Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/

e-ISSN : 2964-2647

IMPLEMENTATION OF THE FORCED ALIGNMENT ALGORITHM FOR AUDIO AND TEXT ALIGNMENT ON THE WEBSITE FOR THE MEANING OF THE JURUMIYAH BOOK

Yoga Ari Cahyadi [1)*], Hidayatus Sibyan [2)], Nur Hasanah[3)]
[1)2)3)] Universitas Sains Al-Qur'an, Indonesia
[1)] yogaaricahyadi@gmail.com, [2)]hsibyan@unsiq.ac.id, [3)]nurhasanah@unsiq.ac.id

* yogaaricahyadi@gmail.com

Submitted : 6 April 2026 | **Accepted** : 28 April 2026 | **Published** : 30 April 2026

Abstract: Learning basic nahwu in studying yellow books still faces various challenges, especially in terms of interpreting books objectively and interactively. This study aims to build an automatic book interpretation system using forced alignment based on Wav2Vec2 + CTC Segmentation for audio and text alignment. This system is designed to provide automatic interpretation with audio and text alignment to facilitate the preparation of students in interpreting books and learning nahwu, especially jurumiyah books. The implementation process involves the extraction and pre-processing of audio and text data, audio and text are then aligned using Wav2Vec2 to produce logits output containing the number of samples, frames, and character tokens, then logits are received by CTC to calculate the alignment, manage blank tokens, calculate sequence probabilities and decoding to text to produce a timestamp array. Then the timestamp is validated and normalized and the final result is TextGrid or JSON. Then the results are integrated in an interactive website interface. The results of this study indicate that the forced alignment algorithm using the Wav2Vec2 model is capable of aligning audio and text with a fairly high level of accuracy. This makes it easier for users to understand the contents of the book through segmented audio playback per sentence or chapter. It is hoped that this research can contribute to the development of learning media for Islamic boarding schools' yellow books based on alignment technology.

Keywords: Prealignment, Jurumiyah Book, Forced Alignment, Wav2Vec2, CTC Segmentation

1. INTRODUCTION

The study of yellow books, especially the Jurumiyah Book, is an important part of the Islamic boarding school education system in Indonesia. This book is the main reference in understanding the science of nahwu, which is the foundation for mastering Arabic. However, in practice, many students face difficulties in understanding the meaning and grammatical structure contained in the book, especially for those who do not yet have a strong foundation in Arabic. One factor that influences students' success in understanding the Jurumiyah book is the learning method used, one of which is sorogan. The Sorogan learning method is a traditional learning method applied in Salafiyah Islamic boarding schools (Sholikhun Muhamad, 2018). One method commonly used in

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

Islamic boarding schools is the sorogan and bandongan methods (Murtafiah, 2021). The Sorogan learning method is implemented by; students face directly with the teacher or the teacher of the Jurumiyah book one by one, in turn (Arifin et al., 2022). Therefore, students must prepare the meaning and recitation precisely before meeting the teacher or recipient of the sorogan. Another activity carried out by students in the Sorogan method is to provide specific signs to support understanding of the meaning in the original manuscript of the Jurumiyah book. This activity is commonly called ngesahi/ngabsahi among students (Rahmawati & Negara, 2021). This method is effective in gradually building students' understanding, but it has several limitations, such as dependence on the teacher's presence, varying speeds of comprehension among students, and lack of access to structured recordings of the meaning (Febrian et al., 2024).

Studying the yellow book is a crucial core activity. The yellow book, generally written in Arabic without harakat, requires a deep understanding of Arabic grammar and the ability to translate accurately. However, not all students are able to interpret the book independently, especially those who are just beginning to learn or have not yet mastered the tools of grammar such as nahwu and sharaf. A common problem in Islamic boarding schools (pesantren) is the students' dependence on others, whether they are religious teachers (ustadz) or more knowledgeable friends, to help them interpret the text. In practice, students often have to wait for others' free time just to ask about the meaning of a sentence or paragraph. This presents a challenge because these individuals are also busy with other commitments, such as teaching, studying, or managing other Islamic boarding school activities. As a result, the learning process is ineffective and slow, limiting the students' independence in understanding the text.

This dependence also makes students tend to be passive, simply waiting for explanations from others, without alternative media to guide them independently. Furthermore, limited access to interactive and structured learning methods makes it difficult for students to consistently deepen their understanding. This situation highlights the importance of innovation in the process of learning the text, one of which is the introduction of digital media that allows students to interpret the text independently, whenever they need it. The use of technology such as text and audio explanations, regional language-based translations, and chapter-based meaning playback systems can provide concrete solutions to address this problem. In this way, students are no longer completely dependent on others and can learn independently and flexibly.

In the digital era, developments in artificial intelligence (AI) technology have provided solutions to improve learning efficiency, one of which is through the use of the Forced Alignment algorithm. Forced Alignment is an algorithm that can automatically align text with audio recordings, enabling the system to match each word in the book with a pre-recorded interpretation voice. With this technology, students can learn independently by listening to the interpretation in a structured and interactive manner. The implementation of Forced Alignment on the Jurumiyah Book Interpretation website allows for the automatic presentation of text and audio synchronization, allowing students to understand the book's meaning more efficiently. Furthermore, the website can provide an automatic chapter-by-chapter audio playback feature, which will help students repeat and deepen their understanding according to their individual learning needs.

With this innovation, it is hoped that learning the yellow books, particularly the Jurumiyah Book, can be more easily accessed and understood by students independently without relying entirely on the presence of a teacher. Therefore, this study aims to develop and implement the Forced Alignment algorithm for audio and text alignment on the Jurumiyah Book interpretation website to assist students' learning in Islamic boarding schools.

2. METHOD

This research uses a system development design (research and development) with a software engineering approach. The development model used is the waterfall model because it has a systematic workflow ranging from needs analysis, design, implementation, testing, and maintenance (Wahid, A, 2020). This design was chosen so that each stage of the development of the audio and text alignment system for the interpretation of the Jurumiyah book can be clearly documented, measured, and evaluated. The research targets two groups: data used in system development and system users. The research data includes the text of the text and the interpretation of the Jurumiyah book, which serves as input text, as well as a collection of audio recordings of the book readings that serve as the main material for the forced alignment process. The target users are students and the Safiinatunnaja Islamic Boarding School, who are potential users of the system and play a role in providing feedback on the usability and accuracy of the alignment results. Data collection in this research was conducted through interviews, observations, and literature studies. Interviews were conducted with students and Islamic boarding school administrators to obtain system requirements related to learning methods, the flow of interpretation of the book,

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

and expectations regarding application features (Sugiyono, 2018). Observations were conducted directly during the book learning process in the Islamic boarding school environment to understand user interaction patterns with text and audio (Spradley, J, 2006). The literature study was conducted by exploring references in the form of books, journals, and scientific articles related to learning the yellow book, forced alignment technology, the CTC Segmentation algorithm, and the use of the Wav2Vec2 model in sound processing (Creswell, J.W, 2014). Figure 1 shows the sequential stages of the research process from data collection to system maintenance.

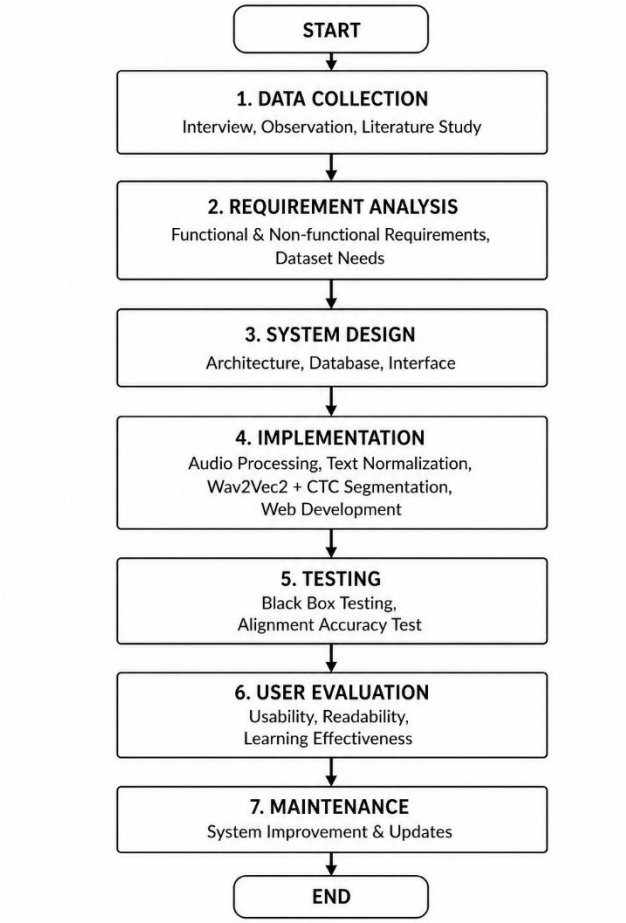

Fig. 1. Research Metodology Flow

The research instruments included a list of interview questions, observation notes, audio processing software, and other supporting tools such as Python, Visual Studio Code, web browsers, and backend development frameworks. The obtained data was then analyzed using two approaches. First, a system requirements analysis was conducted to formulate functional and non-functional requirements, as well as the dataset requirements needed to design a forced alignment system based on Wav2Vec2 and CTC Segmentation. This analysis included text data structures, audio formats, processing capabilities, hardware specifications, and user characteristics (Pressman, R, 2010). Second, a technical analysis was conducted on the forced alignment model. This stage encompassed audio processing, text normalization, acoustic feature extraction using Wav2Vec2, character probability calculations using CTC, trellis construction, forward-backward calculations to determine the best alignment path (Baevski et al., 2020), and determining the timestamp for each word. The alignment results were then validated by checking segmentation accuracy, duration accuracy, and the absence of overlap between segments. Next, the analysis results were integrated into the website using JSON calls to enable synchronous highlighting with the audio. Testing was conducted using a black box method to verify feature feasibility and an alignment accuracy test to assess the accuracy of audio-text alignment. User evaluations were also conducted, involving students and Islamic boarding

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

school administrators, to assess ease of use, interface readability, and the system's effectiveness in supporting the learning process.

3. RESULT AND DISCUSSION

This study produced an audio and text alignment system (Forced alignment) for the interpretation of the book of Psalmist by utilizing the Wav2Vec2 model and the CTC Segmentation algorithm. The developed system allows the audio of the book reading to be displayed in sync with the Arabic Pegon text, so that students can learn the meaning and structure of sentences more interactively. The results of this study include the implementation of alignment, integration into the website, accuracy test results, and functional evaluation of the resulting features. The forced alignment system built is able to produce a stable timestamp for each word in the reading, where the mapping results are then displayed interactively on the web interface. The alignment process is done without manual segmentation, as CTC Segmentation maps audio frames to character sequences by utilizing logits generated by Wav2Vec2. The model shows consistent performance in both short and long-form audio, and provides clear synchronization between voice and text.

Text and Audio Alignment System Design

The system is built by separating the main components into an alignment module and a web display module. The alignment module generates a timestamp per word through a combination of Wav2Vec2 feature extraction and CTC Segmentation. The alignment results are packaged in JSON containing a list of words along with their start time, end time, and duration. In the application interface, the text of the Jurumiyah book along with its Javanese meaning is arranged per chapter. Each word is given a unique identity so that it can be linked to a JSON timestamp. Audio playback uses standard HTML audio elements which are then synchronized with the timestamp. As the audio plays, the system automatically marks words according to chronological order. This display makes it easier for users to read the text while following the audio directly.


```
[
  {
    "tier_name": "Arabic",
    "items": [
      {
        "start": 0.0,
        "end": 1.05,
        "text": ""
      },
      {
        "start": 1.05,
        "end": 1.5625,
        "text": "اَلْكَلَامُ"
      },
      {
        "start": 1.5625,
        "end": 1.6125,
        "text": "هُو"
      },
      {
        "start": 1.6125,
        "end": 1.775,
        "text": "اللَّفْظُ"
      },
      {
        "start": 1.775,
        "end": 1.8875,
        "text": "اَلْمُرَكَّبُ"
```


Fig. 2. Alignment Results

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

Forced Alignment Process Using Wav2Vec2 and CTC

The alignment process begins with processing audio in .wav format, which is then passed through the Wav2Vec2 model to generate Logits based on time frames. These Logits are used as the basis for CTC segmentation to determine words that match acoustic patterns. The CTC algorithm maps text characters to audio frames without requiring manual segmentation. This process generates start and end timestamps, as well as word durations. The segmentation results show that the time distribution of each word is consistent, there is no overlap between segments, and all words fall within the correct audio duration range. Visualization of word durations shows that words with longer phonetic structures have longer durations, while shorter words have shorter durations.

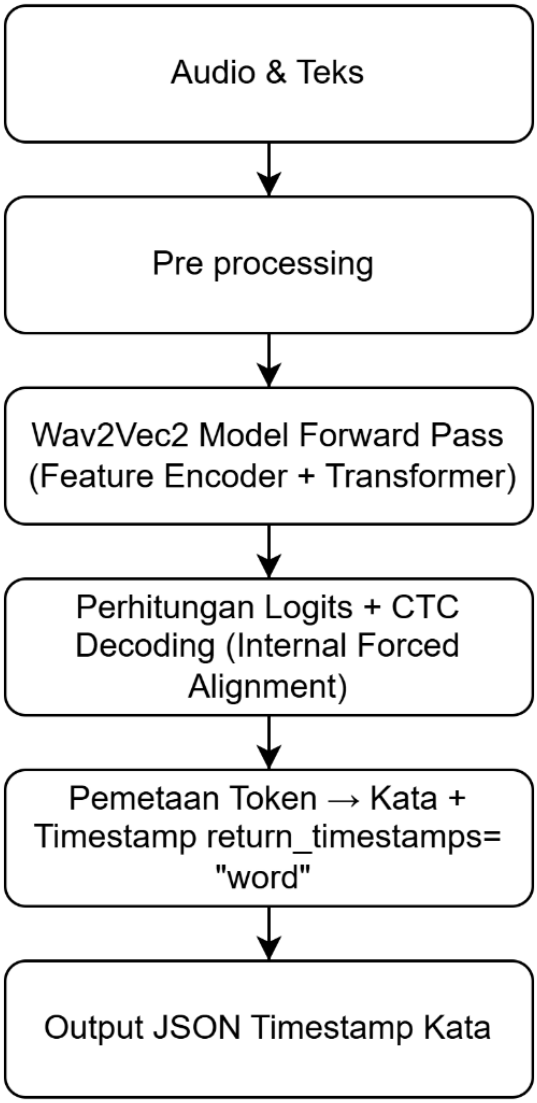

Fig. 3. Forced Alignment

Website Integration

This process integrates alignment results, JSON data, and web interface components to enable the book interpretation system to run interactively. The program code is the main part that regulates the selection and playback of JSON files resulting from forced alignment, the appearance of Arabic text and its meaning, and the

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

synchronization of text with audio. Integration begins with an object structure (babFiles) that stores a list of chapters along with the number of available JSON and audio files. When a user selects a specific chapter, the function (preparebag()) is executed to initialize the application state and save the interface. Next, the data loading process is carried out by (loadBab()), which retrieves each JSON file containing the time-stamp alignment results, filters the tokens, and then displays them on the page in the form of elements () containing Arabic text and Javanese meanings taken from the object (meaningmap). At this stage, synchronization is also performed between each word and the known audio file, so that audio that matches the token can be accessed automatically or manually via the word click feature. This integration ensures that every text element displayed on the web page is directly connected to the original audio using pre-processed alignment results, making the book learning process more intuitive and efficient.

Table 1. Functions and Code Display

No	Parts	Remarks
1	Structure babFiles	Dataset management per chapter
2	Function prepareBab()	Initialization BAB
3	Function loadBab()	Summons JSON + audio
4	Mainingmap & artiAudio	Integration of meaning and audio meaning
5	Word display + meaning	Proof of rendering results
6	Word highlight effects	When auto play is running
7	Control panel display	Play a word, choose a word.
8	Console debug (opsional)	Proof that the JSON file was successfully loaded

System Implementation

The system implementation stage is the stage of realizing the results of the system design into a website that can be operated and run to achieve maximum results according to the system design stages. The dashboard page is the main application display, containing various features accessible to users. On this page, users can select available menus or functions according to their needs. The interface design is kept simple to facilitate user navigation and use of the application.

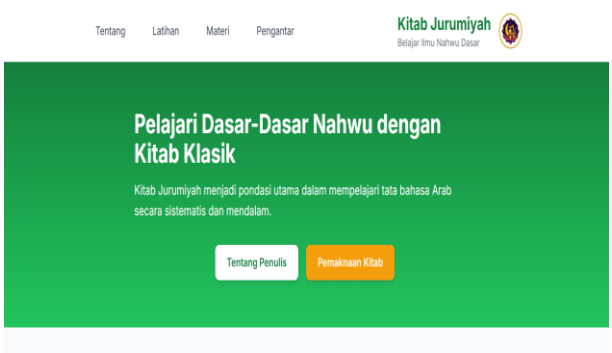

Fig. 4. Dashboard

The "About" page contains a brief explanation of the book Al-Jurumiyah, its history, and its purpose as a foundation for learning grammar. It also provides a general overview of the book's contents and structure, allowing users to understand the context before beginning their study or interpretation.

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

Fig. 5 Description

The Practice Questions page is a feature that provides various questions related to the Jurumiyah book material to help users test their understanding. The questions are neatly arranged so users can practice according to the material being studied. The interface is designed to be simple and interactive to make it easy for users to answer questions, view results, and repeat exercises if necessary.

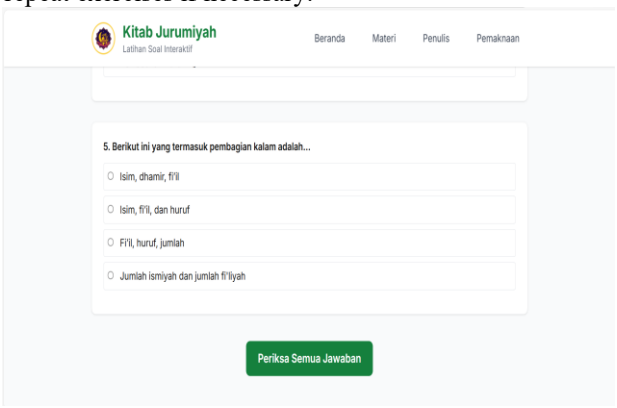

Fig. 6. Practice Questions

The material page provides a brief explanation of each chapter in the Jurumiyah book, complete with an easy-to-understand division of the main topics. On this page, users can select a specific chapter to view the material content, meaning, and basic explanations related to the rules of nahwu. The display is designed to be simple so that users can learn in a structured and comfortable manner.

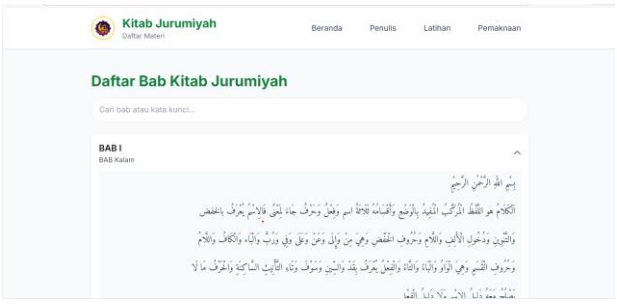

Fig. 7. Material

The Introduction Page of the Jurumiyah Book contains an initial explanation of the book by Imam Ash-Shanhaji which is an important basis in the science of Nahwu. A brief overview of the contents of the book, such as the division of kalam, types of kalimah, and i'rab, so that users get context before studying the following chapters.

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

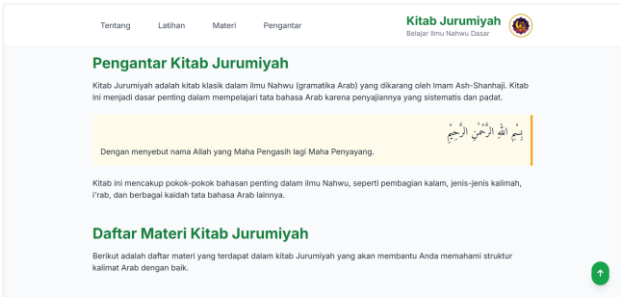

Fig. 8. Introduction

The About the Author page contains brief information about the website's creator, including their identity, educational background or expertise, and their goals in developing this platform for interpreting the Jurumiyah book. The description is presented concisely so users can understand the creator behind the application.

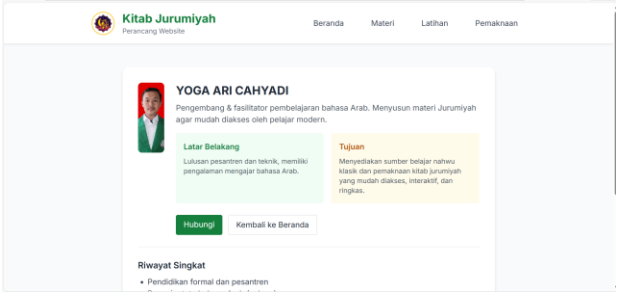

Fig. 9. About the Author

The "*Pemaknaan Kitab*" (Book Meaning) page provides sentence-by-sentence explanations of the Jurumiyah book. On this page, users can select a specific chapter to view the Arabic text along with its Javanese meaning. The system features audio playback and automatic synchronization between text and voice recordings, making it easier for users to follow the learning process in a more interactive and structured manner.

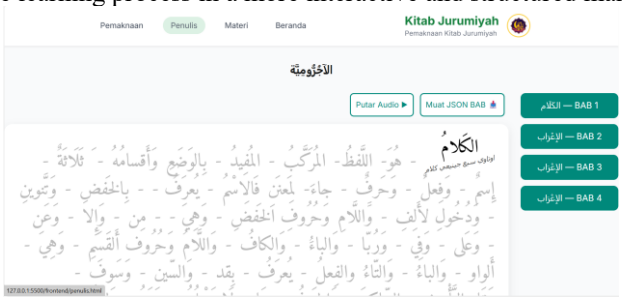

Fig. 10. The Meaning of the Book

The Admin Login page is used to access the website management panel. On this page, administrators enter their username and password to log in. The interface is designed to be simple and secure to facilitate authentication while maintaining data confidentiality. After successfully logging in, admins can manage content, audio, content, and other features on the website.

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

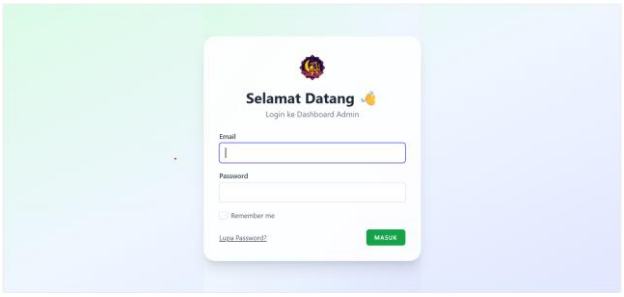

Fig. 11. Admin Login

The Admin Dashboard is the control center for website managers. On this page, admins can view a summary of important data such as the number of materials, audio, interpretations, and recent system activity. The dashboard is designed to be simple for admins to easily access management menus such as chapter management, audio uploads, interpretations, and other settings. This interface makes it easy for admins to efficiently monitor and organize all content.

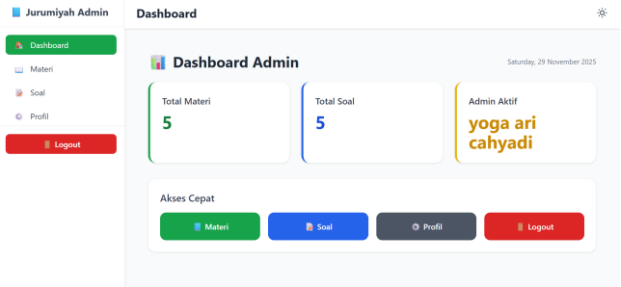

Fig. 12. Admin Dashboard

The Admin Materials page serves as a place for administrators to manage all materials in the Jurumiyah Book. From here, administrators can add, edit, or delete material by chapter, including Arabic text, explanations, and discussion structure. The interface is designed to be simple to manage content quickly and in an organized manner, ensuring that the material displayed in the application is always accurate and up-to-date.

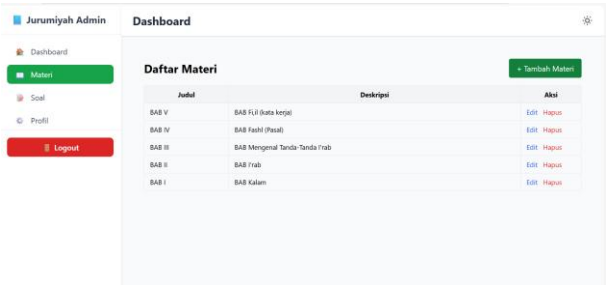

Fig. 13. Admin Material

The Admin Questions page is used by administrators to manage all practice questions in the application. On this page, admins can add, edit, and delete questions based on specific chapters or materials. This feature allows for setting question types, answer keys, and difficulty levels. The interface is designed to be simple to manage questions quickly, structured, and easily monitored, ensuring the practice available to users remains relevant and high-quality.

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

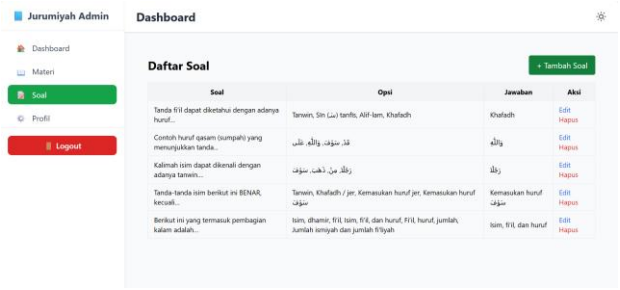

Fig. 14. Admin Questions

The Admin Profile page displays basic information about the administrator account, such as name, email, and other identifying information. On this page, admins can update personal information, change passwords, and set account preferences. The interface is designed to be simple so that admins can easily manage profile data and maintain account security while accessing the system.

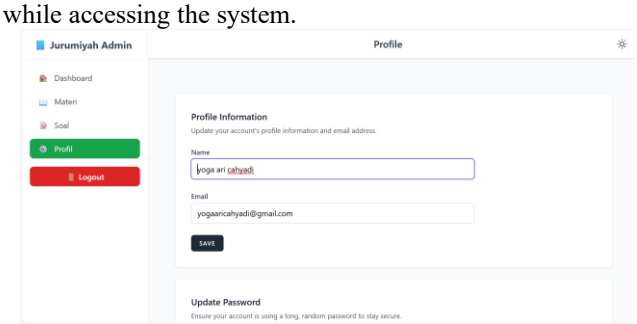

Fig. 15. Admin Profile

Alignment Accuracy Test

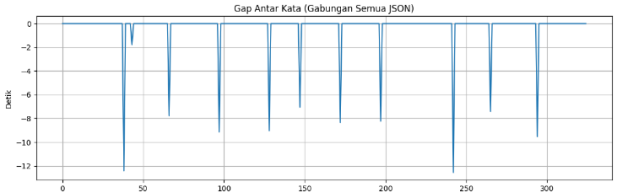

Fig. 16. Gap Graph Between Words

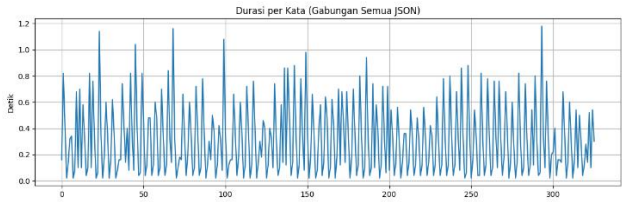

Fig. 17. Graph of Duration Per Word

Alignment test based on how well the existing text matches the audio timestamp using statistical quality metrics for forced alignment:

1. Combined Accuracy Score
2. Total error: 16
3. Accuracy (%): 95.09

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

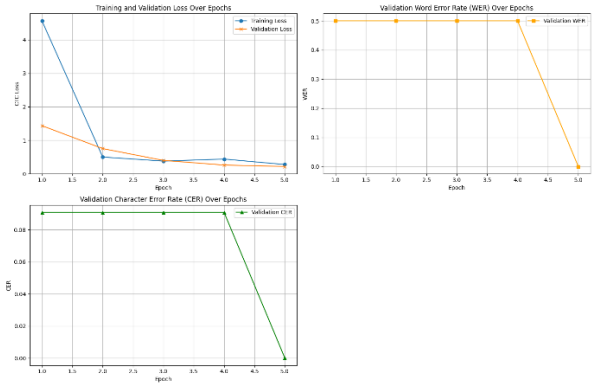

Fig. 18. CTC Loss, WER, CER Graph

Accuracy test how accurate the system is in converting audio into transcription text using ctc loss, wer and cer: for the final results ctc loss: training: 0.48 validation 0.18 wer: long data can reach 0.5 while short data 0.0 cer: 0.09.

Discussion

The results of this study indicate that the implementation of forced alignment using the Wav2Vec2 model combined with CTC Segmentation can effectively support automatic synchronization between audio and text in the Jurumiyah book learning website. The system was able to generate word-level timestamps without requiring manual segmentation, allowing Arabic text, Javanese meanings, and audio playback to be displayed interactively. This result confirms that self-supervised speech representation models such as Wav2Vec2 are suitable for audio-text alignment tasks, especially when the available dataset is limited and the system requires automatic processing of speech signals. The use of CTC-based alignment in this study is also relevant to current developments in automatic speech recognition. Hu et al. (2026) explain that the integration of Wav2Vec2 and CTC in ASR systems is important because CTC enables speech models to handle unsegmented sequence data by mapping audio frames to text tokens. In the context of this research, the CTC mechanism plays an essential role in identifying the most probable alignment path between the audio signal and the corresponding text. Therefore, the success of the system in producing timestamps and synchronized highlighting shows that the Wav2Vec2 + CTC Segmentation approach is technically appropriate for developing interactive learning media based on kitab reading audio.

The alignment accuracy obtained in this study, with a combined accuracy score of 95.09%, indicates that the system can align audio and text with a relatively high level of precision. This finding is strengthened by recent research on speech recognition for low-resource languages. Bootkrajang et al. (2026) found that ASR development for dialectal or low-resource languages requires attention to corpus quality, linguistic characteristics, pre-training, and language-specific information. This is highly relevant to the present study because the system works with Arabic text, Javanese meaning, and pesantren-style reading patterns, which differ from standard speech datasets. Thus, the alignment performance achieved in this study demonstrates that forced alignment can be adapted to specific religious and local learning contexts. From the perspective of audio representation, the use of Wav2Vec2 is appropriate because modern speech systems increasingly rely on self-supervised models to extract robust acoustic features. Ungureanu and Dascalu (2026) state that modern ASR systems commonly include self-supervised encoder-only CTC models such as Wav2Vec2 and XLS-R, convolution-attention models, and sequence-to-sequence models such as Whisper. This indicates that the model selected in this research is aligned with recent trends in speech technology. In addition, McLoughlin et al. (2026) emphasize that spectrogram and speech feature representations remain important in audio and speech analysis because they allow machine learning models to capture meaningful time-frequency patterns. Although this study focuses on forced alignment rather than full speech recognition, acoustic feature extraction remains the foundation of accurate timestamp generation.

The website implementation also shows that forced alignment technology can be transformed into a practical learning tool. The system does not only produce technical outputs in the form of JSON timestamps, but also integrates them into a web interface that supports chapter selection, audio playback, text highlighting, word meaning display, and practice questions. This is important because technology-based learning media should not stop at model development, but must be translated into usable learning features. Ollmann et al. (2026), in their

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

study on self-supervised phoneme tracking for reading assessment, show that speech technology can support reading and pronunciation-related learning by providing automatic tracking of spoken units. This supports the idea that forced alignment can help learners follow text and audio simultaneously, especially in materials that require careful pronunciation and grammatical understanding. The integration of Arabic text and Javanese meaning in this system provides a unique contribution because it adapts speech alignment technology to the traditional pesantren learning environment. In conventional kitab learning, students usually depend on teachers or senior students to understand the meaning and structure of the text. Through this system, students can independently replay each segment, observe the highlighted text, and understand the corresponding Javanese meaning. This supports more flexible and self-paced learning. Xie et al. (2026) explain that speech-based learning systems become more useful when they provide meaningful feedback and guidance to learners, not merely recognition outputs. Therefore, the word meaning display and synchronized playback features in this study can be viewed as an initial form of guided digital learning for kitab interpretation.

The system also has relevance to the development of low-resource speech processing. The Jurumiyah learning context involves Arabic religious text, pesantren reading style, and Javanese translation, which are rarely represented in large public speech datasets. Dar and Pushparaj (2026) show that Wav2Vec2-based ASR can be adapted for low-resource languages by utilizing pre-trained models and fine-tuning strategies. Similarly, Abdulrahman (2026) emphasizes the importance of acoustic feature fusion for low-resource language classification tasks. These studies support the argument that pre-trained speech models can be adapted to specialized linguistic contexts, although additional domain-specific data are still required to improve robustness. However, several limitations should be noted. First, the alignment results depend heavily on the quality of the audio recording, the clarity of pronunciation, and the consistency between the spoken audio and the written text. If the speaker adds, omits, or changes words during reading, the alignment path may shift and reduce timestamp accuracy. Second, the system still requires careful text normalization, especially because Arabic text, Pegon/Javanese meaning, and kitab learning notation may contain special symbols or non-standard orthographic forms. Poncelet and Van Hamme (2026) highlight that subtitle and transcript-based ASR systems require appropriate text handling because differences between spoken audio and written transcripts can affect recognition and synchronization quality. This issue is also relevant to the present system because kitab texts often contain specific formatting and meaning markers.

Another important limitation is related to long audio processing. Although the system can process chapter-based audio, longer recordings may increase the complexity of alignment and the possibility of timestamp drift. This indicates the need for segmentation strategies, audio normalization, and more systematic validation. Vander Eeckt and Van Hamme (2026) show that ASR models need mechanisms to maintain performance when adapting to new tasks, domains, or speakers, especially to reduce degradation when the system is exposed to new data. In this study, this means that future development should include more speaker variations, different recording conditions, and more kitab chapters to ensure that the alignment model remains stable across various learning scenarios. Overall, this research contributes to the development of AI-based Islamic learning media by applying forced alignment technology to the Jurumiyah book interpretation website. The system successfully combines Wav2Vec2, CTC Segmentation, JSON-based alignment output, and interactive web features to support synchronized audio-text learning. Compared with conventional kitab learning, this system offers a more flexible, repeatable, and independent learning experience. Future research can improve the system by expanding the dataset, adding speaker variation, improving text normalization for Arabic and Pegon scripts, and developing automatic feedback for pronunciation or reading accuracy. Such improvements will strengthen the role of forced alignment technology in supporting pesantren-based digital learning.

4. CONCLUSION

Based on the research and implementation results, it can be concluded that the forced alignment method using the Wav2Vec2 + CTC Segmentation model was successfully implemented to support automatic synchronization between audio and text playback. The system was able to determine the timing of each text segment according to the audio without requiring manual timestamps, and the alignment results were relatively accurate. In addition, the developed web system was capable of displaying Arabic text, Javanese meanings, and synchronized audio, thereby supporting more interactive book learning. The results of blackbox testing also showed that all main website features functioned properly, including chapter selection, JSON playback, audio playback, text highlighting, word meaning display, and the practice test system.

Clean Energy and Smart Technology
Volume 04, Number 02, April 2026
DOI : https://doi.org/10.58641/
e-ISSN : 2964-2647

5. REFERENCES

Abdulrahman, A. O. (2026). Pitch-aware multi-feature fusion for classifying statements, questions, and exclamations in low-resource languages. *Computer Speech & Language, 99*, 101941. https://doi.org/10.1016/j.csl.2026.101941

Arifin, A., F., & Hajja Ristianti, D. (2022). Metode sorogan dalam meningkatkan minat dan keterampilan membaca Kitab Kuning Santri Al-Afiyah Bogor Jawa Barat. *Inspiratif Pendidikan, 11*(1), 24–36. https://doi.org/10.24252/ip.v11i1.29195

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS), 33*, 12449–12460. https://arxiv.org/abs/20006.1147

Bootkrajang, J., Inkeaw, P., Chaijaruwanich, J., Taerungruang, S., Boonyawisit, A., Sutawong, B. J. M., Chunwijitra, V., & Taninpong, P. (2026). The development of Northern Thai dialect speech recognition system. *Applied Sciences, 16*(1), 160. https://doi.org/10.3390/app16010160

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.

Dar, M. A., & Pushparaj, J. (2026). A Wav2Vec2 model-based automatic speech recognition system for low-resource Kashmiri language. *International Journal of Speech Technology, 29*(1), 2. https://doi.org/10.1007/s10772-025-10228-7

Febrian, N., Purwanto, P., Syarifah, L., & Muna, N. (2024). Efektivitas metode pembelajaran Sorogan Kitab Jurumiyah di Pondok Pesantren Putri Al Ma'rufiyah Tempuran. *DWIJA CENDEKIA: Jurnal Riset Pedagogik, 8*(1), 83. https://doi.org/10.20961/jdc.v8i1.84564

Hu, H., Tang, C., Tan, P., & Xu, H. (2026). A CTC-based speech recognition network fusing local convolution and global attention. *Sensors, 26*(6), 1865. https://doi.org/10.3390/s26061865

McLoughlin, I., Pham, L., Song, Y., Miao, X. X., Phan, H., Cai, P., Gu, Q., Nan, J., Song, H., & Soh, D. (2026). Spectrogram features for audio and speech analysis. *Applied Sciences, 16*(2), 572. https://doi.org/10.3390/app16020572

Murtafiah, N. H. (2021). Efektivitas penerapan metode sorogan Kitab Al Jurumiyah dalam meningkatkan kemampuan membaca Kitab Kuning. *An Nida, 1*(1), 18–25.

Nijat, M., Wei, Y., & Hamdulla, A. (2026). Perception norm for mispronunciation detection. *Applied Sciences, 16*(7), 3311. https://doi.org/10.3390/app16073311

Ollmann, P., Sonnleitner, E., Kurz, M., Krösche, J., & Selinger, S. (2026). Listen closely: Self-supervised phoneme tracking for children's reading assessment. *Information, 17*(1), 40. https://doi.org/10.3390/info17010040

Poncelet, J., & Van Hamme, H. (2026). Leveraging broadcast media subtitle transcripts for automatic speech recognition and subtitling. *Journal on Audio, Speech, and Music Processing, 2026*, 20. https://doi.org/10.1186/s13636-026-00450-9

Pressman, R. S. (2010). *Software engineering: A practitioner's approach* (7th ed.). McGraw-Hill.

Rahmawati, I., & Negara, T. D. W. (2021). Pelatihan Arab Pegon bagi santri baru guna meningkatkan kualitas pembelajaran Kitab Kuning di Pondok Pesantren Darul Huda Putri. *MA'ALIM: Jurnal Pendidikan Islam, 2*(02), 103–112. https://doi.org/10.21154/maalim.v2i2.3177

Sholikhun, M. (2018). Pembentukan karakter siswa dengan sistem boarding school. *Wahana Islamika: Jurnal Studi Keislaman, 4*(1), 48–64.

Spradley, J. (2006). *Participant observation*. Waveland Press.

Sugiyono. (2018). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Ungureanu, R. D., & Dascalu, M. (2026). Modern speech recognition for Romanian language. *Applied Sciences, 16*(4), 1928. https://doi.org/10.3390/app16041928

Vander Eeckt, S., & Van Hamme, H. (2026). Efficient rehearsal for continual learning in ASR via singular value tuning. *IEEE Transactions on Audio, Speech and Language Processing, 34*, 978–991. https://doi.org/10.1109/TASLPRO.2026.3658931

Wahid, A. (2020). *Rekayasa perangkat lunak dan model waterfall*. Deepublish.

Xie, Y., Zhong, H., Lan, X., & Dong, W. (2026). Mispronunciation detection and diagnosis based on large language models. *Computer Speech & Language, 99*, 101942. https://doi.org/10.1016/j.csl.2026.101942