

Speech Recognition-based for Learning English Pronunciation using the Transformers Wav2Vec 2.0 Algorithm

Megan Medellin¹, Dian Sa'adillah Maylawati², Eva Nurlatifah³, Muhammad Deden Firdaus⁴, Ichsan Taufik⁵

^{1,2,3,4,5}Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati, Bandung, Indonesia

2Email: diansm@uinsgd.ac.id

Proceedings of the Kalijaga International Conference on Education 2026

Abstract. English pronunciation remains a major challenge for Indonesian learners, largely due to phonetic differences and limited exposure to native-like speech. To address this issue, this study proposes an interactive learning system that integrates automatic speech recognition (ASR) technology to assess and support learners' pronunciation skills. This study introduces a speech recognition-based learning technology algorithm to support pronunciation recognition for English. This study used the Transformer-based Wav2Vec 2.0 model to recognize speech and convert into narrative text and conduct a reading assessment automatically. The fine-tuned Wav2Vec2 model demonstrated excellent ASR performance with a final Word Error Rate (WER) of 0.042778 (approximately 4.28%) on validation data, demonstrating high accuracy in speech transcription. The model also demonstrated a significant and consistent reduction in training loss and validation loss across 24,000 training steps, indicating the model consistently shows reductions in both training and validation loss, reflecting effective learning, model stability, and strong generalization capability across unseen data. Furthermore, a gamification mechanism, including learning progress tracking and a feature to "unlock" stories based on a minimum score of 80%, was successfully implemented, effectively motivating users to practice. The accuracy across real-world user interactions in the system indicates the sensitivity to audio quality and user pronunciation. Overall, the system achieves its goal of providing an interactive and innovative learning tool for English pronunciation utilizing natural language processing technology.

Keywords: English Learning; Natural Language Processing; Speech Recognition; Transformers; Wav2Vec

Abbreviations: automatic speech recognition (ASR);

INTRODUCTION

Problems in the teaching and learning process are unavoidable in educational activities at all levels. One of the challenges we frequently encounter is learning English as a foreign language in Indonesia[1]. As an international language, English is a crucial skill for students to master in the era of globalization. However, because Indonesians don't speak English on a daily basis, learning it becomes difficult. Complicated English pronunciation, difficult to remember vocabulary, and confusion in the use of grammar are often obstacles that cause students to lose confidence and have difficulty expressing what they want to say when speaking [2]. The EF English Proficiency Index (EF EPI) 2024 showed that Indonesians still have a low level of English proficiency, ranking 80th out of 116 countries, which reinforces this issue [3]. Java Island has the highest proficiency level, while Maluku has the lowest. Surabaya City has the highest score, while Jambi City has the lowest [3]. Data shows that we need further efforts to improve the English proficiency of people in various regions of Indonesia. Emma Walton, EF's director of academic affairs, stated that "Indonesians' English proficiency is still low," highlighting the need for more efficient learning methods. EF EPI 2024 research, involving 2.1 million participants from 116 countries and territories, shows the level of English proficiency around the world. The number of participants ranged between 18 and 60 years, with 54% women and 46% men, with an average old is 26 [3].

The lack of ability in speaking English among Indonesian students is a multifaceted issue influenced by various linguistic, psychological, and socio-environmental factors [4], [5], [6]. Despite English being a part of the national curriculum from elementary school through university, many students struggle with speaking proficiency. This challenge is compounded by psychological barriers such as anxiety and lack of confidence [7], [8], [9], [10], linguistic difficulties including limited vocabulary and grammar issues [11], [12], [13], and socio-environmental factors like insufficient practice opportunities and teacher-centered instruction [14], [15], [16]. Addressing these challenges requires a comprehensive approach that includes both pedagogical strategies and supportive learning environments [11], [14], [15], [17].

Low motivation to learn, a lack of student awareness of the importance of English mastery, and a lack of varied teaching methods can all contribute to this learning challenge. Learning that relies too heavily on textbooks and traditional lectures often makes it difficult for students to concentrate and prevent them from actively engaging in the learning process. Furthermore, a lack of support from their social and family environment is also a significant contributing factor [18]. As in the teaching materials in the 8th-grade junior high school English textbooks, according to the Merdeka Curriculum, which contains lessons that study narrative texts in English, namely texts that tell a series of events chronologically and are interconnected. This curriculum expects each student to be able to read, understand, and interpret stories provided by the teacher or book, and to be able to express their ideas and feelings effectively through stories [19].

Artificial intelligence (AI)-based speech recognition technology is one innovative way to improve pronunciation skills. A Transformer-based algorithm model developed by Facebook AI Research (FAIR), Wav2Vec 2.0, can be used to accurately identify and analyze student speech [20]. Even with different pronunciation data, this model has proven its success in many automatic speech recognition (ASR) tasks [21]. Wav2Vec 2.0 allows applications to provide immediate feedback on student pronunciation errors by extracting speech representations from audio input and automatically converting them to text. Advances in speech recognition technology are significant, especially for foreign language learning. One of the leading approaches today is the Wav2Vec2.0 model, a Transformer algorithm developed by Facebook AI Research (FAIR). This model uses self-supervised learning techniques, enabling it to automatically learn from large amounts of audio data before fine-tuning with labeled data.

Previous research has shown that technology-based speech recognition and language learning have experienced significant progress through various machine learning and deep learning approaches. The Wav2Vec 2.0-RNN-based model has proven highly effective, achieving 99.36% accuracy and 100% validation, and significantly reducing the word error rate [22]. The development of ASR is also strengthened through the integration of deep transfer learning, pre-training, fine-tuning, Federated Learning, and TSTNN models which improve the accuracy of speech recognition [23]. The RNN-CTC approach shows better performance than GMM-HMM and CNN-CTC in terms of accuracy, testing time, and word error rate [24]. In the context of non-native child speech, Whisper outperforms Wav2Vec 2.0, especially on read speech and native speakers [25]. In addition, the Naive Bayes, CNN, DBN-SVM, and 1D CNN-BiGRU-Attention combination methods are also widely used for English language learning applications, accent recognition, word recognition, and pronunciation error detection, with relatively high accuracy, such as 92.26%, 97.06%, and CNN increases from 90% to 96% when the dataset is enlarged [26], [27], [28], [29], [30], [31]. Several other studies focused on the development of interactive learning applications based on multimedia, augmented reality, speech recognition, quizzes, pronunciation, and answer evaluation, which showed very good user acceptance and speech input accuracy reaching 94% [27], [32], [33]. In general, these findings confirm that the integration of ASR, deep learning, and interactive learning media has great potential in improving the effectiveness of English learning, especially in the aspects of pronunciation, word recognition, automatic evaluation, and user learning experience

This study fine-tunes Facebook's pretrained Wav2Vec 2.0 base model (facebook/wav2vec2-base). This is a key distinction from most previous research. While previous research has also focused on developing English language learning tools and speech recognition (ASR) systems using Wav2Vec 2.0 and Deep Learning to improve accuracy, this study specifically utilizes the fine-tuning process on Facebook's Wav2Vec 2.0 base model. This approach is expected to produce more innovative and effective English pronunciation learning media, while significantly improving the user learning experience.

METHODS

Datasets

Figure 1 provides the process for preparing the datasets. Begin by collecting the data, validating, cleaning, labeling, constructing, and designing the vocabulary. The data collection process focused on acquiring a specific subset of the LibriSpeech dataset, the LibriSpeech ASR train-clean-100. This dataset is renowned for its high-quality collection of English audio and transcripts derived from audiobooks and has become a standard in ASR research due to its scale and cleanliness. The train-clean-100 subset was selected based on its wide accessibility, clear audio quality, and a general English representation suitable for model fine-tuning. In total, 28,539 audio samples were successfully loaded from this dataset. Each sample consists of an audio file in .flac format and a corresponding text transcript in a .trans.txt file. A local overview of the dataset's directory structure, showing the organization of the audio files and transcripts, is shown in Figure 2.

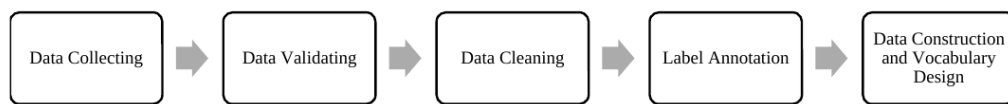


Figure 1. Data preparation process

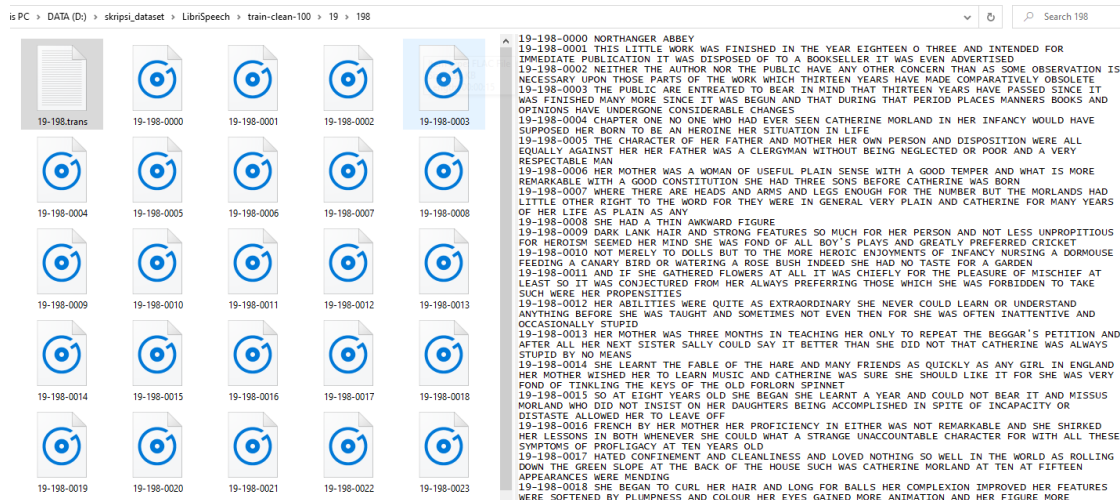


Figure 2. The example of datasets

After data collection, the next step is validation and review to ensure its quality and readiness. In this phase, the data is checked for completeness and basic adjustments are made, including preparation for splitting into training and validation subsets. All audio data in the dataset is resampled to a standard sampling rate of 16,000 Hz. The initial review also ensures that each audio file has a corresponding transcript and that no data is corrupted or significantly missing. This process also includes automatic filtering of audio samples with missing paired transcripts or invalid transcript lines, ensuring dataset integrity. Furthermore, the validated data is then divided into a training subset (90%) and a validation subset (10%) to prepare for the model training and performance evaluation stages.

In the context of the LibriSpeech train-clean-100 dataset, data cleaning primarily involves addressing potential malformed lines or missing files during transcript parsing, as indicated in the training log, which records any warnings. Although the dataset is relatively clean, this step ensures that only valid and complete audio-transcript pairs are processed further. This process serves as a second validation step to ensure the quality of the model input. Furthermore, for Speech-to-Text tasks, data labels are inherently text transcripts of the associated audio files. In the LibriSpeech dataset, these transcripts are already available in .trans.txt files that accompany each audio file. Therefore, the label annotation stage focuses on parsing and mapping the existing text transcripts so they can be accessed and used as labels by the training framework. Manual annotation is not required.

After basic label cleaning and annotation, the data is constructed into a format specific to the Wav2Vec2 model. This process involves vocabulary building and deep audio/text preprocessing. The first step in data construction involves building a vocabulary (a kind of character dictionary) from the text transcripts in the training data. This process involves extracting all unique characters that appear in the transcripts. Special characters that are important to the Wav2Vec2 model's operation are then added: spaces (), used as word boundaries; Unknown Tokens ([UNK]), representing unknown or missing characters in the vocabulary; and Padding Tokens ([PAD]), used to equalize the length of data sequences within a single batch. After this process, the successfully constructed vocabulary has a size of 30 characters. This vocabulary is then saved in JSON format as vocab.json for use by the tokenizer.

The audio and text data undergo a fairly detailed preprocessing process using the Wav2Vec2Processor from the Hugging Face library. This Wav2Vec2Processor integrates functions from Wav2Vec2FeatureExtractor and Wav2Vec2CTCTokenizer, enabling a unified processing flow. First, Wav2Vec2FeatureExtractor converts the raw audio signal into a model-understandable numeric feature representation, including normalization and dimensionality adjustment steps. Second, Wav2Vec2CTCTokenizer translates the text transcript into a sequence of numeric IDs (often called tokens) based on the constructed vocabulary. Third, to prepare the data for efficient batch processing by the GPU, DataCollatorCTCWithPadding is used. This function ensures that all audio sequences and labels in a batch are of uniform length by adding padding tokens as needed. Furthermore, the padding token in the label section was changed to -100 to be ignored by the CTC loss function during training. This data preprocessing process was performed comprehensively on the entire training and validation datasets. Although initial experiments

involved multiprocessing, the process was ultimately run as a single process, taking approximately 5 minutes and 21 seconds for the training data (25,685 samples) and 30 seconds for the validation data (2,854 samples).

Transformers Architecture

Transformers are a network architecture that relies on self-attention mechanisms to obtain representations of input and output sentences during the training process [34], [35]. This neural network was designed to overcome the traditional way of processing sequential data, which is often slow and loses fine details in long contexts. Self-attention is a mechanism that allows the model to look at different parts of a sequence of data simultaneously and determine which parts are most important. This capability makes Transformer more efficient, allowing it to be trained on larger datasets, and more effective at handling long texts where context can significantly impact meaning. In practice, Transformers are often used to solve problems like machine translation, which require encoder-decoder layers [34]. The encoder reads the entire sequence of input data, such as an English sentence, to extract features and convert them into a compact mathematical representation. This representation is a summary that captures the essence of the input. The decoder then takes this summary and, step by step, produces a sequence of outputs, such as a sentence translated into another language. An illustration of the encoder-decoder layer in machine translation in the Transformer architecture is shown in Figure 3.

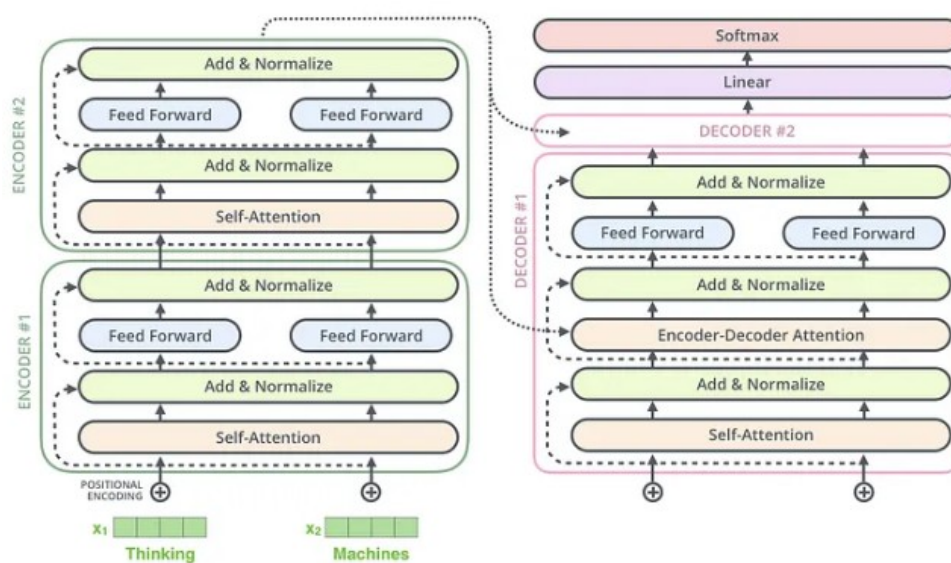


Figure 3. Transformers architecture [36]

Unlike RNN models, which process each word sequentially, self-attention algorithms do not inherently consider word order. Therefore, before entering the self-attention layer, the model adds a positional encoding vector to each input word embedding. This positional encoding provides information about word position, distance, and order between words, allowing the model to identify the location of each input word. The output from the self-attention layer, which has the same size as the input, is combined with the original value before entering the self-attention layer. This combined value then becomes the input to the feed-forward layer, which processes each word individually.

The feed-forward layer has the same number of output neurons as the number of input neurons. The outputs of these feedforward layers are then recombined, and this cycle repeats until the encoder layer is complete. The output of the final encoder layer is then used in the encoder-decoder attention mechanism within the decoder layer. This decoder layer receives the feature extraction results from the encoder and the target values.

Next, the model requires a layer to make concrete predictions, such as selecting the next word in a sequence. The linear layer serves to map the vector space to the original input domain, producing a set of scores, or logits, for each possible token. The softmax function is the final stage that converts these logit scores into a probability distribution. Each element of the softmax output indicates the model's confidence level regarding a particular class or token. The use of the Transformer architecture has been proven to achieve high performance scores compared to other models for machine translation tasks. [37].

Wav2Vec 2.0 Architecture

Wav2Vec 2.0 is an innovative self-supervised learning model framework, developed by Facebook AI (now Meta AI), specifically for speech representation learning [15]. This model offers a significant breakthrough in speech processing due to its ability to be trained on large amounts of raw audio data, even

hundreds to thousands of hours, without the need for transcription labels [15]. This self-learning approach allows Wav2Vec 2.0 to learn deep linguistic and acoustic patterns implicitly from the audio data itself.

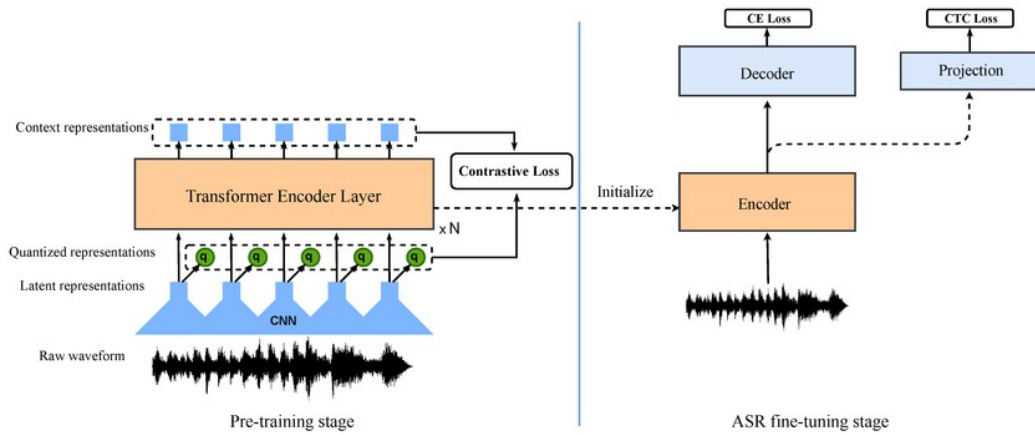


Figure 4. Arsitektur Wav2Vec 2.0 [36]

The core architecture of Wav2Vec 2.0 starts with a convolutional neural network (CNN)-based feature encoder, which transforms a raw audio signal or waveform (X) into a more compact and abstract latent speech representation (Z). This feature representation is then quantized into a series of discrete units called quantized representations (Q), resembling phoneme units in human language, thus creating a discrete representation of speech. The core of the architecture is then a Transformer encoder that processes Q (after being partially masked) and Z to produce context representations (C). In the pre-training phase, a small portion of the units in the discrete representation Q are intentionally masked or hidden. The model is then trained to predict these hidden units based on the context of the surrounding unmasked units. This process is reinforced by a contrastive loss (mathcal{L}) penalty mechanism, which encourages the model to distinguish between correct units and “negative” units (randomly incorrect units), ensuring that the learned representations are truly relevant to the speech structure. A visualization of this architecture can be seen in Figure 4 [38].

The result of this extensive self-training is a model that already “understands” the structure and intrinsic characteristics of speech. The representations generated by Wav2Vec 2.0 can be used as a powerful pre-trained model, which can then be fine-tuned. The fine-tuning process involves adapting this pre-trained model with a small amount of labeled data (e.g., a few minutes or hours of transcription) for a specific downstream task, such as ASR. This capability significantly reduces the need for expensive and time-consuming labeled data, making it highly efficient and flexible for a wide range of speech processing applications [39].

Word Error Rate Matrix

Word Error Rate (WER) is a key metric researchers often use to evaluate the performance of automatic speech recognition (ASR) or machine translation systems. WER measures how accurately a system generates a sequence of words compared to a correct reference [40], [41], [42]. WER values typically range from 0 to 1 or higher, with 0 indicating perfect agreement between the system's results and the reference, while 1 or higher indicates complete discrepancy. The lower the WER value, the better the system's performance. A major challenge in measuring the performance of ASR, or machine translation, systems arises because the word sequences they generate can differ in length from the expected reference word sequence. To address this, researchers derived the WER from the Levenshtein distance, a measure that calculates the difference between two word sequences at the word level, rather than the phoneme level.

The WER calculation process involves aligning the sequence of words that the system recognizes with the sequence of reference words using a dynamic string alignment algorithm. Word Error Rate with formula (1) as follows:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (1)$$

Where:

S is the number of words that have been substituted (replaced).

D is the number of words that have been lost (deleted).

I is the number of words that have been added incorrectly (inserted).

C is the number of correct words.

N is the total number of words in the reference sentence ($N=S+D+C$).

RESULTS AND DISCUSSION

Results and Discussion section contains the explanation of the experimental setting, Wav2Vec 2.0 tuning, the Transformers base model of ASR, the Transformers model with fine-tuned Wav2Vec 2.0, and the evaluation results. Finally, the analysis and interpretation of research findings are available in the Discussion.

Result

Experimental Setting

Data partitioning is crucial for ensuring objective model evaluation. The collected and validated LibriSpeech dataset is then divided into subsets for model training and validation. The Training Set, consisting of 25,685 samples (90%), trains the ASR model, teaching it speech and text patterns. The Validation Set, consisting of 2,854 samples (10%), evaluates model performance during training, monitors loss and WER, and prevents overfitting. While the Test Set is not explicitly used in the main training cycle reported in the logs, in a complete model development cycle, test data is typically set aside from the beginning and used only once at the end to provide an unbiased evaluation of model performance. In the context of application implementation, user-speech data can serve as test data to assess real-world performance.

Wav2Vec2 Model Tuning

The model tuning step involves adjusting or selecting a base model to be fine-tuned. In this study, the base model used was facebook/wav2vec2-base. This model was selected based on its reputation as a highly effective Transformer-based ASR model and its training on large-scale audio data. This initial tuning process ensured that the base model had an architecture suitable for the fine-tuning task on the LibriSpeech dataset.

The Wav2Vec2 model fine-tuning process was performed using Trainer from the Hugging Face library. This stage is the core of the modeling process, where the pre-trained model is adapted to the specific ASR task on the dataset used. The training parameter settings are as follows. The model's output directory is set to ./results, where all training results are stored. The model was trained for 15 epochs, meaning the entire training dataset was processed 15 times. Each GPU processed 8 samples for each training step, with gradient_accumulation_steps=2 set to accumulate gradients for two steps before updating the model weights.

This effectively increases the total batch size to 16, allowing for larger batches without significantly increasing GPU memory requirements. The learning rate was set at 1×10^{-4} , a common value for fine-tuning Transformer-based models. The initial 500 steps were dedicated as a warmup phase for the learning rate, where the learning rate was gradually increased to stabilize the initial training process. Training was performed using the fp16=True (16-bit Floating-Point) format for memory efficiency and computational speed. Model performance evaluation and model checkpoint storage were performed every 100 steps. The Word Error Rate (WER) was chosen as the metric to determine the best model; since a lower WER means better performance, the greater_is_better parameter was set to False. Finally, after the entire training process was completed, the model with the best WER performance on the validation data was automatically loaded.

Evaluation Result

The training process for a Wav2Vec2 model fine-tuned from the facebook/wav2vec2-base model using the LibriSpeech train-clean-100 dataset showed a clear performance improvement over 15 epochs, equivalent to 24,000 training steps. Table 1 displays the training log of this fine-tuned model, including Training Loss, Validation Loss, and Word Error Rate (WER) recorded every 100 evaluation steps. From Table 1 and Figure 5, it is clear that the model shows a significant reduction in loss on both the training data (Training Loss) and validation data (Validation Loss). The Training Loss continues to decrease from 0.2497 at step 1000 to 0.0226 at step 24,000, indicating that the model has successfully learned the patterns in the training data. Similarly, the Validation Loss also shows a consistent downward trend, from 0.188501 at 1000 steps to a low of around 0.081893 (11000 steps), then increasing slightly and stabilizing at around 0.084762 at 24000 steps. This trend indicates that the model maintains its generalization ability well on previously unseen data, indicating a stable and effective learning process.

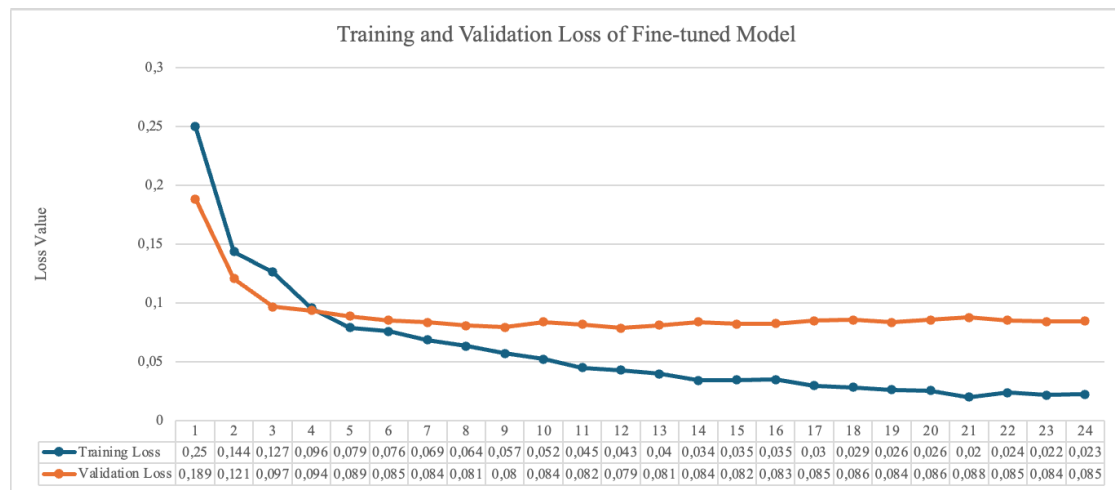


Figure 5. Fine-tuned model result

Table 1. Experiment results

Steps	Fine-tuned Model	Fine-tuned Model	Fine-tuned Model	Base Model	Base Model	Base Model
Steps	Training Loss	Validation Loss	WER	Training Loss	Validation Loss	WER
1,000	0.2497	0.188501	0.161743	0.4905	0.144662	0.144252
2,000	0.1436	0.120844	0.099517	0.1118	0.088665	0.099466
3,000	0.1265	0.096935	0.081814	0.0779	0.078644	0.088452
4,000	0.0959	0.093724	0.074401	0.0658	0.075186	0.082379
5,000	0.0793	0.088964	0.066089	0.0555	0.071694	0.07901
6,000	0.0762	0.085287	0.064072	0.0456	0.074296	0.077528
7,000	0.0688	0.083682	0.062629	0.0414	0.073111	0.074986
8,000	0.0637	0.080841	0.058907	0.0357	0.075463	0.074159
9,000	0.0572	0.079678	0.056043	0.0318	0.076516	0.072504
10,000	0.0523	0.083958	0.05571	0.0278	0.07602	0.071425
11,000	0.0453	0.081893	0.052532	0.0262	0.076201	0.070679
12,000	0.0432	0.078816	0.053299	0.0243	0.075663	0.070083
13,000	0.0399	0.081347	0.050172	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
14,000	0.0343	0.084127	0.050535	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
15,000	0.0348	0.0822	0.049365	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
16,000	0.0352	0.082749	0.048124	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
17,000	0.0298	0.085162	0.047408	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
18,000	0.0287	0.085642	0.046147	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
19,000	0.0264	0.083777	0.045431	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
20,000	0.0256	0.085728	0.044544	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
21,000	0.0201	0.087728	0.043484	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
22,000	0.0241	0.085467	0.043283	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
23,000	0.0218	0.084336	0.04297	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps
24,000	0.0226	0.084762	0.042778	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps	Base model is stopped at 12,000 steps

Steps	Fine-tuned Model	Fine-tuned Model	Fine-tuned Model	Base Model	Base Model	Base Model
Average	0.06020833	0.09067283	0.06004154	0.08619167	0.08217675	0.08374358

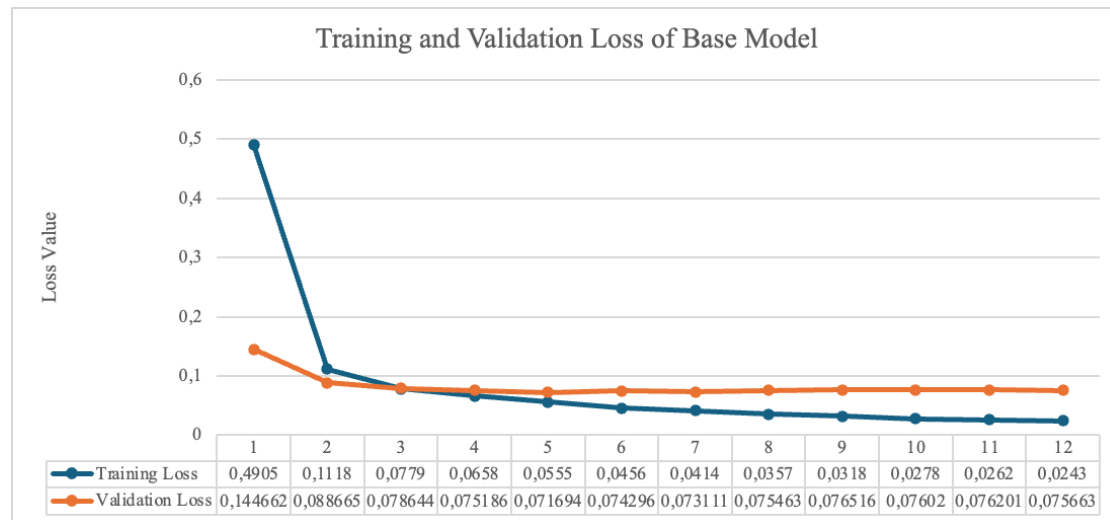


Figure 6. Base model result

As a reference and to further validate the effectiveness of the fine-tuning, a training log from the baseline Wav2Vec2 configuration was analyzed. This configuration also utilizes the LibriSpeech train-clean-100 dataset, but with some differences in text preprocessing (converting text to lowercase) and training parameters (e.g., batch size 32, weight decay 0.005, warmup steps 1000, and logging every 500 steps). Figure 6 shows the training log of the baseline model, which ran for up to 12,000 steps. Similarly, the WER decreased drastically. The WER decreased from 0.161743 (approximately 16.17%) at the 1,000th step to 0.042778 (approximately 4.28%) at the end of training (24,000 steps). This very low WER demonstrates the model's high accuracy in transcribing audio from the LibriSpeech dataset, with an average of only about 4.28% of words being misrecognized. Generally, achieving a WER below 5% is considered excellent for ASR tasks.

Table 1 and Figure 6 show that the baseline model also exhibited a significant reduction in training loss and validation loss as training progressed. The base model's training loss decreased from 0.4905 at 1,000 steps to 0.0243 at 12,000 steps. Meanwhile, the validation loss decreased from 0.144662 at 1,000 steps to 0.075663 at 12,000 steps. Although the base model's initial loss may have been higher than that of the fine-tuned model in this project, both losses consistently decreased until the end of training, demonstrating the base model's learning ability on the dataset.

The base model achieved a WER of approximately 0.070083 (approximately 7.01%) at 12,000 steps. The difference in the number of training steps between this base model (up to 12,000 steps) and the fine-tuned model in this project (up to 24,000 steps) is because the base model's training configuration may have been designed for a shorter duration, or the available logs only covered that range.

Deployment Result

The model deployment plan includes preparing and configuring the environment for the ASR model to function within the application system. The fine-tuned Wav2Vec2 model is loaded into the Flask backend environment upon application startup. The model and processor are stored locally at a predefined path (`./fine_tuned_model_100ds-15`). This loading process occurs once upon application server startup, ensuring the model is ready for inference whenever a transcription request is received. The model is also switched to the GPU to accelerate the inference process, which is essential for real-time response in the application. The backend is developed using the Flask framework (Python), which acts as an API server. The backend is responsible for the core business logic. This includes user management, which handles registration, login, and session authentication, as well as user administration features (adding, changing roles, and deleting) specifically for administrator accounts. The backend also manages Narrative Text story content through CRUD (Create, Read, Update, Delete) operations accessible to administrators.

In addition, the backend handles audio processing. It receives audio files uploaded from the frontend via the `/api/upload-audio` endpoint, temporarily stores them, and prepares them for further processing. The integration of a Speech-to-Text model forms the core of its functionality, where a fine-tuned Wav2Vec2 model is loaded and ready to run. The `/api/transcript-audio` endpoint receives the audio, processes it through the model, and generates a text transcription. Finally, the backend interacts with a MySQL database to store and retrieve user data (users), stories (cerita), and user learning outcomes (hasil_belajar). The frontend is a

web-based user interface built using HTML, CSS, and JavaScript. The frontend acts as a client that communicates with the backend via an API. Key frontend features include a landing page that introduces the application and its key features, as seen in Figures 7 and 8.

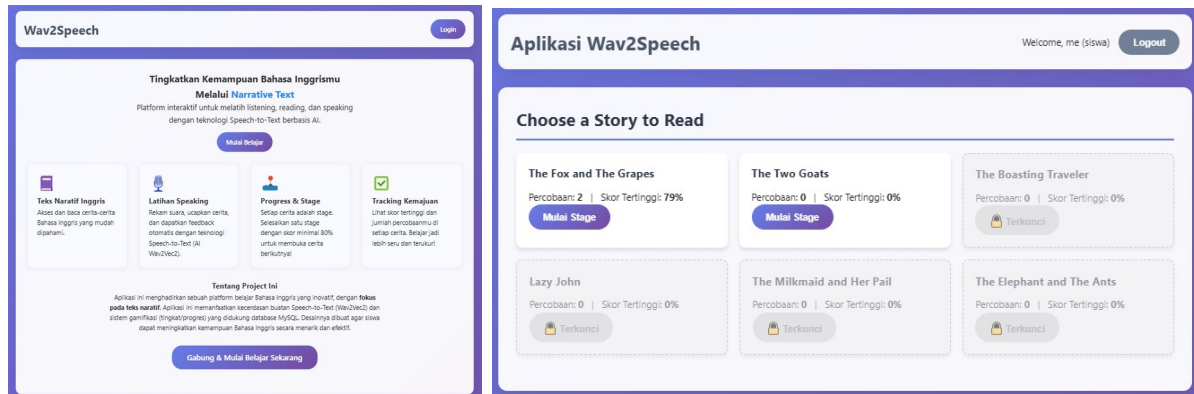


Figure 7. Example of landing page and history of learn

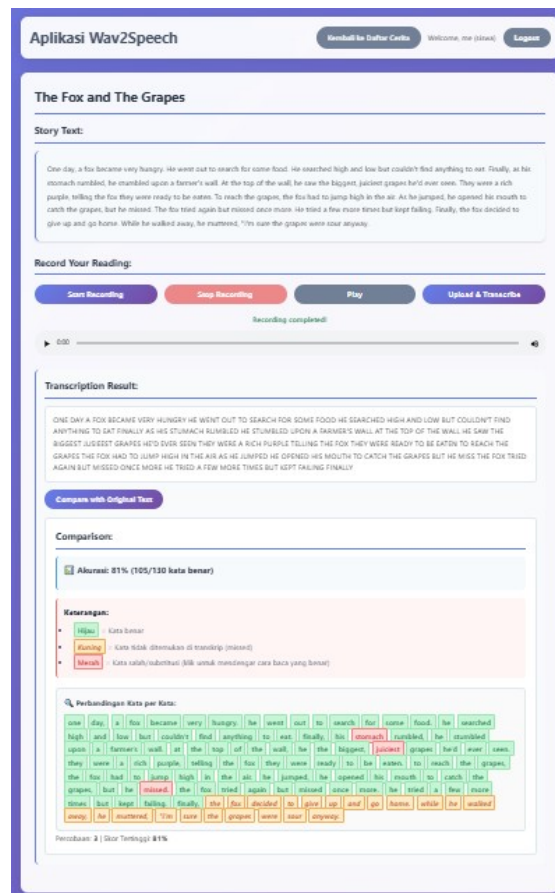


Figure 8. Example of reading detection

Discussion

A direct comparison between the two models shows that the project's fine-tuned model achieves superior performance compared to the baseline model, particularly in the WER metric. While the baseline model demonstrated good learning ability at 12,000 steps, the fine-tuned model achieved a significantly lower WER (approximately 4.28%) at 24,000 steps compared to 7.01% achieved by the baseline model. The comparison graph clearly illustrates this. The Training Loss and Validation Loss curves of the fine-tuned model (orange line) show consistently lower values than those of the baseline model (blue line), indicating that the fine-tuned model learns data patterns more efficiently and accurately. Similarly, the WER curve of the fine-tuned model also lies below the baseline model's curve throughout most of the training process. The faster loss reduction and lower WER of the fine-tuned model confirm the effectiveness of the proposed fine-tuning methodology and the impact of more optimal fine-tuning parameters in this study. The effectiveness of fine-tuning methodologies in reducing loss and lowering WER in speech recognition models is well-documented across various studies [43]. Fine-tuning, when optimized with appropriate parameters and strategies, significantly

enhances model performance, especially in low-resource and specialized contexts. This is achieved through techniques such as hyperparameter optimization [44], [45], selective parameter adaptation [46], and innovative training strategies [47], [48].

The review of the modeling process not only looks at quantitative metrics but also analyzes performance qualitatively. A qualitative evaluation of the model was performed by sampling 10 random samples from the validation dataset. The system took each audio sample along with its original text transcript from the validation dataset. The audio data from each sample was then passed through a trained Wav2Vec2 model. The model generated logits (raw model output) representing character probabilities. These logits were converted into model-predicted token IDs, which were then decoded into a readable text transcription. The model-generated transcriptions were compared to the original (reference) text, and a Word Error Rate (WER) was calculated for each sample, providing a specific error metric at the word level. The resulting model transcription, the original text, and the WER value were displayed for each sample, allowing identification of the types of errors (substitutions, insertions, deletions) likely to occur. This analysis provided a comprehensive understanding of the model's accuracy on real-world examples. The experiment results are proven on 10 samples from the validation dataset to identify the strengths and weaknesses of the model's transcription in more detail (Table 2).

Table 2. Proof of experimental results

No	Original text	Transcript model	WER	Mismatched Words
1	"were before him it was a part of his destiny to become the champion of the anti slavery cause in the national legislature"	"were before him it was a part of his destiny to become the champion of the anti slavery cause in the national legislature"	0	nothing
2	"silvered and pearled on every scented still spray the sky was deep clear strong blue above and against the blue the wallflowers shone bravely from the cracks and crevices of ruined arch and wall and buttress"	"silvered and pearled on every scented still spray the sky was deep clear strong blue above and against the blue the wallflowers shone bravely from the cracks and crevices of ruined arch and wall and buttress"	0	nothing
3	"the magistrate will want your evidence and a summons will be sent you all right then address it to papillon eugene driver"	"the magistrate will want your evidence and e summons will be sent you all ight then address it to papion ugene driver"	0,1818	and a (become e), all right (become all ight), papillon (become papion), eugene (become ugene)
4	"no higher prompting than conquest for conquest's sake and mere race glorification to go far and to endure it must have behind it an ethical impulse a sincerely conceived righteousness"	"no higher prompting than conquest for conquest's sake and mere race glorification to go far and to endure it must have behind it an ethical impulse a sincerely conceived righteousness"	0	nothing
5	"not absolutely based on a knowledge of facts is easily banished for a time by a bright sunny day or some happy outward circumstance and when the brilliant fourteen fine days of october came on"	"not absolutely based on a knowledge of facts is easily banished for a time by a bright sunny day or some happy outward circumstance and when the brilliant fourteen fine days of october came on"	0	nothing
6	"but that would not help for long unless they could abolish the love of it ants seem to care even more for property than we do ourselves we men are inclined to ease up a little when we have all we need"	"but that would not help for long unless they could abolish the love of it ants seem to care even more for property than we do ourselves we met are inclined to ease up a little when we have all we need"	0,0238	men (become met)

No	Original text	Transcript model	WER	Mismatched Words
7	"the work of deduction is the interpretation of these formulas and therefore strictly speaking is not inferential at all the real inference was accomplished when the universal proposition was arrived at"	"the work of deduction is the interpretation of these formulas and therefore strictly speaking is not inferential at all the real inference was accomplished when the universal proposition was arrived at"	0	nothing
8	"the saloon furnishes material to be saved faster than the settlement or residence or rescuemission work can save it virginia suddenly rose and paced the hall rachel answered sadly and yet with a note of hope in her voice"	"the saloon furnishes material to be saved faster than the settlement or residence or rescuemission work can save it virginia suddenly rose and paced the hall rachel answered sadly and yet with a note of hope in her voice"	0,05	rescue mission (become rescuemission)
9	"but all that i can say will be of little avail my revenge is of no moment to you yet while i allow it to be a vice"	"but all that i can say will be of little avail my revengees of no moment to you yet while i allow it to be a vice"	0,0714	revenge is (become revengees)
10	"it was intensely dark the binnacle lamp was extinguished and the light in the cabin burned too dimly to throw the faintest colour upon the hatchway one thing i quickly noticed that the gale had broken and blew no more than a fresh breeze"	"it was intensely dark the vinnical lamp was extinguished and the light in the cabin burned too dimly to throw the faintest color upon the hatchway one thing i quickly noticed that the gale had broken and blew no more than a fresh breeze"	0,0455	binnacle (become vinnical), colour (become color)

In the examples above, the model was able to transcribe most of the sentences accurately, achieving a WER of zero in some samples. However, some transcription errors also occurred, as outlined in the "Inappropriate Words" column. These errors indicate that the model occasionally made phonetic or similar contextual errors, but overall, the sentence structure and the majority of words were correctly recognized. This kind of qualitative evaluation provides deeper insight into the types of errors that still occur and potential areas for future improvement.

CONCLUSION

This research successfully developed a Speech-to-Text (ASR)-based Narrative Text Reading Assessment Automation System using the Wav2Vec2 model. This system specifically aims to improve the English pronunciation skills of junior high school students. The fine-tuned Wav2Vec2 model demonstrated excellent ASR performance with a final Word Error Rate (WER) of 0.042778 (approximately 4.28%) on validation data, demonstrating high accuracy in speech transcription. The model also demonstrated a significant and consistent reduction in Training Loss and Validation Loss across 24,000 training steps, indicating effective learning and good generalization capabilities.

The "Wav2Speech" application system was also successfully implemented, integrating the ASR model for automatic transcription of user audio. The application features a specialized, relevant accuracy scoring system that ignores "religious words" in its calculations to provide scores that focus more on the narrative. Furthermore, a gamification mechanism, including learning progress tracking and a feature to "unlock" stories based on a minimum score of 80%, was successfully implemented, effectively motivating users to practice. Although the model's WER is very low, accuracy across real user interactions in the app shows variation (e.g., 6% accuracy on a single sample), indicating sensitivity to audio quality and user pronunciation. Overall, the system achieves its goal of providing an interactive and innovative learning tool for English pronunciation.

Based on the research results and conclusions, several suggestions for future development and improvement are presented. First, improving the robustness of the ASR model is a priority. The model needs to be developed to be more robust to variations in audio quality and user accents (e.g., non-native speakers).

This improvement can be achieved through data augmentation or fine-tuning with a wider range of real user speech data. Second, improvements to the scoring system are highly recommended. Implementing fuzzy matching can provide tolerance for minor pronunciation errors, while providing more specific pronunciation feedback (e.g., phonemic identification) would be helpful for users. Third, more comprehensive user testing is needed. User testing studies involving more participants can collect aggregate data from the learning_outcomes table. This data can then be used for in-depth statistical analysis of the app's effectiveness and user error patterns, providing valuable insights for optimization. Finally, developing advanced app features can enrich the user experience. This includes adding a variety of story content, including a reference pronunciation feature (text-to-speech), and developing a more interactive and adaptive practice mode to suit user needs.

REFERENCES

- [1] U. Zuhdi, "Identifikasi Pengaruh Keragaman Kultural Terhadap Kesulitan Dalam Proses Belajar Mengajar Bahasa Inggris Sebagai Bahasa Asing Di Sekolah Dasar," *Jurnal Penelitian Pendidikan Guru Sekolah Dasar*, 2020.
- [2] R. Husni and E. N. Saputri, "Kesulitan siswa SMP dalam pembelajaran bahasa Inggris," *Innovative: Journal Of Social Science Research*, vol. 3, no. 3, pp. 8046–8052, 2023.
- [3] "EF EPI EF English Proficiency Index A Ranking of 116 Countries and Regions by English Skills," 2024. [Online]. Available: www.efset.org/www.ef.com/epi
- [4] J. P. Lubis, N. Z. N. Fitri, and S. C. Ridwan, "Pentingnya menguasai bahasa Inggris dan faktor yang mempengaruhi kemampuan berbahasa Inggris," *Karimah Tauhid*, vol. 3, no. 3, pp. 3599–3605, 2024.
- [5] M. Susini, "Strategi meningkatkan kemampuan berbahasa inggris," *Linguistic Community Services Journal*, vol. 1, no. 2, pp. 37–48, 2020, doi: <https://doi.org/10.22225/licosjournal.1.2.2020.37-48>.
- [6] E. Putri Ananda, "Daya Minat dalam Penggunaan Bahasa Inggris dan Pengaruhnya Terhadap Komunikasi Masyarakat Indonesia," *HYPOTHESIS : Multidisciplinary Journal Of Social Sciences*, vol. 2, no. 01, pp. 172–184, Jun. 2023, doi: 10.62668/hypothesis.v2i01.664.
- [7] A. Shodikin, A. F. Asyifa, J. Z. Hafizd, D. Safitri, and E. Setyawan, "Governing Islamic Marriage: The Role of Tawkil Wali Bil Kitabah in Indonesia's Marriage Registration System," *Khazanah Hukum*, vol. 8, no. 2, pp. 33–46, 2026, doi: <https://doi.org/10.15575/kh.v8i2.54248>.
- [8] D. N. Setiawan, M. Faisal, and M. Imamudin, "Fine-Tuned BART Transfer Learning for Abstractive Summarization of Indonesian YouTube Transcripts with ROUGE Evaluation," 2026. doi: 10.2139/ssrn.6128825.
- [9] I. Jauregi Unanue, J. Parnell, and M. Piccardi, "BERTTune: Fine-Tuning Neural Machine Translation with BERTScore," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 915–924. doi: 10.18653/v1/2021.acl-short.115.
- [10] J. Shor, R. A. Bi, S. Venugopalan, S. Ibara, R. Goldenberg, and E. Rivlin, "Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 1–7. doi: 10.18653/v1/2023.clinicalnlp-1.1.
- [11] A. Basir, "Understanding The Difficulties and Solutions in Speaking English Among Indonesian University Students: Case Study of ABC University," *Journal of Development Research*, vol. 9, no. 1, pp. 174–180, May 2025, doi: 10.28926/jdr.v9i1.435.
- [12] N. Noprival, "Speaking Obstacles on English for Medical Purpose Class at an Indonesian College of Health Sciences," *Jurnal Ilmiah Bina Bahasa*, vol. 15, no. 2, pp. 97–106, Nov. 2022, doi: 10.33557/binabahasa.v15i2.1960.
- [13] N. F. Anggrisia and A. Robah, "Exploring challenges and strategies in English speaking among Indonesian university students: A case study of AKM university," *Englisia: Journal of Language, Education, and Humanities*, vol. 11, no. 1, Oct. 2023, doi: 10.22373/ej.v11i1.19156.
- [14] N. Halim, N. A. Kasim, and D. F. Pratiwi, "Developing Speaking Proficiency in Indonesian EFL Classrooms: A Qualitative Study on Challenges and Solutions," *The Academic: English Language Learning Journal*, vol. 10, no. 1, pp. 9–18, Jun. 2025, doi: 10.52208/aellj.v10i1.1384.
- [15] N. A. Lubis, A. Fadhilah, A. M. Azura, S. S. Rambe, M. A. Syafiq, and D. Santoso, "The Students Ignorance In Practicing English Speaking," *Fonologi: Jurnal Ilmuan Bahasa dan Sastra Inggris*, vol. 2, no. 3, pp. 205–211, Jul. 2024, doi: 10.61132/fonologi.v2i3.903.
- [16] S. W. Ali, I. W. Bay, and F. S. Nur, "Barriers to English-Speaking Proficiency Among Indonesian EFL Learners," *International Journal of Integrative Sciences*, vol. 4, no. 6, pp. 1185–1194, Jun. 2025, doi: 10.55927/ijis.v4i6.327.
- [17] I. K. Mohammad, H. Fatsah, and A. Abid, "English Speaking Difficulties Encountered by Tourism Department Students' in EFL Public Interactions: A qualitative study," *Jambura Journal of English Teaching and Literature*, vol. 3, no. 1, pp. 30–40, May 2022, doi: 10.37905/jetl.v3i1.15122.
- [18] J. Fatina, "Kendala dalam Belajar Bahasa Inggris untuk Peserta Didik Kelas VII Sekolah SMP Labschool FIP UMJ."
- [19] K. Pendidikan and D. Teknologi, *English For Nusantara SMP/MTs KELAS VIII*. 2022. [Online]. Available: <https://buku.kemdikbud.go.id>
- [20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." [Online]. Available: <https://github.com/pytorch/fairseq>
- [21] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2Vec2-Large-960h," <https://huggingface.co/facebook/wav2vec2-large-960h>.

- [22] J. Oruh, S. Viriri, and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.
- [23] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey."
- [24] S. Wang, "Recognition of English speech - Using a deep learning algorithm," *Journal of Intelligent Systems*, vol. 32, no. 1, Jan. 2023, doi: 10.1515/jisys-2022-0236.
- [25] Simone Wills, Yu Bai, Cristian Tejedor-García, and Catia Cucchiari, and Helmer Strik, "Automatic Speech Recognition of Non-Native Child Speech for Language Learning Applications," in *OpenAccess Series in Informatics*, Schloss Dagstuhl- Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Aug. 2023. doi: 10.4230/OASICS.SLATE.2023.11.
- [26] Y. Yahya and A. M. Nur, "Pengaruh Penerapan Media Interaktif Berbasis Multimedia Untuk Pembelajaran Bahasa Inggris Di MTs. NW Ketangga Menggunakan Algoritma Naïve Bayes," *Infotek: Jurnal Informatika dan Teknologi*, vol. 5, no. 1, pp. 209–218, Jan. 2022, doi: 10.29408/jit.v5i1.4841.
- [27] D. W. Wibowo, I. K. Putri, and L. Saputri, "Integration of Augmented Reality and Voice Recognition in Learning English for Children," *Journal of Applied Intelligent System*, vol. 7, no. 2, pp. 103–110, Sep. 2022, doi: 10.33633/jais.v7i2.6119.
- [28] W. Ke, "Study on recognition and classification of English accents using deep learning algorithms," *Journal of Intelligent Systems*, vol. 32, no. 1, Dec. 2023, doi: 10.1515/jisys-2023-0174.
- [29] F. Jiao, J. Song, X. Zhao, P. Zhao, and R. Wang, "A Spoken English Teaching System Based on Speech Recognition and Machine Learning," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 16, no. 14, p. 68, Jul. 2021, doi: 10.3991/ijet.v16i14.24049.
- [30] N. Azis, H. Herwanto, and F. Ramadhani, "Implementasi Speech Recognition Pada Aplikasi E-Prescribing Menggunakan Algoritme Convolutional Neural Network," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 460, Apr. 2021, doi: 10.30865/mib.v5i2.2841.
- [31] A. Alsobhani, H. M. A. Alabbodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," *J. Phys. Conf. Ser.*, vol. 1973, no. 1, p. 012166, Aug. 2021, doi: 10.1088/1742-6596/1973/1/012166.
- [32] M. M. I. Putra, S. Sompie, S. Paturusi, T. Elektro, and U. Sam, "Implementasi Speech Recognition pada Aplikasi Pembelajaran Bahasa Inggris untuk Anak," *Jurnal Teknik Informatika*, vol. 15, no. 4, pp. 247–256, 2020, doi: <https://doi.org/10.35793/jti.v15i4.30426>.
- [33] I. Hadiatma, H. A. Nugroho, and E. Nugroho, "Desain dan Implementasi Speech Recognition Sebagai Media Pembelajaran Pronunciation Bahasa Inggris," *J. Teknol. Elektro*, vol. 9, no. 1, pp. 39–45, 2018, doi: <https://doi.org/10.22441/jte.v9i1.3138>.
- [34] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
- [35] A. Vaswani et al., "Attention Is All You Need," Jun. 2017.
- [36] X. Yue, X. Gao, X. Qian, and H. Li, "Adapting Pre-Trained Self-Supervised Learning Model for Speech Recognition with Light-Weight Adapters," *Electronics (Switzerland)*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010190.
- [37] Ida Novindasari, "LLM Transformer: Kamu Hanya Butuh Attention."
- [38] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." [Online]. Available: <https://github.com/pytorch/fairseq>
- [39] A. Baevski, M. Auli, and A. Conneau, "Wav2vec 2.0: learning the structure of speech from raw audio (2020)," URL: <https://ai.meta.com/blog/wav2vec-20-learningthe-structure-of-speech-from-raw-audio>.
- [40] M. Popović and H. Ney, "Word error rates: Decomposition over POS classes and applications for error analysis," in *Proceedings of the second workshop on statistical machine translation*, 2007, pp. 48–55.
- [41] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates, "An empirical analysis of word error rate and keyword error rate," in *Interspeech*, 2008, pp. 2070–2073.
- [42] A. Ali and S. Renals, "Word Error Rate Estimation for Speech Recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 20–24. doi: 10.18653/v1/P18-2004.
- [43] J. Yang et al., "Effective Fine-tuning Method for Tibetan Low-resource Dialect Speech Recognition," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, Oct. 2023, pp. 960–965. doi: 10.1109/APSIPAASC58517.2023.10317230.
- [44] H. Wang, J. Bin, C. Gou, L. Yang, B. Hou, and M. Qin, "Low-Resource Speech Recognition by Fine-Tuning Whisper with Optuna-LoRA," *Applied Sciences*, vol. 15, no. 24, p. 13090, Dec. 2025, doi: 10.3390/app152413090.
- [45] A. Almahmood, H. Al-Ammal, and F. Albalooshi, "Enhancing Speech-to-Text Transcription Accuracy for the Bahraini Dialect," in *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, IEEE, Nov. 2024, pp. 508–514. doi: 10.1109/3ict64318.2024.10824280.
- [46] Y. Han, H. Chen, J. Du, C.-Q. Kong, S. Xiong, and J. Pan, "Layer-Adaptive Low-Rank Adaptation of Large ASR Model for Low-Resource Multilingual Scenarios," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, Nov. 2024, pp. 696–700. doi: 10.1109/ISCSLP63861.2024.10800407.
- [47] T. Tan et al., "CBA-Whisper: Curriculum Learning-Based AdaLoRA Fine-Tuning on Whisper for Low-Resource Dysarthric Speech Recognition," in *Interspeech 2025*, ISCA: ISCA, Aug. 2025, pp. 3309–3313. doi: 10.21437/Interspeech.2025-1705.

- [48] K. Takahashi, K. Hojo, T. Sakai, Y. Wakabayashi, and N. Kitaoka, "Fine-tuning Parakeet-TDT for Dysarthric Speech Recognition in the Speech Accessibility Project Challenge," in *Interspeech 2025, ISCA: ISCA*, Aug. 2025, pp. 3304–3308. doi: 10.21437/Interspeech.2025-1484.