Jurnal Sains dan Teknologi
Volume 14 Number 1, Tahun 2025, pp. 200-209
P-ISSN: 2303-3142 E-ISSN: 2548-8570
Open Access: https://doi.org/10.23887/jstundiksha.v14i1.94184

Machine Learning Framework for Automatic Speech Transcription and Summarization Using HMM and TextRank

Yusuf Kurnia[1*], Kristen[2], Ardiane Rossi[3], Junaedi[4], Aditiya Hermawan[5]

[1,2,5]Informatics Engineering, Buddhi Dharma University, Tangerang City, Indonesia
[3,4]Information Systems, Buddhi Dharma University, Tangerang City, Indonesia

ARTICLE INFO

Article history:
Received January 10, 2025
Accepted April 08, 2025
Available online April 25, 2025

Kata Kunci:
Speech to Text, Speech Recognition, Text Summarization, TextRank, Machine Learning.

Keywords:
Speech to Text, Speech Recognition, Text Summarization, TextRank, Machine Learning.

ABSTRAK

Studi ini dilatarbelakangi oleh meningkatnya kebutuhan untuk memproses data audio secara efisien, seperti dalam rapat, kuliah, dan wawancara, yang biasanya masih dilakukan secara manual. Proses manual ini memakan waktu dan rentan terhadap kesalahan manusia, sehingga diperlukan sistem otomatis yang dapat mengubah ucapan menjadi teks dan merangkum informasi secara akurat. Tujuan utama penelitian ini adalah mengembangkan sistem otomatis yang mengintegrasikan Hidden Markov Model (HMM) untuk transkripsi ucapan dan TextRank untuk peringkasan teks, serta mengevaluasi kinerja sistem tersebut. Penelitian ini menggunakan pendekatan eksperimen kuantitatif dengan subjek penelitian berupa data audio dalam format MP3 yang diperoleh dari berbagai kegiatan, seperti rapat, kuliah, dan wawancara. Data audio tersebut diproses dengan metode ekstraksi fitur menggunakan Mel-Frequency Cepstral Coefficients (MFCC), kemudian ditranskripsikan menggunakan HMM dan diringkas menggunakan algoritma TextRank. Analisis data dilakukan dengan mengukur akurasi transkripsi menggunakan Word Error Rate (WER) dan mengevaluasi kualitas peringkasan menggunakan metrik ROUGE. Sistem ini diuji pada tiga kategori audio dengan kompleksitas yang bervariasi. Hasil menunjukkan bahwa sistem mencapai akurasi transkripsi yang tinggi, terutama untuk audio wawancara (WER: 7,6%) dan kinerja ringkasan yang efektif (ROUGE-1: 0,78, ROUGE-L: 0,74). Lebih jauh, alur kerja otomatis menunjukkan peningkatan efisiensi waktu hingga 96% dibandingkan dengan metode manual. Temuan ini menunjukkan kelayakan praktis menggabungkan algoritma probabilistik dan berbasis grafik untuk mengotomatiskan pemrosesan data audio skala besar. Pendekatan ini secara signifikan mengurangi beban kerja manusia sekaligus memastikan akurasi dan konsistensi. Penelitian ini memiliki implikasi yakni berkontribusi pada kemajuan sistem pemrosesan bahasa alami hibrida dan menyediakan landasan yang kuat untuk integrasi masa depan dengan ringkasan abstraktif berbasis transformator dan skalabilitas multibahasa.

ABSTRACT

This study is motivated by the increasing need to process audio data efficiently, such as in meetings, lectures, and interviews, which are usually still done manually. This manual process is time-consuming and prone to human error, so an automated system is needed that can convert speech into text and summarize information accurately. The main objective of this study is to develop an automated system that integrates the Hidden Markov Model (HMM) for speech transcription and TextRank for text summarization, and to evaluate the performance of the system. This study uses a quantitative experimental approach with research subjects in the form of audio data in MP3 format obtained from various activities, such as meetings, lectures, and interviews. The audio data is processed using the feature extraction method using Mel-Frequency Cepstral Coefficients (MFCC), then transcribed using HMM and summarized using the TextRank algorithm. Data analysis is carried out by measuring the accuracy of the transcription using the Word Error Rate (WER) and evaluating the quality of the summary using the ROUGE metric. This system is tested on three audio categories with varying complexity. The results show that the system achieves high transcription accuracy, especially for interview audio (WER: 7.6%) and effective summarization performance (ROUGE-1: 0.78, ROUGE-L: 0.74). Furthermore, the automated workflow shows up to 96% time efficiency improvement compared to the manual method. These findings demonstrate the practical feasibility of combining probabilistic and graph-based algorithms to automate large-scale audio data processing. This approach significantly reduces human workload while ensuring accuracy and consistency. This research has implications for contributing to the advancement of hybrid natural language processing systems and providing a solid foundation for future integration with transformer-based abstractive summarization and multilingual scalability.

*Corresponding author.
E-mail addresses: yusuf.kurnia@ubd.ac.id (Yusuf Kurnia)

1. PENDAHULUAN

In the rapidly developing digital era, the volume of audio-based data has increased significantly, especially in the context of business communication, education, and media. Audio recordings of meetings, lectures, interviews, and seminars have become an integral part of everyday activities and play an important role as a source of information that must be processed efficiently. However, the main challenge in managing audio data lies in the summarization process, which is generally still done manually. This process takes a long time, requires a lot of energy, and is prone to human error, especially when dealing with large and complex data.These inefficiencies have a direct impact on productivity and accuracy, as delays in accessing critical information can impact strategic decision-making. Additionally, manual summaries often vary in quality and consistency, depending on the skills of the individuals processing the data. As the demand for automated natural language processing increases, systems that can understand human speech are increasingly needed. Previous studies have highlighted the potential of NLP-based systems in education, such as the implementation of a web-based chatbot at Universitas Internasional Batam using Natural Language Processing and the Knuth-Morris-Pratt algorithm, which successfully automated campus information services with 86% accuracy(Fadlilah et al., 2022; Verma et al., 2019). This application reflects the growing need for intelligent language-based interfaces in educational contexts. In line with thisIndonesian. Previous researchshows that the development of an automatic speech recognition system based on the XLSR-53 model that has been pre-trained in various languages is able to produce competitive transcription performance with a Word Error Rate (WER) of only 12% using relatively limited training data, emphasizing the urgency and effectiveness of ASR technology in supporting the efficiency of audio data processing in various sectors(Bain et al., 2019; Vora et al., 2020).

Machine Learning technology offers an innovative approach to address these challenges through the development of automated systems capable of recognizing, transcribing, and summarizing information from audio recordings. One of the most widely used methods in speech recognition is the Hidden Markov Model (HMM), which is effective in modeling sequential data and has proven to be reliable in speech recognition applications.(Al-Dabet et al., 2019; Erline & Christian, 2019). The use of HMMs in speech recognition is based on their ability to model stochastic processes that are temporal in nature. HMMs allow for a probabilistic representation of a sequence of observations, such as acoustic features extracted from a speech signal, which allows them to capture temporal dynamics and variations in pronunciation. This approach allows the system to learn and recognize patterns in sequential data, which is important in speech recognition. The application of Hidden Markov Model (HMM) in speech recognition system has shown significant results. The use of HMM for the recognition and pronunciation of Hijaiyah letters has achieved a high level of accuracy, which shows the reliability of HMM in recognizing complex sound patterns (Arisaputra & Zahra, 2024; Isyanto et al., 2022). Another study showed that the use of HMM for emergency vehicle siren recognition also achieved high accuracy, further emphasizing the effectiveness of HMM in handling complex sound patterns(J. Liu et al., 2021; Zieve et al., 2024). Thus, the use of HMM in speech recognition offers an effective and efficient approach to transcribe and summarize information from audio recordings, which can be integrated into automated systems for various applications. Apart from speech recognition, text summarization also plays an important role in extracting important information from audio transcription results (Ardianti et al., 2019; Ubaidi & Dewi, 2024).One of the most effective and widely used methods for automatic text summarization is the TextRank algorithm. TextRank is a graph-based ranking algorithm inspired by PageRank, the algorithm used by Google's search engine. It models text as a graph, where sentences or words are represented as nodes, and the relationships between them are represented as edges. By applying the principle of ranking to this graph, TextRank identifies the most important sentences for extractive summarization (Gumilang et al., 2021; Russell et al., 2021). Several previous studies have demonstrated the effectiveness of TextRank in various text summarization contexts. For example, the application of TextRank in summarizing electronic product reviews showed that the algorithm can extract important information but is less effective in capturing subjective opinions(Rajasa & Prasetio, 2020; Tridarma & Endah, 2020).

Another study compared TextRank's extractive and abstractive methods and found that the extractive approach was more stable and produced more accurate summaries for texts with clear structure.(Yugandhar dkk., 2023) Additionally, a study proposed the integration of TextRank with K-Means clustering, which was shown to reduce information redundancy in the summary results(Deshmukh, 2025; Silva Passos et al., 2024). Based on this research, TextRank is considered as a promising method for automatic text summarization, especially in the context of audio transcription data. This research offers a novelty by integrating HMM for speech transcription and TextRank for text summarization within a machine learning framework, with the aim of creating an automated and efficient system for processing large-scale audio data. The novel value offered by this research is the implementation of the combination of HMM and TextRank in one integrated system, which not only transcribes speech accurately but also

summarizes information effectively by utilizing graph-based algorithms. This approach also allows processing data in various audio categories with varying complexity, and significantly reduces human workload. The urgency of this research is very high, especially considering the increasing volume of audio data that must be processed in various sectors, including business, education, and media. The manual process that is still applied in many institutions can reduce efficiency and consistency, while the automated system developed in this research is expected to offer a faster, more accurate, and more reliable solution for managing large amounts of audio data. The main objectives of this research are to develop an automated system that integrates HMM for speech transcription and TextRank for text summarization, and to evaluate the performance of the system in various audio categories with varying complexity, as well as to measure the time efficiency achieved compared to manual methods.

2. METHOD

This study uses a quantitative experimental approach using machine learning-based software development to automate the Speech-to-Text process using the Hidden Markov Model (HMM) algorithm and data summarization using the TextRank algorithm. (Bain et al., 2019; Nada et al., 2024). The data source for this study consists of audio recordings of various activities containing important information related to a particular discussion topic, such as meeting results or classroom learning sessions. These recordings will be used as a basis for testing the performance of the automatic transcription and summarization system. The data format used in this study is MP3, as it offers wide compatibility and efficient storage without significantly sacrificing audio quality. To ensure processing efficiency and system performance, the input data for the application is limited to a maximum duration of 15 minutes and a file size of no more than 50 Megabytes (MB)(Joshi, 2024; Matsuura et al., 2020). These limits are set to maintain system stability, reduce the computational load during transcription and summarization processes, and ensure the application remains responsive and accurate when handling large audio data.

Figure 1. Flowchart of Audio Processing into Text and Summary

The research methodology is shown in Figure 1, where audio recordings are transcribed and summarized into text using the Hidden Markov Model (HMM) algorithm. TextRank is also used to determine which sentences are most important for extractive summarization.This brief research procedure beginsThe first stage in system development is the process of collecting audio data from various sources that represent real-life situations, such as meeting discussions, academic lectures, and interview sessions. The MP3 audio format was chosen due to its wide compatibility and efficiency in terms of data storage and transmission. In this context, the collected audio data must meet the criteria of a maximum duration of 15 minutes and a file size not exceeding 50MB to ensure system stability and processing efficiency. This stage is critical to ensure the diversity of contexts and voice characteristics, including variations in intonation, articulation, accent, and background noise, which will affect the accuracy of the designed transcription and summarization system.(Nath & Roy, 2024; Ramadhan et al., 2020). Second, praproses focuses on transforming raw audio signals into structured input features suitable for modeling. This process begins by segmenting the audio into smaller units to facilitate sequential analysis. Next, the main features extracted are Mel-Frequency Cepstral Coefficients (MFCC), which have been empirically proven to accurately represent the phonetic characteristics of human speech.(Jollyta et al., 2024; Paisey et al., 2024). MFCC has the advantage of approximating human auditory perception because it is based on the Mel scale, which adjusts the frequency resolution. In addition to feature extraction, the process also involves signal normalization and noise removal to ensure that the extracted acoustic features are robust to variations in the recording environment. Effective preprocessing plays a vital role in improving the performance of speech recognition models at later stages.

Third, moThe main machine learning model used in this study is the Hidden Markov Model (HMM), which is trained using manual transcription data and MFCC features obtained in the previous stage. HMM is an effective statistical model in modeling sequential data, especially in speech recognition applications, because it can represent the probabilistic relationship between hidden states (phonemes) and actual observations (acoustic features)(Erline & Christian, 2019; Kommey et al., 2020; Matsuura et al., 2020). The training process involves estimating the transition parameters between observation states and emissions through the Baum-Welch approach or similar algorithms. By optimizing the probability of a sequence of observations corresponding to a sequence of phonetic states, the HMM model can build a representation

capable of recognizing speech patterns despite articulation variations between individuals. This stage serves as the main foundation for building a probabilistic-based automatic transcription system. Once trained, the HMM-based STT model is used to infer transcriptions from previously unseen audio files. Decoding is performed using the Viterbi algorithm, which identifies the most likely sequence of hidden states that could produce the observed MFCC sequence. Transcription accuracy is evaluated using the Word Error Rate (WER), which is calculated as the number of substitutions, insertions, and deletions divided by the total number of words in the reference transcription as shown inequation 1 under (Russell dkk., 2024). Evaluation of the three types of audio yielded varying WERs, reflecting the influence of acoustic quality and linguistic complexity, as confirmed by recent similar findings. Word Error Rate (WER) measures the accuracy of the model in converting audio to text. A lower WER value indicates that the model is able to produce a more accurate transcription that matches the actual speech. Conversely, the higher the WER, the more errors occur, either in the form of substitutions, deletions, or additions of words that should not be in the transcription result.Once the text transcription is obtained, the next step is to extract key information from the text using the TextRank algorithm, a graph-based automatic summarization method. In this approach, each sentence is represented as a node in the graph, while the semantic relationships between sentences are represented as edges. The PageRank algorithm is then used to calculate the importance weight of each node based on its connectivity in the graph, so that sentences with a central role can be identified as summarized. The evaluation of the summarization accuracy is performed using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, including ROUGE-1 (unigram match), ROUGE-2 (bigram match), and ROUGE-L (longest common subsequence). These metrics are widely accepted in the comparison of automatic summarization. Overall, the stages described in the HMM model enable the system to transcribe audio recordings into text and then summarize the text with high efficiency. HMM plays a vital role in recognizing speech patterns by probabilistically mapping the relationship between phonemes and words, which enables accurate speech-to-text conversion. Meanwhile, TextRank analyzes the relationship betweenThe combination of HMM and TextRank offers several advantages, especially in handling large amounts of audio data automatically and consistently. HMM is effective in recognizing variations in speech, including accent and intonation differences, while TextRank provides an objective summary without the need for manual interpretation. Furthermore, this model-based approach reduces the time and human resources required for transcription and summarization. This method is especially useful in practical applications, such as processing meeting or lecture recordings, where speed and accuracy in data processing are critical for decision making and documentation. To evaluate the system, two types of evaluation instruments are used.presented in Table 1.

Table 1 Validity Evaluation Instrument

No	Evaluation Instrument	Aspects measured	Indicator	Validity
1	Word Error Rate (WER)	Accuracy of transcription results compared to manual reference text	Substitution (S), Deletion (D), Insertion (I), and Word Count (N)	Valid because it refers to international standard metrics in ASR(Russell dkk., 2024)
2	RED (RED-1, RED-2, RED-L)	Compatibility of system summary content with manual summary	Matching of unigrams, bigrams, and longest subsequences between system summary and manual reference	Valid because it is widely used in NLP for summary evaluation.(Verma dkk., 2023)

3. RESULT AND DISCUSSION

Result

This study successfully developed a machine learning-based system that is able to automatically transcribe audio recordings into text and produce concise summaries. The workflow of this system is illustrated in Figure 2 and follows a series of structured steps designed to simplify the processing of audio input—from user interaction to output delivery. The system integrates the Google Speech-to-Text API for transcription and the Google AI Summarization API for summarizing the transcribed content. Here is a detailed explanation of each step in the flowchart.

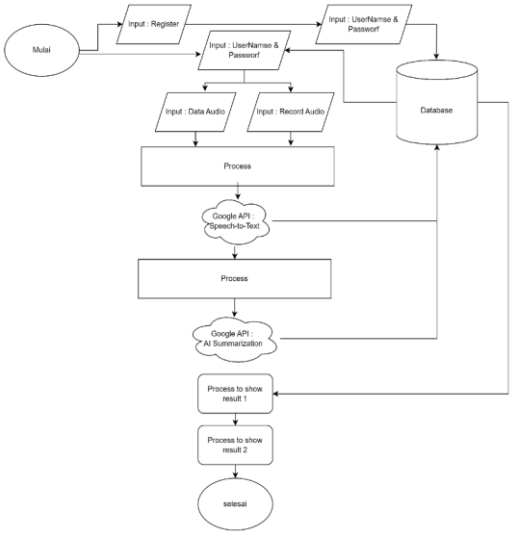

Figure 2. Application System Planning Flow Diagram

The process begins with the system initialization step, when the platform is prepared for user interaction. At this stage, new users are required to register by creating an account with a username and password. Existing users can log in using their credentials. This authentication process ensures secure and personalized access to the system. Once logged in, users are given two options for providing audio input. The first option is to upload a pre-recorded audio file, such as in MP3 format. The second option is to record audio directly through the system interface. The input audio then undergoes a preprocessing stage, which includes segmentation, normalization, and feature extraction—such as Mel-Frequency Cepstral Coefficients (MFCC)—to ensure compatibility with speech recognition engines. The preprocessed audio is then sent to the Google Speech-to-Text API, which automatically transcribes the spoken content into written text using machine learning models such as Hidden Markov Models (HMMs) or neural networks. The resulting transcription is then cleaned and structured in preparation for the summarization phase. This can include detecting sentence boundaries and normalizing the text to make it more readable. The cleaned text is then processed through the Google AI Summarization API, which uses graph-based (such as TextRank) or transformer-based summarization techniques to extract the most relevant sentences and generate a concise summary. The system presents an initial summary view to give the user a quick overview of the audio content. In addition to the short summary, a detailed view is also provided, which displays the full transcription and the resulting summary, allowing the user to perform a more thorough review and validation. The workflow ends with the option for the user to download, save, or use the transcription and summary for further purposes such as documentation, reporting, or analysis. This incremental automation not only reduces the time and human effort required for manual transcription and summarization, but also improves consistency, accuracy, and user accessibility across domains such as business meetings, educational lectures, and media interviews. HThe main page displays the main application interface which is equipped with main function menus, including the data input menu, data processing menu, and data processing results display menu. Home page showed in Figure 3.

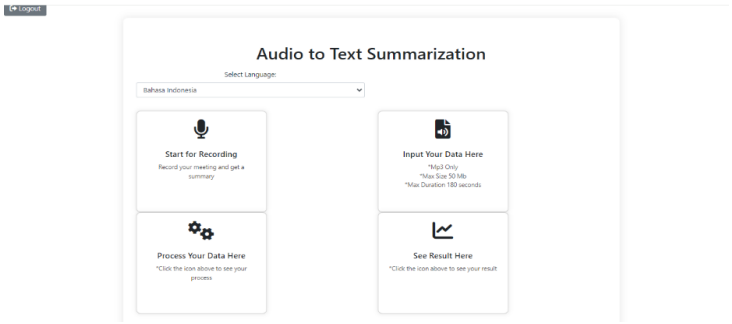

Figure 3. Home Page

TDisplay the results of the trial process of converting audio files to text that has been carried out, which is presented as a summary only.A detailed view of the audio to text conversion process along with a summary presenting the essence of the converted audio is presented in Figure 4. 5Audio to Text Detailed Conversion Results Page showed in Figure 5.

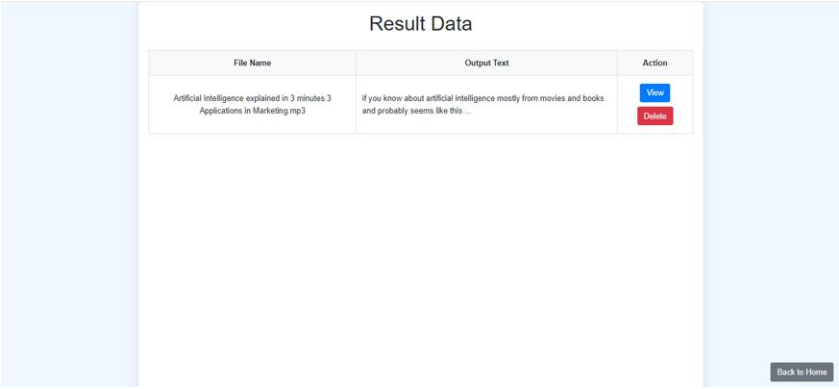

Figure 4. Audio to Text Conversion Results Page

Figure 5. Audio to Text Detailed Conversion Results Page

After the system was developed, a series of experiments and evaluations were conducted to measure the effectiveness of the transcription and the quality of the resulting summary. The performance of the system in converting audio to text was evaluated using the Word Error Rate (WER), which measures transcription errors based on the number of incorrect words compared to the total words in the manual transcription. The evaluation results are presented in Table 2.

Table 2 Speech to Text Evaluation (Google API)

No	Category	Audio Duration (Minutes)	Google API Accuracy (%)	WHO (%)
1	Audio 1 (Meeting)	8	87.2	12,8%
2	Audio 2 (Lecture)	15	81.6	18,4%
3	Audio 3 (Interview)	5	92.4	7.6% of total

HGoogle Speech-to-Text API performance evaluation results based on three audio categories with different durations. The evaluation measures speech recognition accuracy and Word Error Rate (WER) as an indicator of transcription errors. The results show that the performance of Google Speech-to-Text API varies depending on the type and duration of the audio. This difference in accuracy can be attributed to external factors such as noise level and speech intelligibility. Interview audio tends to have a more controlled environment and clearer speech, resulting in more accurate transcription. In contrast, lecture audio may contain more background noise or intonation variations, which makes speech recognition more challenging. Overall, Google Speech-to-Text API performs well in capturing audio content, especially for audio with high speech intelligibility and minimal background noise.

In addition, an evaluation is conducted to assess the quality of the summaries generated by the tested system using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. This metric compares the automated summaries with the human-generated reference summaries. These metrics consist of ROUGE-1 (unigram or single word matching), ROUGE-2 (bigram or word pair matching), and ROUGE-L (longest common sequence matching), which provide a comprehensive view of the accuracy and alignment of the automated summaries with the reference summaries.

Table 3. Automatic Text Summarization Evaluation

No	Category	Transcription Text Length (Words)	Red-1	Red-2	Red-L
1	Audio 1 (Meeting)	year 1173	0.75	0.68	0.71
2	Audio 2 (Lecture)	1852	0.72	0,65	0.69
3	Audio 3 (Interview)	year 843	0.78	0.72	0,74

The evaluation results show that the system achieves the best accuracy on interview audio, with a ROUGE-1 score of 0.78, a ROUGE-2 score of 0.72, and a ROUGE-L score of 0.74. This high accuracy is attributed to the structured and straightforward nature of the conversation in the interview, making it easy for the system to identify important information. In meeting discussions, even though there are many speakers, the system is still able to capture key information with fairly good accuracy. For meeting discussion audio, the system recorded a ROUGE-1 score of 0.75, a ROUGE-2 score of 0.68, and a ROUGE-L score of 0.71. Although slightly lower than the interview category, these results indicate that the system can summarize important points accurately. The system experienced a decrease in accuracy for lecture audio, with a ROUGE-1 score of 0.72, a ROUGE-2 score of 0.65, and a ROUGE-L score of 0.69. This decrease may be due to the complexity of the sentences, the use of technical terms, and the longer transcription length (1852 words), which made it difficult for the system to capture the relationships between sentences and produce a cohesive summary.

The evaluation results based on the ROUGE metrics show that the system achieves the best performance on interview recordings, followed by meeting discussions and lectures. This variation is mainly influenced by the discourse structure and background noise. However, beyond accuracy, another important dimension in evaluating the practical contribution of the system is the speed of information processing, especially when compared to traditional manual transcription and summarization methods. To assess the efficiency improvements provided by the proposed system, a time-based comparison is performed between manual and automated processes for converting audio to text and summarizing its contents. The manual process usually involves several steps such as listening to the entire audio file (often repeatedly), manually identifying important information, and constructing a coherent summary. In contrast, the automated system runs these stages through a machine learning pipeline with minimal human intervention. AA comparative analysis of information processing times for three audio categories, showing significant time savings achieved through automation is presented in Table 4.

Table 4 Automatic Text Summarization Evaluation

No	Category	Audio Duration (Minutes)	Manual Processing Time (Minutes)	Automatic Processing Time (Minutes)	Time Efficiency (%)
1	Audio 1 (Meeting)	8	30	5.2	94,8%
2	Audio 2 (Lecture)	15	60	9.2	90,8%
3	Audio 3 (Interview)	5	20	3.6	96,4%

Automated systems significantly reduce the time required for audio processing. On average, the automation process achieved a 93% reduction in processing time across all audio categories. This efficiency is particularly impactful in the context of organizations with high data throughput, such as academic institutions, government meetings, or corporate environments, where timely access to summarized information directly impacts operational response and decision-making accuracy.

Discussion

The findings of this study underscore the efficacy of integrating Hidden Markov Models (HMM) and TextRank algorithms within a machine learning framework to automate speech transcription and text summarization. Empirical evaluations conducted using three different audio categories: meetings, lectures,

and interviews show significant performance differences in transcription accuracy and summary quality, with the systems achieving the highest Word Accuracy Rate (WAR) and ROUGE scores on interview data. This trend suggests that speech data characterized by structured dialogue and minimal background noise is more suitable for probabilistic decoding and sentence extraction techniques. In contrast, lecture audio, which tends to involve complex terminology, less structured discourse, and higher acoustic variability, yields the lowest performance, specifically higher Word Error Rate (WER) and reduced ROUGE-L scores. These results are consistent with previous studies showing that speech recognition systems are sensitive to noise, prosodic variation, and linguistic complexity(Jollyta et al., 2024; Verma et al., 2019).

The results of this study indicate that the developed automatic audio transcription and summarization system can increase time efficiency by up to 96.4% compared to the manual process. This level of efficiency is in line with the results of previous studies., which reports that Google Assistant achieves a 95% success rate in responding to voice commands, highlighting the substantial potential of speech recognition technology in improving productivity (Al-Dabet et al., 2019; Silva Passos et al., 2024). In addition, user satisfaction levels exceeding 95%, particularly in terms of perceived usability and accuracy, support the findings.who found that students have a positive attitude towards ASR (Automatic Speech Recognition) technology in English oral training, because this technology helps them understand their speaking ability more intuitively(Deshmukh, 2025; Isyanto et al., 2022). The main advantage of this study lies in the integration of automatic transcription and text summarization, which has received little attention in previous studies. Previous studiesshows that the whole-speech summarization approach can reduce the error propagation from ASR to summarization, although the model still faces training challenges. The system developed in this study offers a practical solution by combining the HMM model for transcription and the TextRank algorithm for summarization, making it applicable to real-world scenarios such as education, journalism, and business intelligence(Maulidia Sari & Siti Fatonah, 2024; Silva Passos et al., 2024).

These findings are in line with recent research highlighting the importance of data quality in the performance of speech recognition systems(Joshi, 2024; W. Liu et al., 2002). Emphasizing that models like Whisper show high accuracy on good quality audio transcriptions, but are prone to “hallucinations” or inaccurate text generation on low quality data. This phenomenon suggests the need for special attention to the quality of audio input in practical implementations. Compared with previous studies, the implementation of HMM in this study strengthens its robustness in modeling sequential acoustic data, a characteristic that has long been recognized in classical speech processing literature (Deshmukh, 2025; Silva Passos et al., 2024).]Current results extend this understanding by showing that, even when compared to more state-of-the-art deep learning architectures, HMMs remain a competitive solution under conditions of data and computational resource constraints.

In the context of previous literature, the use of HMMs in speech recognition has long been recognized due to its solid statistical framework and the availability of effective training algorithms. However, recent developments indicate a shift towards deep learning models such as Transformers, which offer improved performance in speech recognition tasks. Similarly, the TextRank algorithm has been widely used in extractive text summarization, but recent studies have shown that improvements in damping factors and similarity measures can improve the quality of summarization.(Al-Dabet et al., 2019; Arisaputra & Zahra, 2024). That the synergy between probabilistic models (HMM) and graph-based algorithms (TextRank) can be effectively operated in a modular architecture that supports large-scale cloud-based implementations. Second, the results show that the system performance is not only algorithm-dependent but also significantly affected by discourse structure and acoustic quality, emphasizing the importance of domain-specific preprocessing and tuning to improve generalization. Third, while extractive summarization is adequate for structured discourse, these findings open up opportunities to explore abstractive models based on transformers, such as BERTSum and T5, to capture semantic nuances in complex free speech. Furthermore, the integration of predictive modeling such as the ensemble-based approach proposed by previous researchers to analyze the behavior of LMS can improve the system's adaptability by anticipating user engagement patterns and optimizing the robustness of the language model. Finally, this study provides an empirical benchmark that can serve as a basis for future comparative studies, especially those aimed at improving the system's robustness to language and dialect variations.(Ardianti et al., 2024) Implications of this studycovers various aspects, both in academic and practical domains. Academically, this research provides an important contribution to the development of natural language processing (NLP) technology, especially in the integration of the Hidden Markov Model (HMM) model for speech transcription and the TextRank algorithm for automatic summarization. This approach can be the basis for further studies in the fields of speech-to-text and automatic summarization. From a practical perspective, this framework has great potential to be applied in various sectors, such as education, customer service, journalism, and the legal field, where efficient and accurate transcription and

summarization of conversations or speeches are required. The use of this technology can save time, increase productivity, and facilitate access to information for users. In addition, the results of this study can also encourage the development of virtual assistant systems and AI-based applications that are smarter in handling voice input and producing concise and relevant information. This study has several limitations that need to be considered. First, the use of the Hidden Markov Model (HMM) model for speech transcription still has limitations in handling variations in intonation, accent, and background noise, which can affect the accuracy of transcription, especially on unclean or informal audio data. In addition, the TextRank algorithm used for summarization is extractive, so the resulting summary tends to only extract important sentences without doing deep contextual understanding, which can reduce the coherence and completeness of the summary. To overcome these limitations, it is recommended that further research explore the integration of deep learning models such as LSTM or transformer-based models (e.g. BERT or Whisper) that are more adaptive to language and context variations. In addition, abstractive summarization approaches based on neural models should be considered to produce more natural and informative summaries. The use of larger and more diverse datasets is also recommended to improve the generalization and overall performance of the system.

4. CONCLUSION

This study successfully developed an automated framework that integrates Hidden Markov Model (HMM) for speech transcription and TextRank algorithm for text summarization. The system provides an efficient solution to the challenges in audio data processing, which often requires significant time and effort when done manually. By automating the transcription and summarization process, the system not only improves accuracy and consistency but also allows information processing in a much shorter time. The findings emphasize the importance of discourse structure and audio quality in influencing model performance, indicating that probabilistic and graph-based approaches can produce effective results in automatic transcription and summarization. This study also paves the way for further development by integrating transformer-based abstractive summarization models and multilingual evaluation, which is expected to improve the semantic depth and applicability of the system across different languages and application contexts.

5. REFERENCES

Al-Dabet, S., Tedmori, S., & Al-Smadi, M. (2019). Enhancing Arabic Aspect-Based Sentiment Analysis Using Deep Learning Models. *Computer Speech And Language*, *69*(28), 101224. Https://Doi.Org/10.1016/J.Csl.2021.101224.

Ardianti, M., Nurhayati, O. D., & Warsito, B. (2019). Model Prediksi Kinerja Siswa Berdasarkan Data Log Lms Menggunakan Ensemble Machine Learning. *Jst (Jurnal Sains Dan Teknologi*, *12*(3), 562–571. Https://Doi.Org/10.23887/Jstundiksha.V12i3.59816.

Arisaputra, P., & Zahra, A. (2024). Indonesian Automatic Speech Recognition With Xlsr-53. *Ingenierie Des Systemes D'information*, *27*(6), 973–982. Https://Doi.Org/10.18280/Isi.270614.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2019). Whisperx: Time-Accurate Speech Transcription Of Long-Form Audio. *Proceedings Of The Annual Conference Of The International Speech Communication Association, Interspeech*, 4489–4493. Https://Doi.Org/10.21437/Interspeech.2023-78.

Deshmukh, A. M. (2025). Comparison Of Hidden Markov Model And Recurrent Neural Network In Automatic Speech Recognition. *European Journal Of Engineering Research And Science*, *5*(8), 958–965. Https://Doi.Org/10.24018/Ejers.2020.5.8.2077.

Erline, M., & Christian, Y. (2019). Web-Based Chatbot With Natural Language Processing And Knuth-Morris-Pratt (Case Study: Universitas Internasional Batam. *Jst (Jurnal Sains Dan Teknologi*, *11*(1), 132–141. Https://Doi.Org/10.23887/Jstundiksha.V11i1.43258.

Fadlilah, M. F., Atmadja, A. R., & Firdaus, M. D. (2022). *Pemanfaatan Transformer Untuk Peringkasan Teks : Studi Kasus Pada Transkripsi Video Pembelajaran* (Vol. 6, Issue 3, Pp. 2111–2119). Https://Doi.Org/10.47065/Bits.V6i3.6342.

Gumilang, K., Nugraha, A. F., Savitri, I., Yani, N. F., Puspitasari, Y. A., Sekartaji, J. G., Satria, A., Nadeak, C. T., & Lailani, A. (2021). Klasifikasi Suara Katak Menggunakan Model Deep Learning Modified Densenet-121 Dan Densenet-169 Dengan Fitur Ekstraksi Mfcc. *Prosiding Seminar Nasional Sains Dan Teknologi Seri Iii*, *2*(1), 627–637.

Isyanto, H., Arifin, A. S., & Suryanegara, M. (2022). Performance Of Smart Personal Assistant Applications Based On Speech Recognition Technology Using Iot-Based Voice Commands. *International Conference On Ict Convergence*, 640–645. Https://Doi.Org/10.1109/Ictc49870.2020.9289160.

Jollyta, D., Oktarina, D., & Johan, J. (2024). Tinjauan Kasus Model Speech Recognition: Hidden Markov Model. *Jurnal Edukasi Dan Penelitian Informatika (Jepin*, *6*(2), 202. Https://Doi.Org/10.26418/Jp.V6i2.39231.

Joshi, P. (2024). *An Introduction To Text Summarization Using The Textrank Algorithm (With Python Implementation*. Https://Www.Analyticsvidhya.Com/Blog/2018/11/Introduction-Text-Summarization-Textrank-Python/.

Kommey, B., Addo, E. O., & Tamakloe, E. (2020). A Hidden Markov Model-Based Speech Recognition System Using Baum-Welch, Forward-Backward And Viterbi Algorithms. *Jordan Journal Of Electrical Engineering*, *9*(4), 509–536. Https://Doi.Org/10.5455/Jjee.204-1675950756.

Liu, J., Liu, X., & Yang, C. (2021). A Study Of College Students' Perceptions Of Utilizing Automatic Speech Recognition Technology To Assist English Oral Proficiency. *Frontiers In Psychology*, *13*(December), 1–9. Https://Doi.Org/10.3389/Fpsyg.2022.1049139.

Liu, W., Sun, Y., Yu, B., Wang, H., & Peng, Q. (2002). *Automatic Text Summarization Method Based On Improved Textrank Algorithm And K-Means Clustering* (P. 39).

Matsuura, K., Ashihara, T., Moriya, T., Tanaka, T., Kano, T., Ogawa, A., & Delcroix, M. (2020). Transfer Learning From Pre-Trained Language Models Improves End-To-End Speech Summarization. *Proceedings Of The Annual Conference Of The International Speech Communication Association, Interspeech*, 2943–2947. Https://Doi.Org/10.21437/Interspeech.2023-1307.

Maulidia Sari, Y., & Siti Fatonah, N. (2024). Peringkasan Teks Otomatis Pada Modul Pembelajaran Berbahasa Indonesia Menggunakan Metode Cross Latent Semantic Analysis (Clsa. *Jurnal Edukasi Dan Penelitian Informatika*, *7*(2), 153–159. Https://Doi.Org/10.26418/Jp.V7i2.47768.

Nada, Q., Ridhuandi, C., Santoso, P., & Apriyanto, D. (2024). Speech Recognition Dengan Hidden Markov Model Untuk Pengenalan Dan Pelafalan Huruf Hijaiyah. *Jurnal Al-Azhar Indonesia Seri Sains Dan Teknologi*, *5*(1), 19. Https://Doi.Org/10.36722/Sst.V5i1.319.

Nath, S. S., & Roy, B. (2024). Towards Automatically Generating Release Notes Using Extractive Summarization Technique. *Proceedings Of The International Conference On Software Engineering And Knowledge Engineering*, 241–248. Https://Doi.Org/10.18293/Seke2021-119.

Paisey, E. K., Santosa, E., Kurniawati, A., Supijatno, & Matra, D. D. (2024). Long-Reads-Based Transcriptome Dataset From Leaves Of Lime, Citrus Aurantiifolia (Christm.) Swingle Treated By Ethephon And Abscisic Acid. *Data In Brief*, *48*(3), 109167. Https://Doi.Org/10.1016/J.Dib.2023.109167.

Rajasa, M. F., & Prasetio, B. H. (2020). *Penerapan Metode Hidden Markov Model Pada Sistem Pengenalan Suara Sirene Kendaraan Darurat* (Vol. 1, Issue 1, Pp. 1–8).

Ramadhan, M. R., Endah, S. N., & Mantau, A. B. J. (2020). Implementation Of Textrank Algorithm In Product Review Summarization. *Icicos 2020 - Proceeding: 4th International Conference On Informatics And Computational Sciences*, 41. Https://Doi.Org/10.1109/Icicos51170.2020.9299005.

Russell, S. O. C., Gessinger, I., Krason, A., Vigliocco, G., & Harte, N. (2021). What Automatic Speech Recognition Can And Cannot Do For Conversational Speech Transcription. *Research Methods In Applied Linguistics*, *3*(3), 100163. Https://Doi.Org/10.1016/J.Rmal.2024.100163.

Silva Passos, R., Rocha, C. A. A. C., Carvalho, A. P. O., Silva, L. B., & Silva, R. L. A. (2024). Environmental Noise Exposure Assessment From Fireworks At Festivals And Pilgrimages In Northern Portugal. *Applied Acoustics*, *181*(28), 108143. Https://Doi.Org/10.1016/J.Apacoust.2021.108143.

Tridarma, P., & Endah, S. N. (2020). Pengenalan Ucapan Bahasa Indonesia Menggunakan Mfcc Dan Recurrent Neural Network. *Jurnal Masyarakat Informatika*, *11*(2), 36–44. Https://Doi.Org/10.14710/Jmasif.11.2.34874.

Ubaidi, U., & Dewi, N. P. (2024). Penerapan Hidden Markov Model (Hmm) Dan Mel-Frequency Cesptral Coefficients (Mfcc) Pada E-Learning Bahasa Madura Untuk Anak Usia Dini. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(6), 1111–1120. Https://Doi.Org/10.25126/Jtiik.2020722477.

Verma, J. P., Bhargav, S., Bhavsar, M., Bhattacharya, P., Bostani, A., Chowdhury, S., Webber, J., & Mehbodniya, A. (2019). Graph-Based Extractive Text Summarization Sentence Scoring Scheme For Big Data Applications. *Information (Switzerland*, *14*(9), 1–28. Https://Doi.Org/10.3390/Info14090472.

Vora, A., Jain, R., Shah, A., & Sonawane, S. (2020). *Extractive Summarization Using Extended Textrank Algorithm* (P. 38).

Zieve, M., Gregor, A., Stokbaek, F. J., Lewis, H., Mendoza, E. M., & Ahmadnia, B. (2024). Systematic Textrank Optimization In Extractive Summarization. *International Conference Recent Advances In Natural Language Processing, Ranlp*, 1274–1281. Https://Doi.Org/10.26615/978-954-452-092-2_135.