T-ROI) in long-duration video lectures poses a significant challenge. Therefore, lecture archive summarization becomes essential from an online perspective. The results of lecture archive summarization still require further improvement. This research aims to distinguish T-ROI using a speech processing approach h. Our plan is divided into collecting instructors'/presenters' voice datasets, clarifying the T-ROIs through sound processing technology, and building a suitable deep neural network architecture to detect the T-ROIs in the actual lecture archives automatically. We will inevitably encounter various challenges to achieve the objective, such as individual differences. This article describes the experimental dataset collection design considering individual differences and lecture room environments. It summarizes how such efforts will be effective in realizing personalized voice style detection and improving the accuracy of speech processing in real environments.**

Keywords— Voice Style, Lecture Archives, Dataset, Deep Learning, T-ROI

I. INTRODUCTION

Distance education, which originated in the 19th century, has experienced a resurgence in the 21st century, fueled by the rapid advancements in computer and internet technologies. The introduction of Open Educational Resources (OER) and Massive Open Online Courses (MOOCs) has expanded educational accessibility to a broader audience [1]. Furthermore, in response to the global COVID-19 pandemic, UNESCO reported that around 107 countries had implemented complete school closures, impacting approximately 860 million school-age children, which accounts for nearly half of the world's school-age population [2]. The position of distance education before the COVID-19 pandemic was characterized as occasional and non-routine learning, providing opportunities for learning and experiences that were not typically available in regular classrooms. However, many schools have attempted to utilize distance education as a substitute for traditional classrooms to reduce the risk of infection [3]. In other words, school activities normally take place are now being conducted entirely through distance education. Distance education with "anytime, anywhere, with anyone" might be anticipated to be utilized in routine and non-routine learning aspects in the future "post-COVID" era.

One powerful application of distance education is lecture archives, providing online asynchronous delivery of recorded lectures. Choe et al. investigated the impact of various lecture archives on student engagement, satisfaction, and learning outcomes [4]. The findings indicated that multimedia learning could effectively engage learners. However, it is essential to consider the time and cost required to accommodate all lectures. Therefore, this study focuses on lecture archives in which face-to-face lectures are recorded in fixed equipment. The Japan Advanced Institute of Science and Technology (JAIST), for example, more than lecture archives recorded from face-to-face lectures have been stored and delivered since April 2006 as supplementary learning materials through their learning management system (LMS) [5]. These archives capture both the voice of the instructor and the students. However, it might be necessary to support a sense of engagement with the students in learning with such archives. Unlike face-to-face lectures, where instructors can interact with students in real-time, delivering 100-minute lecture archives 'as is' without any modifications does not adequately maintain learner engagement. In other words, students should struggle to stay focused and attentive throughout lengthy archives in an online environment. Therefore, lecture archive summarization becomes essential from an online learning/education perspective.

Lecture archive summarization is necessary to improve learning with archives, which are boring with little interaction, and to preserve important sections within lectures. One of the essential keywords in archive summarization is Temporal ROI (Region of Interest) which means scenes or periods that learners should watch in the archives. Sheng et al. provided focal periods of the lecture archives as T-ROIs using the instructor's movements, voice intensity, and slide differences [6]. However, several limitations should be improved. Firstly, the original sound in the archives includes room noise, which does not contribute to T-ROI detection and may introduce unwanted distractions. Secondly, specific features derived from the original sound can also play an important role in accurately identifying T-ROI. For example, teachers modulate their voice to emphasize style when important knowledge is coming[7]. Similarly, teachers tend not to use emphasis voice

in non-important parts. It is reasonable to think that the change in voice style is often related to the difference in the importance of the content taught. This also means that teacher's voice style is strongly associated with T-ROI. Based on these considerations, this study focuses on voice style as a "way" to express something in communication [8].

We hypothesize that instructors change their voice style when discussing T-ROI parts during lectures. Unfortunately, the previous study did not cover sound processing enough since estimating appropriate sound features from the lecture archives is complicated due to the positional relationship between the recording units and the instructors, the noise in the lecture room, etc. Therefore, the available speech dataset for estimating T-ROIs during lectures is also inadequate. Consequently, it is necessary and challenging for further research in the area of automatic instructors' voice style detection.

The following article is organized as follows: Chapter II introduces the related work about lecture archive summarization and voice datasets; Chapter III claims the objective and difficulties of this research; Chapter IV describes how to implement our approach; Chapter V proposes the experimental design for data collection; Chapter VI investigates the initial analysis results; Chapter VII discusses possible challenges in this research; and Chapter VIII summarizes this study and emphasizes its contributions.

II. RELATED WORK

A. Lecture archive summarization

Lecture archive summarization has undergone important advancements over the years. Hulens et al. presented an automated system designed to record lecture archives in classrooms [9]. The system employed specialized hardware, including cameras and microphone arrays, to capture multiple viewpoints of the lectures, enhancing the learning experience through digital zooming and improving recording quality with audio source localization and face detection for audience interaction. Andra et al. proposed an automatic summarization method for lecture archive transcripts using an attention-based Recurrent Neural Network (RNN) [10]. The approach incorporates segmentation and linguistic-based features to enhance topic coherency and improve summary quality. Experimental results show significant improvement in the ROUGE score compared to baseline models. However, their approach depends on the semantic analysis of the lecture content. Such text is not always available for lecture archives. Chand et al. and Soares et al. respectively analyzed the sound characteristics to get the text information and divide the instructional video according to the text information content [11] [12]. However, for mathematics teaching content such as number formula, it is difficult to accurately divide through text information only. Our previous study proposed a deep neural network architecture for detecting focal periods as T-ROIs in lecture archives [6]. The research focused on leveraging combined features of image and sound within the pre-trained deep neural network architectures to improve the reliability of T-ROI detection. This approach enables the detection of specific focal periods using minimal computational resources without relying on semantic features. However, there is still potential for further accuracy improvement, especially in sound processing since they used voice intensity and did not consider detailed sound features.

B. Voice datasets

The key to a successful speech processing approach is the speech data set. The "interactive emotional dyadic motion capture database" (IEMOCAP) is a well-known corpus used for emotion estimation from speech, collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC) [13]. It includes recordings of ten actors engaged in dyadic sessions, meticulously capturing their facial expressions and hand movements during both scripted and spontaneous communication scenarios. The actors performed emotional scripts and improvised scenarios to evoke various emotional states, including happiness, anger, sadness, frustration, and a neutral state. With approximately 12 hours of data, the IEMOCAP corpus provides valuable resources for studying multimodal and expressive human communication. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a validated multimodal database with recordings from 24 professional actors [14]. It includes emotions expressed in speech and song formats, covering a wide range of emotions. The database offers various options for stimuli selection and contains many recordings. It has been thoroughly evaluated for emotional validity and reliability. This dataset is freely available under a Creative Commons license, making it a valuable resource for emotion-related research. These studies are worth datasets for estimating emotion from speech. However, it is not obvious which emotional expressions indicate important lecture content as T-ROIs. In addition, although there are many archives of lectures and presentations, they are not labeled to indicate the important parts. Therefore, constructing a dataset that can be used for estimating T-ROIs as the voice style in lectures or presentations is a crucial element of this research.

III. OBJECTIVE

Our final goal is to distinguish T-ROIs from the sound processing approach by defining instructors' voice styles, analyzing sound features of different voice styles to clarify the relationship between voice styles and T-ROIs with the ideal and actual sound environment, and building a suitable deep neural network architecture to detect T-ROIs in actual lecture archives automatically. This research can help to improve the accessibility and provide quick access to the T-ROIs of the archives as a part of lecture archive summarization.

To achieve this goal, we proceed with the following research questions/challenges:

1) *Research Question 1:* How to define/collect instructors'/presenters' voice style data in the lecture? Challenge 1: Collecting voice style data with appropriate labels is difficult.

2) *Research Question 2:* How can the relationship between the T-ROIs and the voice styles be clarified? Challenge 2: Relating the sound features and the limited voice style data is difficult. Challenge 3: Different instructors have different voice styles as individual differences.

3) *Research Question 3:* How to apply the proposed method to the actual lecture archives? Challenge 4: Very few research focus on speech in the real environment. It is hard to refer to any solution to cover the actual situation.

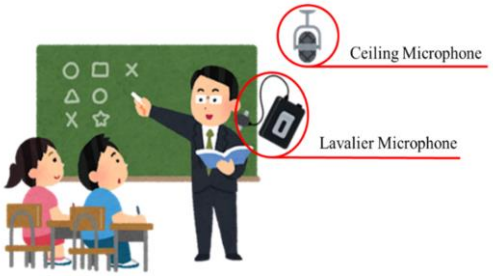

Fig. 1. The concept plan of experimental environment.

IV. METHODOLOGY

In response to the aforementioned specific questions / challenges, we have devised the following corresponding methods at the current stage.

A. Data Collection

Based on our current investigation, we have yet to find an existing database regarding teachers' speaking styles. Therefore, we plan to collect data on different voice styles to establish a database through experiments.

Regarding Research Question/Challenge 1, we plan to record participants' speaking styles as the data collection experiment during lecture/presentation interactions. In lectures, instructors use various expressions to teach the content more accurately. One of them is voice style. We distinguished it according to what is important, neutral, or unimportant in their talk rather than voice emotion as described in related work. This is because we believe the relationship between emotion and the importance of what is taught is not necessarily one-to-one.

Each participant will be asked to speak sample sentences and their own topics in different (important, neutral, and unimportant) styles. The experiment environment should follow the actual lecture condition, as shown in Fig. 1. It uses the ceiling microphone for the actual environment voice and the lavalier microphone for pure instructors' speech. The detailed content is explained in Chapter V. In the future application, we will use both microphones, but if we want to use the existing lecture video data, we usually cannot use the lavalier microphone's data. Because of the environmental noise, the ceiling microphone cannot collect pure voice. Kaur et al. processed the audio signals of lecture video to identify the location and duration of various types of noise [15]. Therefore, we plan to draw on Kaur et al. 's research and collect two databases using two microphones. This will help to modify / improve the data quality of ceiling microphones for future applications, making them more suitable for real lecture archives which only have one microphone.

B. Relationship Analysis

Chand et al. and Soares et al. are using loudness, pitch, and pause rate to divide the instructional video [11] [12]. As an initial stage of research, we also decided to analyze these characteristics. For example, for loudness, we analyzed the short-term energy of each voice style for some participants who participated in the data collection experiment.

By detecting the teacher's voice style, the lecture archives can be segmented into different levels of importance. This enables students to efficiently access and retrieve their T-ROIs.

Regarding Research Question/Challenge 2, we need to identify a viable approach to analyze the relationship between the T-ROIs and the voice styles. Our current limited research has not found related studies on the classification of voice styles in actual lectures. However, the features used in studies of voice emotion estimation, such as MFCC and SVM, may be applicable to the estimation of teaching styles. One of the related research is using MFCC (Mel Frequency Cepstral Coefficients) and SVM (Support Vector Machine to classify the voice emotion[16]. Our research aims to classify voice styles in an actual lecture. Therefore, we plan to apply MFCC-SVM as a baseline to detect the voice style. Voice style data will be input to MFCC to gather the voice spectral features. Then, SVM performs voice style classification by processing the voice spectral features.

Additionally, since this technique has not been thoroughly validated and applied in real-world environments, its applicability to actual teaching scenarios may be limited.

As shown in Fig. 2, we plan to relate the sound features (including pitch, tempo, and loudness) and the voice styles by using regression statistics to establish a regression equation and deep learning methods like voice transformers to correspond non-linear relationships. The above steps will provide assistance in addressing Challenge 3 in the future.

C. Multi-Stream Network

Fig. 3 shows how to apply the proposed method to the actual lecture archives. The input consists of sounds recorded by the lavalier and ceiling microphones. First, speech data are divided into 1-minute segments to increase data. Then, the

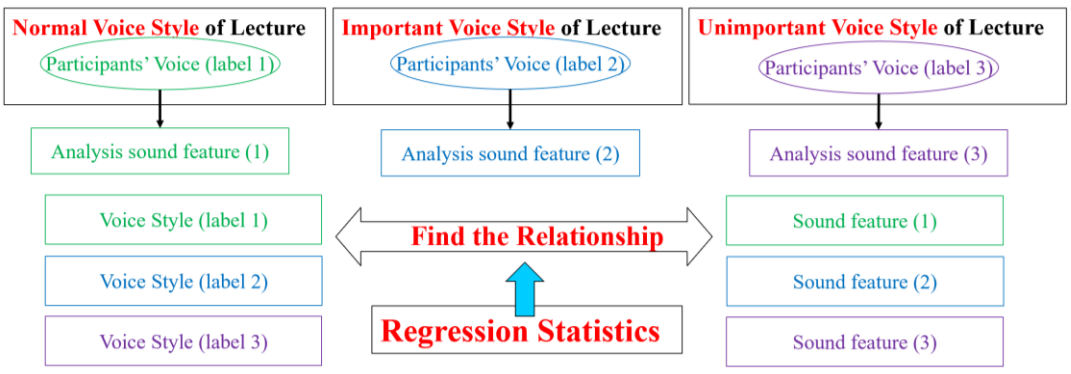

Fig. 2. Overall of relation analyze.

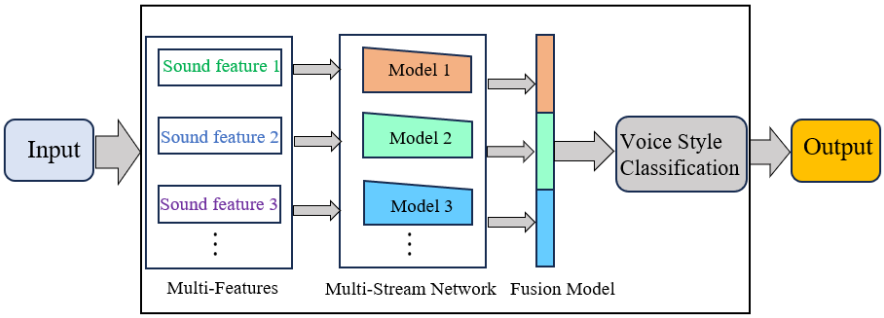

Fig. 3. Feature-based fusion model.

multiple stream feature-based fusion model analyzes the sound features extracted from the segments. Considering the result of the previous step, compensate for the sound from the ceiling microphone by using the lavalier microphone to extract different sound features. The output is the classification of voice styles, which helps identify and differentiate between different voice styles to detect T-ROIs within the archives.

V. DATA COLLECTION EXPERIMENT DESIGN

In the first step, we intend to conduct an experiment to collect data featuring distinct voice styles to differentiate the T-ROIs in lecture archives. It is essential to gather voice samples in authentic environments to ensure the robustness of our results. Therefore, we require collecting voice data during instructional sessions to examine different voice styles. In addition, for a more accurate analysis of voice characteristics, we aim to collect voice data that exhibits relatively pure voice data at the same time. Our goal is to collect more than two thousand minutes of audio data based on these requirements.

A. Experimental Overview:

The participants will deliver presentations in an environment resembling a lecture room, and their speeches will be recorded. During the speeches, they will be instructed to be mindful of important, neutral, and unimportant content. Each speech will last approximately 20 seconds, with breaks inserted as needed. The primitive voice data is recorded with a duration of 20 seconds. Finally, employing a moving average technique, we process this 20-second voice data to generate voice segments lasting several minutes. The total recording speech duration of around 100 minutes for one participant. The overall estimated duration of the experiment, including informed consent, procedure explanations, speech recordings, breaks, and post-experiment questionnaires, is approximately 150 minutes. After the experiment, the post-experiment questionnaire will be administered to gather information related to the participants' gender, age, English skills, experience in lectures and presentations, consciousness of each voice style, and any observations or insights regarding the experiment. In specific research, collecting voice style data proves to be challenging, as participants of different genders, ages, and English skills may perform varying voice styles. To enhance the data annotation process, we have thus taken into consideration the variables mentioned in the post-experiment questionnaire.

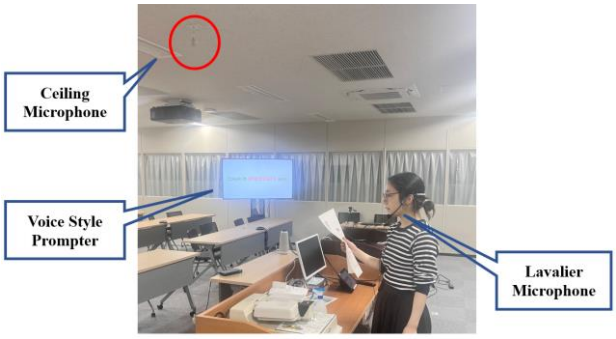

Fig. 4. The experiment settings.

B. Participants and Data Collection Method:

This study aims to analyze voice styles in English lectures and presentations among a target population of 20 students with English communication skills at our university. The participants will deliver presentations in three different voice styles based on specific texts displayed on the screen and their research summaries. The experiment will use two types of microphones: ceiling and lavalier. The experiment setting is depicted in Fig. 4.

C. Data and Variables:

The data collection will involve recorded audio data using lavalier microphones (.wav format) and ceiling microphones (.mp3 format). In addition, a post-experiment questionnaire will be administered to gather information such as gender, age, English skills, experience in lectures and presentations, consciousness of each voice style, and free-text responses.

These experimental procedures have been applied to the JAIST Life Science Committee.

VI. RESULTS / INITIAL ANALYSIS

A. Short-Time Energy Analysis

We analyzed the short-time energy of three voice styles for three participants from the data collection experiment. The comparison results show in Fig. 5. The short-time energy feature shows more noticeable differences in the important voice style compared to the normal and unimportant voice styles. Therefore, the short-time energy feature may be helpful in identifying the important voice style. However, for the normal and unimportant voice styles, their differences in the short-time energy feature are not significant, which suggests that using the short-time energy feature alone may not effectively distinguish between the normal and unimportant voice styles.

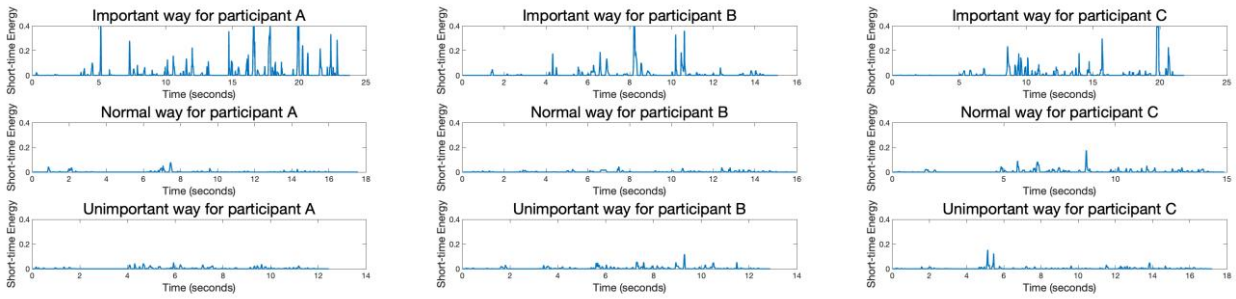

Fig. 5. Short-Time energy of voice styles.

B. *MFCC-SVM Analysis*

Using MFCC-SVM, we classified three voice styles based on recordings captured by the ceiling and lavalier microphones, we obtained accuracies of 0.39 for the ceiling microphone database and 0.46 for the lavalier microphone database are presented in TABLE I. The results indicate that the lavalier microphone database outperforms the ceiling microphone database in terms of accuracy. This difference may be attributed to the lavalier microphone database recording purer speech data, while the ceiling microphone database includes environmental noise and reverberations, which potentially affect the accuracy. Moreover, in the classification of the three voice styles, the accuracy for the normal style is higher than that of important and unimportant styles. Particularly, the accuracy for the unimportant style is the lowest, with a maximum value of only 0.29. Thus, using MFCC-extracted features for recognizing unimportant style data may yield unsatisfactory results. Additionally, as this study is in the initial stage, the available data is limited. The insufficient data may also contribute to suboptimal results.

VII. DISCUSSION

In collecting the data set, we will ask the participants to be aware of the three voice styles. However, we should also investigate whether others can be aware of the differences in the styles. Therefore, we will evaluate the quality of voice style labels after this experiment. In addition, we should consider that speech recorded under instruction or acting conditions may differ from natural speech settings [13]. This raises concerns regarding the collected data's representativeness and applicability to real-world scenarios. Therefore, we need to acknowledge and address this challenge in our study as it can impact the validity and generalizability of our findings.

During actual lectures, such as mathematics and chemistry classes, the understanding of semantics may be relatively limited, especially when dealing with equations and other specialized content. This is another challenge for our future research.

VIII. CONCLUSION

Our study aims to explore the classification of voice styles based on actual lectures. Classifying instructors' voice styles in educational settings can improve lecture archive efficiency and enhance student's learning experience by keeping engagement and better understanding. To achieve our objectives, we need to address several challenges:

Firstly, we recruit participants to gather enough datasets for voice style detection and investigate whether others can perceive these differences. One of the essential points is to facilitate the restoration of participants' natural state as much as possible during data collection. Thus, this article introduces an experimental environment that encourages authentic speech, reducing the impact of instruction or acting conditions. By adopting this strategy, we can enhance the representativeness and applicability of the collected data, thereby ensuring the validity and generalizability of our study findings.

Additionally, we are aware that sound contains rich acoustic features to carry important information. Therefore, we will explore the possibility of speech classification in real lecture room environments through sound processing techniques.

Lastly, this study proposes a machine/deep learning method to estimate the T-ROI based on individual speakers. Many acoustic machine/deep learning models assume that training data and test data are identically distributed. Therefore, it is easy to reflect features common to all speakers. However, it is difficult to reflect the characteristics specific to each speaker. This point enables the realization of personalized models. Findings from this research will contribute to the field of speech analysis and interpretation.

ACKNOWLEDGMENT

This work received support from the Japan Science and Technology Agency and the JSPS KAKENHI Grant Number 23H03506.

TABLE I. RESULTS OF ACCURACY

Model	Dataset	Voice Style	Precision	Recall	F1-score	Accuracy
MFCC-SVM	Ceiling microphone	Important Normal Unimportant	0.36 0.50 0.14	0.33 0.75 0.08	0.35 0.60 0.11	0.39
	Lavalier microphone	Important Normal Unimportant	0.50 0.55 0.29	0.38 0.75 0.25	0.43 0.63 0.27	0.46

REFERENCES

[1] A. Saykili, “Distance education: Definitions, generations and key concepts and future directions,” International Journal of Contemporary Educational Research, vol. 5, no. 1, pp. 2-17, 2018.

[2] S. Abuhammad, “Barriers to distance learning during the covid-19 outbreak: A qualitative review from parents' perspective,” Heliyon, e05482, 2020.

[3] A. Schleicher, “The Impact of COVID-19 on Education: Insights from "Education at a Glance 2020",” Web site: http://www.oecd.org, 2020.

[4] R. C. Choe, Z. Scuric, E. Ehkol and S. Cruse, “Student Satisfaction and Learning Outcomes in Asynchronous Online Lecture archives,” CBE—Life Sciences Education, Vol. 18, No. 4, 2019.

[5] S. Hasegawa, Y. Tajima, M. Matou, M. Futatsudera, and T. Ando, “Case studies for self-directed learning environment using lecture archives,” International Conference on Web-basedEducatio, pp. 299-304, 2007.

[6] R. Z. Sheng, “Automatic focal period detection of lecture archives,” Japan Advanced Institute of Science and Technology, master thesis, March 2022.

[7] X. Y. Che, S. Luo, H. J. Yang, and C. Meinel, “Sentence-Level Automatic Lecture Highlighting Based on Acoustic Analysis,” IEEE International Conference on Computer and Information Technology (CIT), pp. 328–334, 2016.

[8] D. S. Mokere, I. Y. Rawung, “The English Lecturers’ Voice style in Teaching English,” sintuwumarosoJET, Vol. 5, No. 1, August 2019.

[9] D. Hulens, B. Aerts, P. Chakravarty and A. Diba, “The CAMETRON Lecture Recording System: High Quality Video Recording and Editing with Minimal Human Supervision,” Lecture Notes in Computer Science, vol.10704, Springer, Cham, 2018.

[10] M. B. Andra and T. Usagawa, “Automatic lecture archive content summarization with attention-based recurrent neural network,” International Conference of Artificial Intelligence and Information Technology (ICAIIT), IEEE, pp. 54-59, 2019.

[11] D. Chand, H. Oğul, “A Framework for Lecture Video Segmentation from Extracted Speech Content,” 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics.

[12] E. R. Soares, E. Barrére, “An Optimization Model for Temporal Video Lecture Segmentation using Word2vec and Acoustic Features,” Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, pp. 513–520, October 2019.

[13] C. Busso, M. Bulut, C. C. Lee and A. Kazemzadeh, “IEMOCAP:

[14] Interactive emotional dyadic motion capture database,” Language resources and evaluation 42 (2008): 335-359.

[15] R. L. Steven, A. R. Frank, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLoS ONE 13(5): e0196391, 2018.

[16] P. C. Kaur, L. Ragha, “Audio de-noising and quality assessment for various noises in lecture videos,” International Conference on Paradigm Shifts in Communications Embedded Systems, 05-06 April 2023.

[17] H. Aouani, Y. B. Ayed, “Emotion Recognition in Speech Using MFCC with SVM, DSVM and Auto-encoder,” International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 21-24, 2018.

[18] B. Schuller, B. Vlasenko, F. Eyben and M. Wollmer, “Cross-corpus acoustic emotion recognition: Variances and strategies,” IEEE Trans. Affective Computing, vol.1 no.2 pp.119-131, 2010.