Exploring Advances in Meeting Minutes Generation and Face Attendance Systems: A Comprehensive Literature Survey

Prof. Seema Mane[1], Anushka Khadtare[2], Swamini Hingmire[3], Prathamesh Pawar[4], Shritej Vasal[5]

[1]*Department of Information Technology, Sinhgad Institute of Technology and Science*
[2]*Department of Information Technology, Sinhgad Institute of Technology and Science*
[3]*Department of Information Technology, Sinhgad Institute of Technology and Science*
[4]*Department of Information Technology, Sinhgad Institute of Technology and Science*
[5]*Department of Information Technology, Sinhgad Institute of Technology and Science*

Abstract - The existing literature survey offers a thorough exploration of automatic text summarization, speech-to-text conversion, and face recognition technologies, all of which are integral to the proposed model named as ConvoLogix. Historically, traditional methods for managing meetings and collaboration involved manual note-taking and attendance tracking. These processes were time-consuming, error-prone, and not conducive to optimization. In today's fast-paced business environment, the absence of automated solutions hinders efficiency and collaboration. The survey underscores the significance of embracing advanced techniques in Machine Learning with Traditional and Deep Learning Models for Audio and video processing for process automation, emphasizing their pivotal role in streamlining meetings and attendance tracking. A key theme within the survey is the identification of limitations associated with conventional approaches to meeting minute generation and attendance recording. The ConvoLogix model's core objective is to leverage these technologies to automate meeting minutes generation and attendance tracking, resulting in time savings, improved collaboration, and data-driven insights.

Key Words: Automated Meeting Summarization, Face Attendance Tracking, Natural Language Processing, Machine Learning, Deep Learning, Neural Network Models

1. INTRODUCTION

Traditional methods in the past involved manual note-taking during meetings, which required individuals to transcribe and summarize the discussions. This process was prone to human errors, consumed significant time and effort, and hindered optimization. Additionally, attendance tracking was typically done manually, relying on sign-in sheets or manual headcounts, which were also prone to inaccuracies. The lack of automated solutions hinders optimization and collaboration in today's fast-paced business environment. The proposed project aims to overcome these challenges by leveraging advanced technologies. The remaining paper is organised as follows: Section 2 introduces the author's research. Section 3 comprises of methodologies and approaches relevant to the two processes of the proposed ConvoLogix Model. Section 4 represents the identified Trends, Challenges and Gaps in the Literature Survey. Section 5 proposes a Future Research Direction, followed by Conclusion in Section 6.

2. LITERATURE SURVEY

The survey is divided into seven modules which is relevant to each step in the process of generating automated meeting minutes and face attendance tracking.

A. Module 1: Audio Extraction from Video

Sasavade et al. [4] expounded on the merits of converting video content into audio, including benefits like the

conservation of storage space, and introduces the MoviePy library as a potent tool for video editing. Furthermore, the document underscores the utilization of Tkinter for the development of a user-friendly graphical user interface (GUI) designed for the video-to-audio conversion procedure. It offers a comprehensive guide with a step-by-step breakdown of instructions on how to extract audio from a video using Python, with accompanying code snippets to facilitate comprehension. The document concludes by hinting at the potential for program improvement, including the ability to accommodate various file formats and the provision of more comprehensive information about the converted audio files.

B. Module 2: Speech Extraction from Audio

Rivet et al. [8] proposed that the separation of audio and visual components in speech can be accomplished through various approaches, which encompass techniques like blind source separation and the incorporation of visual cues in addition to audio data. The paper also underscores the constraints associated with relying solely on audio-based methods and explores potential use cases for audiovisual speech source separation.

Bronkhorst et al. [6] worked on the problem of Target Speech Extraction. The ”cocktail-party effect” alludes to humans’ remarkable capacity to isolate and comprehend a particular speaker in environments filled with noise and numerous voices. An important challenge within this context revolves around the differentiation of the desired speaker from other speakers who often share similar speech characteristics. In recent years, Targeted Speech Extraction (TSE) has emerged as a burgeoning field of research, offering a promising solution to the cocktail-party problem. In audio-based TSE systems, an audio clue, such as an enrollment utterance, serves as a key element for identifying and extracting the speech of the target speaker from the acoustic mixture. The clue encoder in these systems extracts crucial information from this audio clue, which in turn aids the speech extraction process. This speech extraction module encompasses various components, including a mixture encoder, a fusion layer, and a target extractor. The mixture encoder handles the processing of the complex audio mixture and generates relevant features, which are then integrated within the fusion layer. Here, the fusion layer merges these features from the mixture with the information obtained from the audio clue. Ultimately, the target extractor calculates an estimate of the target speaker’s speech by leveraging the fused features. To train the TSE system, a fully supervised training methodology is employed. This entails the optimization of parameters to minimize the dissimilarity between the estimated target speech and the actual target speech, serving as the ground truth.

Gu et al. [7] discuss about the primary discoveries and underscore the significance of voice traits and spatial separation in aiding speech comprehension. Studies have shown that when talkers are of different genders and positioned at least 10° apart, it leads to improved performance in tasks related to speech perception. Furthermore, the combination of interaural level differences (ILDs) and interaural time differences (ITDs) also contributes to better discrimination of speech sounds. The role of attention is pivotal in choosing the target speech, with both inherent and cognitive processes influencing where one’s attention is directed. However, there remains a need for further research to gain a comprehensive understanding of the intricate mechanisms involved in perceiving speech in situations where multiple individuals are speaking simultaneously.

C. Module 3: Speech to Text Conversion

Vinnarasu et al. [9] researched on Speech to text conversion and summarization for providing a user-friendly and efficient approach that converts spoken language into written text while simultaneously generating a concise summary of the resulting text. The intended applications of this method are diverse, including the creation of lecture notes and summarizing lengthy documents. The proposed model integrates both speech recognition and text summarization to provide a comprehensive solution. In this process, the model extracts features from the spoken words and employs natural language processing techniques to convert them into written text. Subsequently, a ranking algorithm is applied to produce a summary of the text, prioritizing words based on their frequency. The effectiveness

of this proposed method is rigorously evaluated through extensive experimentation.

Shivakumar et al. [11] conducted a study that delves into the influence of language models on enhancing the precision of speech-to-text conversion systems. It is a comprehensive analysis of the technologies employed in speech recognition systems with small, medium, and large vocabularies, while also assessing their respective advantages and drawbacks. The primary objective of the experiment is to investigate how language models can augment accuracy, with a specific emphasis on challenging scenarios like noisy sentences and incomplete words. The findings reveal that randomly selected sentences exhibit superior performance in comparison to sequentially structured sentences.

Ghadage et al. [10] proposed a speech-to-text conversion system that is purposefully designed to extract, characterize, and identify information from spoken language signals. In the context of this research paper, the system leverages the Mel-Frequency Cepstral Coefficient (MFCC) feature extraction method, along with the Minimum Distance Classifier and Support Vector Machine (SVM) techniques for speech classification. The system is implemented using MATLAB and is put to the test with a database that the researchers themselves created. This database encompasses sentences in Marathi, English, and a fusion of both languages (Marathi-English). Impressively, the system attains a high level of accuracy, with 93.625% accuracy for Marathi, 91.6667% for English, and 90.625% for sentences that mix Marathi and English.

D. Module 4: Text Summarization

The growing volume of online text content has amplified the significance of automatic text summarization. The processes stated in El-Kassas et al. [12] research work, aims to save time and effort while condensing substantial text volumes into concise summaries. There are various approaches to automatic text summarization, including extractive, abstractive, and hybrid methods. Extractive techniques pick out essential sentences from the source document(s), while abstractive methods craft new sentences that encapsulate the primary ideas in the text. Hybrid methods, on the other hand, combine both extractive and abstractive approaches.

Nonetheless, abstractive methods remain a challenging area that necessitates further development. This survey offers a comprehensive overview of automatic text summarization, encompassing its various approaches, methods, techniques, evaluation procedures, and potential research directions.

The technique put forth in Fattah et al. [13] presented research introduces a trainable summarization system that integrates a range of features. These features encompass sentence position, the presence of positive and negative keywords, sentence centrality, how closely a sentence aligns with the document title, the inclusion of named entities and numerical data, the sentence's length relative to the document, and the complexity of a sentence referred to as the "bushy path." These features collectively contribute to the calculation of a combined similarity score for each sentence.To assess the effectiveness of this approach, the research applies it to a dataset comprising 100 religious articles written in English. The results obtained from this experimentation demonstrate promising outcomes. [13]

Neto et al. [14] conducted their research that uses a machine learning approach for automatic text summarization. This approach employs a collection of features that are directly extracted from the source text. These features include statistical attributes derived from word frequency and linguistic characteristics based on the text's argumentative structure. The research paper assesses the outcomes generated by the trainable summarizer in comparison to several baseline summarization techniques. Computational results are gathered from established text databases for this comparison. The findings reveal that the trainable summarizer, particularly when coupled with the Naive Bayes classifier, outperformed the baseline methods concerning precision and recall.

E. Module 5: Face Recognition

The real-time video processing system proposed by Yang et al. [15] for face recognition in attendance management is

designed to elevate the precision of recognizing individuals during check-ins. Its key goals are to ensure the system's stability, minimize absenteeism, and enhance the user interface settings. According to the experimental findings, this video-based face recognition system achieves an accuracy rate of 82% and reduces absenteeism by around 60%. It results in improved operational efficiency, effectively curbing instances of students leaving early or skipping classes, and it plays a vital role in steering the evolution of attendance management systems.

Deng et al. [16] worked on collaborative representation techniques in face recognition which underscored the significance of how well distinct classes can be separated to achieve precise clustering and classification. The paper introduces a novel approach called the superposed linear representation classifier (SLRC). This method leverages a combination of class centroids and intra-class differences to enhance the model's capacity to generalize effectively in collaborative representation. Through experimentation on diverse datasets, the results demonstrate that SLRC surpasses other dictionary learning methods and attains a level of performance that sets the current benchmark.

The Trunk-Branch Ensemble Convolutional Neural Networks (TBE-CNN) framework is introduced to tackle the challenges associated with video-based face recognition (VFR) by Ding et al. [17]. This framework incorporates a holistic training strategy that leverages both still images and synthetically generated video frames to develop robust face representations, specifically resistant to blurriness. It also employs a Trunk-Branch Ensemble CNN model to enhance its ability to handle variations in pose and facial occlusions. Furthermore, the framework introduces an improved triplet loss function aimed at further enhancing the discriminatory power of the learned representations in the TBE-CNN. Experimental results conducted on three prominent video face databases confirm the effectiveness of these proposed techniques, setting a new standard for performance in the field.

F. Module 6: Speaker Identification

Conventional speaker identification systems necessitate the meticulous design of features, whereas deep learning permits the automatic acquisition of these features. This research work conducted by Jalil et al. [18] introduces a CNN (Convolutional Neural Network) structure that utilizes Mel-spectrograms as input for speaker identification. The study includes experiments conducted on the TIMIT dataset to assess the effectiveness of the proposed CNN architecture, with a particular focus on comparing its performance against state-of-the-art systems in scenarios involving both clear and noisy speech samples.

Bai et al. [19] integrated deep learning and I-vector technology representing a substantial advancement for speaker recognition systems. There are two key research areas of significance: refining the parameters of traditional features and developing effective deep learning models for speaker recognition. This research delves into enhancing the performance of speaker recognition systems by examining various input types and neural network architectures, while also identifying the most optimal feature parameters. The study assesses established deep learning-based speaker recognition algorithms, such as Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN). Additionally, it introduces and compares several improved models. Experimental results underscore the considerable improvements achieved through the amalgamation of these techniques, highlighting the effectiveness of this approach in enhancing speaker recognition rates.

The central focus of Pawar et al. [20] revolved around speaker identification through the use of neural networks and their study underscores the importance of speaker identification in various applications, including telecommunications, financial transactions, and retrieving information from speech databases. The primary emphasis of the paper lies in the realm of text-dependent speaker identification. In this context, users provide a speech sample, which is subsequently compared to stored templates for the purpose of speaker recognition. The system in question makes use of features derived from speech signals,

such as LPC coefficients, AMDF, and DFT, which serve as input parameters for training a neural network. This neural network fine-tunes its weights to attain the most accurate match and determine the identity of the speaker. Additionally, the document explores different paradigms of speech recognition, ultimately concluding with test results that highlight the effectiveness of the system.

G. Module 7: An Integrated Approach for generating Automatic Meeting minutes and a Face Attendance System

As a response to the worldwide transition to online interactions brought about by the COVID-19 pandemic, Bharti et al. [1] introduced a technique for generating both written and audio summaries from recorded video content. This method involves the conversion of the audio from videos into text, followed by the application of text summarization algorithms to create concise summaries. The practical uses of this approach encompass the creation of lecture notes, meeting minutes, subtitles, and the development of storylines. The implementation is carried out in the Python programming language, and its performance is assessed using short videos from YouTube. Since there is a lack of standardized benchmarks and specific datasets for evaluation, a manual validation process is undertaken.

Meetings serve as a prevalent means of communication and idea exchange. However, the manual transcription and summarization of meeting content can be quite time consuming. In response to this challenge, the SmartMeeting tool proposed by Song et al. [2] has been developed. This tool automates the processes of recording, transcribing, summarizing, and managing information during inperson meetings. It makes use of advanced natural language processing techniques, including automatic speech recognition, speaker identification, and meeting summarization. The SmartMeeting system is structured around three core components: ASR-based Transcription, Transcript Enrichment, and Meeting Summarization. The system's performance was assessed using various metrics, including Character Error Rate, Speaker Attribution Accuracy, and ROUGE scores. The results showed promise, although the performance of the Automatic Speech Recognition (ASR) component was influenced by the number of attendees. SmartMeeting provides an array of features, including user registration, meeting management, real-time transcription, meeting summarization, and the capability to search through summaries.

An alternative system for automatic meeting summarization and topic detection developed by Huang et al. [3] uses speech recognition to transcribe spoken content. It leverages a combination of latent Dirichlet allocation and the TextTiling algorithm to identify topic transitions and assess the topics within each segment. Results demonstrate an impressive 85 percent similarity between machine-generated summaries and human-authored records. The core objective of this system is to reduce the human effort necessary for creating meeting reports by autonomously recording and analyzing meeting content.

3. METHODOLOGIES AND APPROACHES

A. Part 1: Relevant to Automated Minutes Generation

For the the Audio Extraction from Video various techniques and methods have been employed, such as utilizing learning analytics to identify exploratory dialogue, implementing voice cloning techniques for various applications, conducting usability evaluation assessments, utilizing the MoviePy library for video editing, and employing Tkinter for the creation of a graphical user interface. Additionally, sociocultural discourse analysis has been utilized to understand classroom interactions and communication [4].

Incorporating visual information alongside audio signals, various techniques have been employed in audiovisual speech source separation, such as video-based voice activity detection, spectral subtraction with visual features, AV postprocessing of audio ICA, AV scene analysis, and full joint AV modeling. These methods collectively aim to enhance speech separation by leveraging both audio and visual data [8]. Another paper delves into a range of techniques and concepts related to the extraction of the target speaker's speech from a mixture,

incorporating audio, visual, and spatial cues. These encompass generative and discriminative models, training and evaluation criteria using signal level metrics, deployment challenges in the context of TSE systems, the utilization of i-vectors and neural network-based embeddings, categorization based on clues and the number of microphones, and the extension of TSE techniques to various other speech processing tasks [5]. The research on speech perception in multi-talker situations includes psychoacoustic models, auditory scene analysis, attentional control, and grouping based on voice characteristics and spatial separation. The document also highlights the need for quantitative models and further research on other grouping cues [6]. The paper ”Neural spatial filter for target speaker speech separation” introduces directional features, such as Directional Power Ratio (DPR) and Directional Signal-to-Noise Ratio (DSNR), along with power spectra and inter-channel spatial features. An attention mechanism is also used to address spatial overlap. The network structure includes LSTM layers and evaluation metrics include SI-SNRi and SDRi [7].

We can find various techniques and strategies for Speech to Text Conversion, including speech recognition through the Google API, text summarization methods based on word frequency and ranking, pre-processing methods to introduce punctuation marks, a comparison of summarization time with the Gensim library, and thorough experimentation aimed at validating the efficiency of the proposed approach [9]. Multilingual speech-to-text conversion system utilizes MelFrequency Cepstral Coefficient (MFCC) feature extraction and a combination of Minimum Distance Classifier and Support Vector Machine (SVM) for speech classification. The system is trained and tested using a self-generated database containing Marathi, English, and Marathi-English mix sentences. The accuracy of the system is evaluated for each language, achieving high accuracy rates [10]. Another paper compares the technologies used in small, medium, and large vocabulary speech recognition systems. It evaluates the benefits and limitations of different approaches and focuses on the role of language models in improving accuracy. The experiment involves testing speech data with noisy sentences and incomplete words, and the results show that randomly chosen sentences yield better accuracy compared to sequential sentences [11].

Automatic Text Summarization can be categorized into extractive, abstractive, and hybrid approaches. Extractive methods select important sentences from the input document(s), while abstractive methods generate new sentences that convey the main ideas of the text. Hybrid methods combine both extractive and abstractive approaches. The paper also explores template-based, ontology-based, semantic-based, and deep learning-based methods within the abstractive approach [12]. Another research work discusses about utilizing features such as sentence position, positive and negative keywords, sentence centrality, sentence resemblance to the title, sentence inclusion of name entities and numerical data, sentence relative length, Bushy path of the sentence, and aggregated similarity. The approaches include a trainable summarizer, genetic algorithm model, and mathematical regression model to generate summaries based on these features [13]. ATS also includes using machine learning algorithms such as Naive Bayes and C4.5 decision tree, employing statistical and linguistic features extracted from the original text, and utilizing heuristics and features based on the argumentative structure of the text. These approaches aim to improve the performance and objectivity of text summarization algorithms [14].

The convolutional neural network (CNN) architecture for speaker identification uses Mel-spectrogram as input. The experiments are conducted on the TIMIT dataset in both clean and noisy environments. The CNN architecture is compared with the conventional UBM-GMM SID system using well known acoustic features. The performance of the proposed CNN architecture is evaluated and discussed [18]. Some techniques in another work that can be inferred include ivector extraction, the combination of DNN (Deep Neural Networks) and i-vector system, and the baseline system. These techniques involve methods for feature extraction, machine learning, and data analysis [19]. The paper ”Speaker Identification using Neural Networks” discusses three approaches to speaker identification: the acoustic phonetic approach, the pattern recognition approach, and the artificial intelligence approach.

It explains the steps involved in each approach, such as preprocessing, feature extraction, training, and pattern comparison. The paper also mentions the use of neural networks in speaker identification [20].

B. Part 2: Relevant to Face Attendance Tracking

Face recognition attendance system based on real-time video processing employs various methods such as geometric feature analysis, subspace analysis, neural network, and support vector machine for face recognition. The system includes modules for login, recognition, check-in, and background management. The accuracy rate, stability, and truancy rate of the system are analyzed, and the interface settings are optimized [15]. Another paper discusses the discriminant nature of collaborative representation methods in face recognition. It proposes a superposed linear representation classifier (SLRC) that combines class centroids and intra-class differences to improve generalization ability. The paper also evaluates the performance of SLRC compared to other dictionary learning techniques on various datasets [16]. Trunk-Branch Ensemble Convolutional Neural Networks (TBE-CNN) framework for video-based face recognition includes training the network with both still images and simulated video frames to learn blurrobust face representations. The framework also incorporates a Trunk-Branch Ensemble CNN model to enhance robustness to pose variations and occlusion. Additionally, an improved triplet loss function is used to improve the discriminative power of the learned representations. Experimental evaluations on three video face databases demonstrate the effectiveness of the proposed techniques [17].

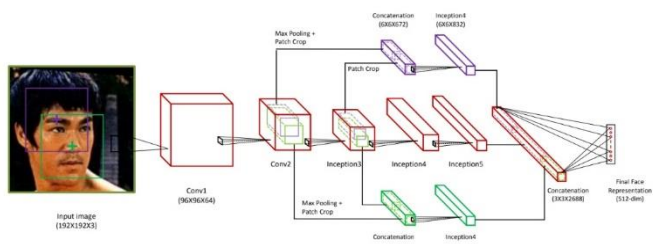

Fig. 1. Model architecture for Trunk-Branch Ensemble CNN (Source: Figure from [17])

C. Part 3: An Integrated Approach for generating Automatic Meeting minutes and a Face Attendance System

Generating audio/text summaries from recorded videos involves extracting the audio from the video, converting it to text, and then using text summarization algorithms to generate a summary. The proposed method also includes features such as translation and text-to-voice conversion. The scheme is implemented in Python and evaluated using short videos from YouTube [1].

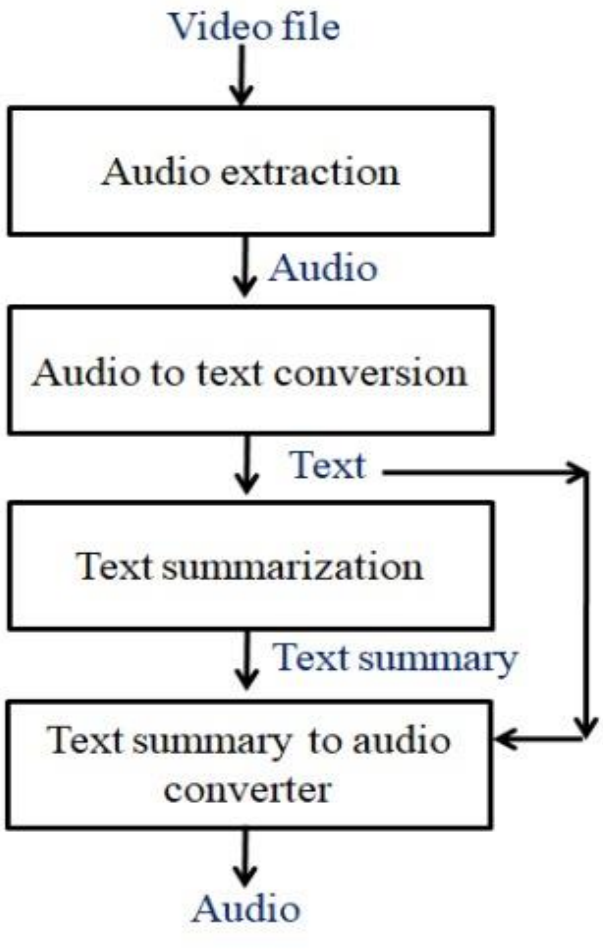

Fig. 2. Meeting Minutes Generation Prcocess (Source: Figure from [1])

The SmartMeeting system integrates multiple natural language processing techniques for automatic meeting transcription and summarization. The methodologies include hybrid ASR models for transcription, speaker identification for voice separation, and a weakly supervised pre-training mechanism for meeting summarization. The approaches involve the use of advanced neural networks, such as transformers and BERT, to enhance the performance of transcription and summarization tasks [2]. The framework for automatic meeting summarization and topic detection utilizes speech recognition, latent Dirichlet allocation (LDA), and a TextTiling algorithm to segment text and identify topic boundaries. Extractive summarization is then used to generate concise summaries. The approach is evaluated using precision, recall, and F-scores [3].

4. TRENDS, CHALLENGES AND GAPS

In the domain of audio extraction from video, a notable trend involves utilizing Python and tools like MoviePy for efficient audio extraction. However, assessing the efficiency of video-to-audio conversion techniques remains a challenge, often necessitating user studies for quality and usability evaluation. Furthermore, the paper's solution for audio extraction from video files is described as basic and lacking advanced features, indicating a gap in the availability of more comprehensive functionalities. To address this gap, there's a need for more user-friendly and advanced methods that not only streamline the process but also enhance accessibility to audio content without requiring the viewing of accompanying videos [4].

Current trends in target speech extraction (TSE), highlight the adoption of multiple clues and neural networks in the field. It underscores the practicality of TSE as a solution to the cocktail-party problem. Despite the progress, notable challenges persist in dynamic scenarios where the target speaker's location is not fixed. Additionally, the paper raises awareness of a gap in research, as investigations into these dynamic cases are relatively rare, emphasizing the need for further exploration in this area to enhance the applicability of TSE methodologies [5].

In a certain research work the proposed approach follows a significant trend in using speech recognition for efficient speech-to-text conversion, offering potential benefits in transcription and enhancing content understanding, particularly in fields like lecture note archiving. This model seamlessly integrates speech recognition technology, providing a comprehensive solution for transcribing spoken language. Nonetheless, a notable challenge lies in its limited focus, primarily summarizing sentences that conclude with a full stop or contain brief pauses marked by commas, overlooking other punctuation marks. Rectifying this gap is essential for advancing the performance and versatility of speech-to-text conversion systems [9].

Automatic text summarization techniques align with the contemporary trend by providing a comprehensive survey, addressing methodological nuances, evaluation complexities, and the growing demand for diversified datasets and standardized assessment measures. This expansive overview covers various facets of automatic text summarization, including approaches, methodologies, techniques, evaluation criteria, and outlines future research directions. However, a notable challenge lies in the absence of an in-depth analysis of specific algorithms or techniques applied in the realm of automatic text summarization, highlighting a gap in the literature. This void underscores the necessity for more focused investigations into these specific methods to advance the field's knowledge and capabilities [12].

A current prominent upswing involves the utilization of real-time video processing to enhance the efficiency of face recognition attendance systems, resulting in a remarkable reduction in truancy rates, with an approximate decrease of 60%, and achieving a notable accuracy rate of 82%. This real-time video processing approach allows for consistent and effective attendance monitoring. Nevertheless, it also presents substantial challenges, notably the system's susceptibility to inaccuracies in recognizing faces due to variations in facial features, accessories, cosmetics, and lighting conditions. Addressing these challenges and enhancing the system's resilience against such factors reveals a significant gap in the existing state of face recognition technology [15].

The field of speaker identification introduces a CNN architecture that effectively utilizes Mel-spectrograms as input, yielding improved accuracy, especially under challenging low signal-to-noise conditions, in contrast to traditional UBM-GMM systems. This shift towards feature learning instead of manually engineered features marks an evolution in speaker identification methodologies. However, a conspicuous gap exists as the paper omits information regarding the computational complexity and efficiency of the proposed CNN architecture. Addressing this gap is vital for assessing the practicality and scalability of the CNN-based approach, thus representing a crucial aspect of future research and development in this domain [18].

The paper ”An Approach for Audio/Text Summary Generation from Webinars/Online Meetings” responds to a prevalent trend in the field by introducing a method for generating audio and text summaries from webinars and online meetings. This innovative approach involves audio extraction, transcription into text, and the creation of concise summaries, aligning with the increasing demand for efficient information retention in such settings. Nevertheless, a substantial challenge lies in the limited scope of the evaluation process, primarily utilizing short YouTube videos, which may not fully represent the complexity and diversity found in real-world webinars and online meetings. This disparity underscores a critical gap in the research, emphasizing the need for the method to be refined and adapted to address a broader range of content, ensuring its practicality and effectiveness in diverse and more complex scenarios [1].

5. FUTURE RESEARCH DIRECTION

Based on the literature survey and problems addressed, the future research directions can be:

1. Enhancing meeting summarization: Further improve the existing text summarization algorithms to capture the nuances of meetings more effectively. This could involve incorporating contextual information, speaker identification, and sentiment analysis to generate more accurate and comprehensive meeting summaries.

2. Advanced attendance tracking: To improve attendance tracking with facial recognition, explore advanced algorithms for handling lighting variations, expressions, and occlusions. Consider integrating voice or fingerprint recognition for enhanced accuracy. Also, incorporate multimodal data (audio, video, text) to gain a holistic understanding of meetings, improving tracking and summarization.

3. Real-time meeting summarization: Develop techniques for real-time meeting summarization, where the system can generate concise summaries as the meeting progresses. This would enable participants to have access to key points and action items in real-time, facilitating better collaboration and decision-making during the meeting.

4. Integration with collaboration tools: Explore the integration of the automated system with existing collaboration tools such as video conferencing platforms. This would allow seamless generation of meeting summaries and attendance tracking within the same interface, enhancing user experience and productivity.

5. User customization and personalization: Develop methods to allow users to customize and personalize the generated meeting summaries and attendance tracking features according to their preferences. This could involve incorporating user feedback mechanisms and adaptive algorithms to tailor the system to individual needs.

6. Evaluation metrics and benchmark datasets: Establish standardized evaluation metrics and benchmark datasets for meeting summarization and attendance tracking. This would enable fair comparisons between different systems and facilitate advancements in the field.

By focusing on these future research directions, the project can contribute to the development of more efficient and accurate meeting summarization systems, as well as advanced attendance tracking techniques, ultimately enhancing efficiency, collaboration, and user experience in the business environment.

6. CONCLUSION

To date, the integration of the six modules discussed in the literature survey has not been accomplished. These modules include audio extraction from video, speech extraction from audio, speech-to-text conversion, text summarization, face recognition and speaker identification. The proposed ConvoLogix model aims to address this gap by integrating these modules together to automate meeting minutes generation and face attendance tracking. However, it is important to note that the module related to speaker identification remains unexplored and poses a significant challenge in the field. The problem of accurately identifying speakers in meetings is yet to be fully solved, and further research is needed to develop effective techniques in this area. The ConvoLogix model fills a crucial void by offering a holistic solution for automating the creation of meeting minutes and tracking attendee faces. This innovation enhances efficiency, fosters collaboration, and empowers data-driven insights within the business landscape.

REFERENCES

[1] Bharti, N., Hashmi, S. N., & Manikandan, V. M. (2021, September). “An Approach for Audio/Text Summary Generation from Webinars/Online Meetings” In 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 6-10). IEEE.

[2] Song, Y., Jiang, D., Zhao, X., Huang, X., Xu, Q., Wong, R. C. W., & Yang, Q. (2021, October). “SmartMeeting: Automatic Meeting Transcription and Summarization for In-Person Conversations” In Proceedings of the 29th ACM International Conference on Multimedia (pp.2777-2779).

[3] Huang, T. C., Hsieh, C. H., & Wang, H. C. (2018). “Automatic meeting summarization and topic detection system” Data Technologies and Applications, 52(3), 351-365.

[4] Sasavade, S., Sutar, T., Barale, K. & Kambale, D. (2023, June). “Extract the Audio from Video by using python” In 2023 International Research Journal of Engineering and Technology (IRJET).

[5] Zmolikova, K., Delcroix, M., Ochiai, T., Kinoshita, K., Cernock ˇ y, J., & ´ Yu, D. (2023). “Neural Target Speech Extraction: An overview” IEEE Signal Processing Magazine, 40(3), 8-29.

[6] Bronkhorst, A. W. (2015). “The cocktail-party problem revisited: early processing and selection of multi-talker speech” Attention, Perception, & Psychophysics, 77(5), 1465-1487.

[7] Gu, R., Chen, L., Zhang, S. X., Zheng, J., Xu, Y., Yu, M., ... & Yu, D. (2019, September). “Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information” In Interspeech (pp.4290-4294).

[8] Rivet, B., Wang, W., Naqvi, S. M., & Chambers, J. A. (2014). “Audiovisual speech source separation: An overview of key methodologies” IEEE Signal Processing Magazine, 31(3), 125-134.

[9] Vinnarasu, A., & Jose, D. V. (2019). “Speech to text conversion and summarization for effective understanding and documentation.” International Journal of Electrical and Computer Engineering, 9(5), 3642.

[10] Ghadage, Y. H., & Shelke, S. D. (2016, April). “Speech to text conversion for multilingual languages.” In 2016 International Conference on Communication and Signal Processing (ICCSP) (pp. 0236-0240). IEEE

[11] Shivakumar, K. M., Jain, V. V., & Priya, P. K. (2017, April). “A study on impact of language model in improving the accuracy of speech to text conversion system.” In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 1148-1151). IEEE.

[12] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). “Automatic text summarization: A comprehensive survey” Expert systems with applications, 165, 113679.

[13] Fattah, M. A., & Ren, F. (2008). ” Automatic text summarization” World Academy of Science, Engineering and Technology, 37(2), 192.

[14] Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002). ” Automatic text summarization using a machine learning approach” In Advances in Artificial Intelligence: 16th Brazilian Symposium on Artificial Intelligence, SBIA 2002 Porto de Galinhas/Recife, Brazil, November 11–14, 2002 Proceedings 16 (pp. 205-215). Springer Berlin Heidelberg.

[15] Yang, H., & Han, X. (2020). ” Face recognition attendance system based on real-time video processing” IEEE Access, 8, 159143-159150.

[16] Deng, W., Hu, J., & Guo, J. (2017). ” Face recognition via collaborative representation: Its discriminant nature and superposed representation” IEEE transactions on pattern analysis and machine intelligence, 40(10), 2513-2521.

[17] Ding, C., & Tao, D. (2017). ” Trunk-branch ensemble convolutional neural networks for video-based face recognition” IEEE transactions on pattern analysis and machine intelligence, 40(4), 1002-1014.

[18] Jalil, A. M., Hasan, F. S., & Alabbasi, H. A. (2019, December). ”Speaker identification using convolutional neural network for clean and noisy speech samples” In 2019 first international conference of computer and applied sciences (CAS) (pp. 57-62). IEEE.

[19] Bai, Z., & Zhang, X. L. (2021). ” Speaker recognition based on deep learning: An overview.” Neural Networks, 140, 65-99.

[20] Pawar, R. V., Kajave, P. P., & Mali, S. N. (2005, August). ” Speaker Identification using Neural Networks.” In Iec (prague) (pp. 429-433)