IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Advances in Automatic Meeting Minute Generation: A Survey

Jaisal Shah[1] and Neelam Jain[2]
Department of Computer Science
S.V. K. M's Mithibai College of Arts, Chauhan Institute of Science and
Amrutben Jivanlal College of Commerce and Economics (Autonomous), Mumbai, Maharashtra, India
Affiliated to University of Mumbai, Mumbai, Maharashtra, India
jaisal13shah@gmail.com[1] and neelam.jain@mithibai.ac.in[2]

Abstract: *We faced the largest crisis of the twenty-first century at the start of 2020: the COVID-19 pandemic. In the midst of the turmoil, the generation ultimately found a method to get the job done by using automation in many aspects of life. Following the epidemic, we saw an 87% increase in video conferencing technologies for daily communications. Almost everything, from online gatherings to college lectures to business meetings, was housed on the internet, which, because it was virtual, increased the odds of ineffective interactions. In reality, statistics collected from employees across all domains reveal that people frequently miss essential points since taking minutes of meetings is a time-consuming, distracting, and extremely dull chore, and that over 37 billion dollars is squandered on ineffective meetings. Keeping track of significant decisions and agreements that were reached at a meeting requires the use of meeting minutes. The issues addressed and the choices made must be recorded in order to be reviewed at the start of the following meeting and for future reference. Many businesses retain salaried personnel to take minutes of meetings, using up valuable time and resources. We provide a method to enable staff members to have productive conversations that will increase a company's productivity by making greater use of the tools and technical improvements that are now accessible. Our approach extracts crucial information from significant debates using Deep Learning methods. The suggestion is for an automated method to record minutes and transcripts of a meeting with the benefit of speaker identification. The model we suggest will be able to recognise the speaker using Mel Frequency Cepstral Coefficient (MFCC)[12], convert an audio file into plain text using Deep Neural Networks (DNN), and summarise the meeting transcript into condensed minutes with the aid of Transformers.*

Keywords: Automatic Meeting Minute Generation.

I. INTRODUCTION

Numerous companies have relied heavily on meetings for a variety of purposes, including decision-making, problem-solving, planning, and brainstorming. A gathering of two or more individuals for productive communication, discussion, and decision-making is known as a meeting.

An individual employee attends eight to twelve meetings on a weekly basis, each lasting between 30 and an hour, in accordance with a poll conducted by Booqed in 2022. Keeping effective meeting notes i.e. minutes is essential to running effective meetings. They act as a reference for both people who weren't there at the session and at a later point in time. They may also be employed as corporate defence in specific circumstances where written evidence is necessary.

Given that the majority of data nowadays is disseminated online, summarising the scattered data has become crucial. The text summary proves to be quite important[1]. This summarising of lengthy business meetings may be done using text summarization. Compilation of a meeting summary NLP may be used to hear consumer questions and concerns and respond with findings, something that Internet service providers have been doing for the previous few years.

Because it eliminates interruptions, irregularities, repairs, and repetitions that are typical of speech, summarised speech ought to be easier to grasp than a straight transcription of speech [7].

There are basically two approaches to summarising business information.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

1. Abstractive Summarization: By analysing the text using advanced regular language computations to create a new, more constrained message that fetches the most data—parts of which may not be included in the initial record—the Abstractive technique tries to provide an overview. In contrast, the extractive approach just makes use of the original text's phrases[10].
2. Extractive Summarization: In an extractive summary, we pick out the most important terms and phrases from the source text and take just those. The summary would include these phrases that were taken out[9].

Original Text: "lagos, nigeria (cnn) a day after winning nigeria's presidency, muhammadu buhari told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. buhari said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability." [8]
Abstractive summary: "muhammadu buhari says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals." [8]
Extractive summary: "muhammadu buhari told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued Nigeria. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability." [8]

II. BACKGROUND

The expanding availability of documents necessitates in-depth research in the field of text summarization. Deep learning has been studied and used in a variety of different study subjects since it emerged as a new and appealing machine learning area in the previous ten years [2]. Speech was a logical early application for
deep learning, and several research papers have been published to this day on the topic of using deep learning for voice-related applications, notably speech recognition [3] - [4] - [5] - [6]. The foundation of traditional speech recognition systems is the use of hidden Markov models-based Gaussian Mixture Models (GMMs) to describe speech signals (HMMs)[11].

III. METHODOLOGY

This research examined a number of publications on "speech recognition" "speaker verification" and "text summarization" that were chosen from prestigious computer science magazines. Each proposed methodology was read in its entirety. In the subject of deep learning, the research issues around these approaches were discovered, and the techniques were organised into groups based on the algorithms used.

3.1 Approaches and Methods

The essential points of the meeting are extracted from the meeting tape. Various articles have written about automating the generation of minutes. This study concentrates on application areas, speech characteristics, and training corpora while attempting to synthesise the current research on speech summarization, taking into account the effects of emerging approaches.
Various techniques are implemented for achieving the desired outcome from proposed research problems.

1. Megha Manuel et. al. proposed model for the Automated Minute Book Creation (AMBOC) system is divided into three components: speech-to-text, speaker verification, and text summarization. It delivers improved results through the utilisation of Google Speech API, MFCC, and Transformers technology, MFCC and Transformers. For the use case of automatic generation of minutes, AMBOC devised a multi-step approach that involves a series of transformations to the data. This process begins with recognizing recordings from speakers and then converting speech to text transcripts. Finally, the transcripts are summarised and attributed to each speaker in the form of minutes[14].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

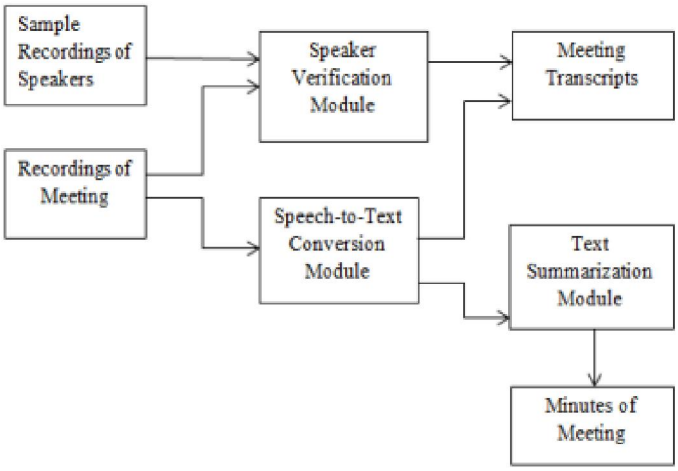

Figure: Stages of proposed AMBOC Model

The limitation of this study is that it is restricted to the Indonesian language. The use of the Google API has resulted in a reduced accuracy when applied to other languages. Further experimentation is necessary to enhance the performance of the AMBOC model. With continued testing, it is believed that the AMBOC model has the potential to demonstrate significant improvement in the future.

2. Haitan Rachman et al. worked in the area of automatic generation of meeting minutes written in Indonesian. The authors aimed to balance the instances per class by using the Synthetic Minority Over-sampling Technique (SMOTE) and resampled the current phrases. To achieve this, they employed four different classifiers - Naive Bayes, Support Vector Machine Linear, IBk, and J48 tree - and used a transcript dataset to train their model.

   The authors reported an F-measure of 85.22 percent and found that resampling the model improved the performance to 94.52 percent. It is also noted that speaker recognition was not taken into account while reviewing the meeting minutes, which may have affected the overall outcome of the study.

   Overall, the text provides a summary of the results of the study by Haitan Rachman et al. on automatic generation of meeting minutes in Indonesian.

3. Justin Jian Zhang et. al.'s study discusses automated minutes generation from legislative speeches. The study discussed in the given statement aimed to find meeting minutes in legislative speeches. In order to achieve this goal, the text of the speeches was divided into smaller chunks and the prominent characteristics of each chunk were extracted using a machine learning classifier known as the Conditional Random Field (CRF). The chunks were then stored in a tree structure using a logical syntax tree, which made it easier to apply classifiers to extract important features from sentences.

   The accuracy of the experiment was evaluated using the ROGUE-L F-Measure, which showed an accuracy of 73.2%. Despite this relatively high accuracy, the study has certain limitations. One of the main limitations is that the remarks in parliament are typically prepared in advance, so the results from this study may not be accurate for unplanned sessions, such as routine office meetings. Additionally, the training process for the CRF classifier is computationally complex, making it challenging to retrain the model with new data if necessary[15].

4. Beam Tasbiraha Athaya et. al. conducted a research to manage meeting minutes and schedule meetings automatically. To encrypt the information and size of the files kept in their database, they employed the Base64 technique. Their work has the drawback of being unable to manage picture and video files [16].

5. Atsuki Yamaguchi, et. al. proposes an automatic minuting system AutoMin2021 - This research was a shared task organised by Hitachi Systems. The task involved developing an automated minute-taking system that can summarise a transcript into topic-based blocks and generate a summary of the meeting. The approach used in this task was a reference-free method that utilised a pre-trained BART model calibrated using a summary corpus of chat discourse. In the task, if a transcript or another minute was provided, the system used multiple

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

relevance ratings to determine if the minute was taken from the same meeting. This approach received the highest adequacy score among all entries and performed well in terms of fluency and grammatical accuracy.

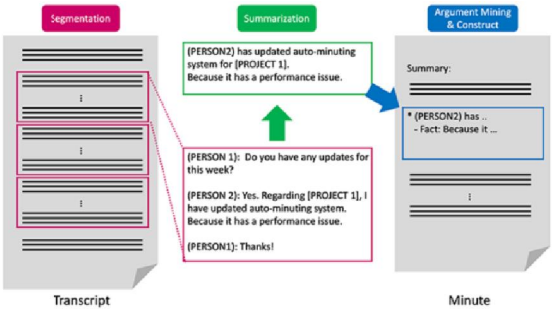

Figure: Multiple interdependent segments transformed into summaries

The suggested model uses segmentation of phrases which outperformed the majority vote baseline models. The model requires training a machine learning model to perform various tasks and make final decisions[17].

6. Praribha Thorat, et. al. describes in this paper that performs the reverse of the desired output. It focuses on constructing a voice-based text summarizer. The proposed model summarises large amounts of textual data and converts the summarised output into audio signals. The proposed approach in this research paper is novel in several ways. Firstly, it employs extractive text summarization to identify the most important paragraphs in the text. Secondly, it introduces a statistically innovative technique for ranking sentences, which is used to determine which sentences are to be selected and summarised by the summarizer. Thirdly, the selected sentences are combined to create a written summary, which is finally converted into audio form.
   This approach differs from traditional text summarization methods, which typically employ abstractive summarization, where the summarizer generates new sentences to represent the most important information in the text. However, extractive summarization has been shown to be more effective for text summarization, as it preserves the meaning of the original text and reduces the risk of introducing errors or inaccuracies[23].
7. J.N. Madhuri, et. al. outlines the strategy that employs extractive text summarization to identify the most important paragraphs in the text. This study introduces an extractive-based text summary approach that is accompanied by a statistically innovative technique for ranking sentences. This ranking system is used to determine which sentences are to be selected and summarised by the summarizer. The selected sentences are then combined to create a written summary, which is finally converted into audio form. This innovative approach not only simplifies the process of summarising large amounts of text data but also makes it easier for the user to listen to the summary instead of having to read through the entire text[18].
8. Josef Steinberger. et. al., presents a novel approach for evaluating the quality of text summarization algorithms and summaries. The authors propose the use of latent semantic evaluation (LSE), which is based on latent semantic analysis (LSA), to overcome the limitations of traditional evaluation methods.ease, without having to go through the entire document in detail. The authors first provide a comprehensive overview of the current state of the field of text summarization, including both extractive and abstractive approaches. They also discuss existing evaluation methods and their limitations, such as the difficulty in obtaining reference summaries and the subjectivity of human evaluations. The authors then describe the LSE approach, which utilizes LSA to measure the semantic similarity between a summary and the original text.
   The results of the experiments conducted by the authors show that LSE is effective in evaluating the quality of summaries generated by different algorithms, and that it provides more reliable results compared to traditional evaluation methods. The results also demonstrate the potential of LSE to serve as a universal evaluation method for different summarization algorithms. The use of LSE has the potential to significantly improve the accuracy and reliability of summary evaluation, and to provide a more objective and consistent way to compare the performance of different summarization algorithms. The authors' novel approach to evaluating summaries using LSE has the potential to greatly advance the field and provide more objective and accurate evaluation results[19].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

9. M.N. Inglole and et.al. presents a pioneering study on automatic consumer audio and video summarization, which achieves efficient summarization by logically segregating the analysis process. The authors make use of saliency models, which are widely recognized for their effectiveness in summarization tasks. The work focuses on creating automatic video summaries in the user domain, where previous methods face several challenges due to the complexity of content material analysis. The videos used in this study were recorded under unstructured settings with lighting, clutter, and significant camera motion, as well as poor quality sound due to the combination of multiple sound sources and significant background noise. These factors present unique difficulties for summarization and make the results of this study particularly significant. The authors' approach to overcome these challenges through the use of saliency models and logical segregation of the analysis process is novel and holds the potential to have a significant impact on the field of consumer audio and video summarization[20].
10. Athanasia Zlatintsi an et. al. have performed an approach that involves the combination of both audio and text information to identify and summarise the key and important events within the audio data. The results of the study demonstrate that using this synergistic approach can lead to more accurate and effective summarization compared to methods that only use audio or text information alone.

    One of the key insights from this study is the recognition of the human auditory system and its ability to detect important events. The authors take advantage of this by utilising auditory and perceptual signals such as noise level, roughness, and teaser power to measure the hearing capacity of the human auditory system. These signals are then used to identify and summarise the most important events in the audio data[21].

    This approach has the potential to significantly advance the field of audio summarization. By combining audio and text information, the authors have created a method that can effectively capture the key events in audio data and present them in a summarised format. The use of auditory and perceptual signals to measure human hearing capacity is an innovative contribution that holds the potential to lead to further breakthroughs in the field[21].

    In our experiments, we discovered commonalities in the corpora utilised by previous research studies. It is not surprising that the same corpora can be used for both speech recognition (speech to text) and audio/text summarization. The corpora used in the evaluated publications are listed in a table and are either publicly accessible or can be obtained upon request. The majority of the speech domain corpora are in English, feature one or two speakers, and vary in size from small to moderate, with very few exceeding 500 hours.

Table: Characteristics of Corpus

Corpus	Summarised Content	Language	No. of speakers	Size
AMI	Meeting	English	More than 2	100 hours
ICSI	Meeting	English	More than 2	70 hours
MATRICS	Multimodal meeting	English	More than 2	10 hours
TEDe	Lecture	English	1	50 hours+75 hours
CSJ	Lecture; Task-oriented dialogue	Japanese	1 or 2	658 hours
TDT2	Broadcast news	English	1 or 2	518-1036 hours

IV. MAIN FINDINGS

This preliminary survey of the speech recognition and summarization research literature has revealed several important gaps. Despite the availability of strategies for automatic speech summarization, there remains a substantial difference in quality between the automatic summarization and manual summarization performed by humans. Additionally, there has

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

been limited research on abstractive summarization, which has the potential to be highly valuable. This lack of research is partly due to the absence of appropriate corpora, resources, and benchmark summaries in the audio domain.

Another issue that has been identified is the limited number of task-based or extrinsic assessments in this field. Most of the existing research has focused on traditional summarization without considering the specific use case for which the summarization is intended. This makes it difficult to replicate and compare studies, as different researchers may use different corpora or different batches of the same corpus.

Finally, the reliability of the speech-to-text conversion, the technique of feature selection, and the overall quality of summarization are all impacted by various factors, including the audio quality, the organised speech, and the speaker count[22]. To address these challenges and make progress in the field, future research should focus on developing robust and reliable methods for speech recognition and summarization, as well as creating appropriate corpora, resources, and benchmark summaries in the audio domain.

4.1 Open Problems and Challenges

1. Inaccuracies in Speech-to-Text Processing - Despite advancements in automatic speech recognition, these technologies still experience issues that affect the summarization of speech content. Machine learning and deep learning techniques, as well as language models, have the potential to improve ASR output accuracy, however, there is limited work in this area to date, including the use of indexing techniques.
2. Speaker Segmentation - Summarising speech from multiple speakers remains a complex task. Diarization refers to the process of dividing an audio stream into segments associated with different speakers. Deep learning methods employed by tech companies such as Google and Microsoft have speaker diarization capabilities, but none of the existing automatic speech recognition (ASR) engines, including commercial ones, have fully solved the diarization problem.Summarizing speech with multiple speakers is difficult and may result in incorrect speaker identification or missing important information if sentences are not accurately detected. This is particularly problematic in question-answer sequences where short answers refer to previous speaker's utterances and the summarization process may not capture the full exchange.
3. Discontinuities in Speech - It is normal for various types of disfluencies to occur in conversations, such as interruptions, overlapping speech, incorrect starts (e.g. "I'll, let's talk about it"), filler expressions (e.g. "of course", "ok", "you know"), non-lexical filled pauses (e.g. "umm", "uh"), and repetitions. Meetings and interviews are challenging for summarization due to disfluencies, filler phrases, redundancies, and lack of structure.
4. Verbatim Text Summarization - Speech summarization consists of two steps - transcribing speech into text and then summarising the transcription. Another method is to create a summary directly from speech without transcribing it first. This can be achieved through the use of deep learning algorithms, computer vision methods, or by giving greater emphasis to the most commonly occurring audio patterns

V. CONCLUSION

In conclusion, the field of automatic meeting minute generation has made significant progress in recent years, driven by the increasing demand for more efficient and effective methods of summarising and documenting meetings. A wide range of techniques and approaches have been proposed and evaluated, including traditional NLP methods, deep learning models, and multimodal approaches that incorporate audio and visual information.

Despite these advances, several challenges still remain in this field. For example, accurately transcribing speech, identifying speakers and their roles, and effectively summarising meeting content, especially in the presence of disfluencies and multiple speakers, remain difficult tasks. Additionally, the limited availability of annotated training data, and the lack of standard evaluation metrics, make it difficult to compare different methods and assess their relative strengths and weaknesses.

Future research in this area should focus on addressing these challenges and developing more robust and effective methods for automatic meeting minute generation. Additionally, standardisation of evaluation metrics and the availability of large-scale annotated training data will be crucial for further progress in this field.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

REFERENCES

[1]. Virender Dehru, Pradeep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi and Pawan Kartik "Text Summarization Techniques and Applications", IOP Conference Series: Materials Science and Engineering, 2021.

[2]. Y. Xie, L. Le, Y. Zhou, and V. V. Raghavan, ''Deep learning for natural language processing,'' in Handbook of Statistics. Amsterdam, The Netherlands: Elsevier, 2018.

[3]. H. Singh and A. K. Bathla, ''A survey on speech recognition,'' Int. J. Adv. Res. Comput. Eng. Technol., no. 2, no. 6, pp. 2186–2189, 2013.

[4]. M. A. Anusuya and S. K. Katti, ''Speech recognition by machine: A review,'' Int. J. Comput. Sci. Inf. Secur., vol. 6, no. 3, pp. 181–205, 2009.

[5]. Y. Zhang, ''Speech recognition using deep learning algorithms,'' Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013, pp. 1–5. [Online]. Available: https://scholar.google.com/scholar?as_q=Speech+Recognition+ Using+ Deep+Learning+Algorithms&as_occt=title&hl=en&as_sdt=0%2C31

[6]. I. Shahin, A. B. Nassif, and S. Hamsa, ''Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments,'' Neural Comput. Appl., to be published

[7]. Goldman, J., et al., Accessing the spoken word. International Journal on Digital Libraries, 2005. 5(4): p. 287-298

[8]. P.J. Liu, and C.D. Manning, Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368, 2017.

[9]. Banerjee, S. and A.I. Rudnicky. An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. in Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. 2008. IEEE.

[10]. Dana Rezazadegan1,2,*, Shlomo Berkovsky2 , Juan C. Quiroz3,2, A. Baki Kocaballi4,2, Ying Wang2 , Liliana Laranjo5,2, Enrico Coiera. Automatic Speech Summarisation: A Scoping Review.

[11]. Y. Zhang, ''Speech recognition using deep learning algorithms,'' Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013, pp. 1–5. [Online]. Available: https://scholar.google.com/scholar?as_q=Speech+Recognition +Using+ Deep+Learning+Algorithms&as_occt=title&hl=en&as_sdt=0%2C31.

[12]. L. Deng et al., ''Recent advances in deep learning for speech research at Microsoft,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 8604–8608.

[13]. G. H. Rachman and M. L. Khodra, "Automatic rhetorical sentence categorization on Indonesian meeting minutes," 2016 International Conference on Data and Software Engineering (ICoDSE),2016.

[14]. Megha Manuel1, Amritha S Menon1 , Anna Kallivayalil1 , Suzana Isaac1 and Lakshmi K.S2, Automated Generation of Meeting Minutes Using Deep Learning Techniques. March 2022.

[15]. Zhang, Justin Jian Fung, Pascale Chan, Ricky "Automatic minute generation for parliamentary speech using conditional random fields. Acoustics, Speech, and Signal Processing".

[16]. Beam Tasbiraha Athaya, Tasbiraha Munira, Sirajum Zaman, Afsana Zaman Hossain, Syed Kabir, Col. "A Proposed Algorithm and Architecture for Automated Meeting Scheduling and Document Management,2018.

[17]. Yamaguchi, A., Morio, G., Ozaki, H., Yokote, K.-i., Nagamatsu, K. (2021) Team Hitachi @ AutoMin 2021: Reference-free Automatic Minuting Pipeline with Argument Structure Construction over Topic-based Summarization. Proc. First Shared Task on Automatic Minuting at Interspeech 2021.

[18]. J.N.Madhuri, Ganesh Kumar R., Extractive text summarization using sentence ranking, 2019 IEEE.

[19]. Josef Steinberger., KarelJezek., "Using latent semantic evaluation in textual content summarization and summary evaluation", Department of computing and Engineering, 2014.

[20]. M.N. Ingole, M.S. Bewoor, S.H. Patil "Text summarization using expectation maximisation cluster based algorithms " International Journal of Engineering Research and Applications 2012.

[21]. Athanasia Zlatintsi, Elias Iosif, "Audio salient occasion detection and summarization the usage of audio and textual content modalities ".

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

[22]. Furui, S. and T. Kawahara, Transcription and distillation of spontaneous speech, in Springer Handbook of Speech Processing. 2008, Springer.

[23]. Pratima Mohan Thorat , Prof. Dr. M. S. Bewoor, A Novel Approach for Voice based Text Summarizer, 2020 IEEE.

[24]. S. and Y. Liu, Using N-best lists and confusion networks for meeting summarization. IEEE Transactions on Audio, Speech, and Language Processing,