Jurnal Elektronika dan Telekomunikasi (JET), Vol. 24, No. 1, August 2024, pp. 1-11
Accredited by KEMDIKBUDRISTEK, Decree No: **158/E/KPT/2021**
doi: 10.55981/jet.595

Designing Human-Robot Communication in the Indonesian Language Using the Deep Bidirectional Long Short-Term Memory Algorithm

Suci Dwijayanti [*]**, Ahmad Reinaldi Akbar, Bhakti Yudho Suprapto**

Dept. of Electrical Engineering
Universitas Sriwijaya
Indralaya, 30662
Indonesia

Abstract

Humanoid robots closely resemble humans and engage in various human-like activities while responding to queries from their users, facilitating two-way communication between humans and robots. This bidirectional interaction is enabled through the integration of speech-to-text and text-to-speech systems within the robot. However, research on two-way communication systems for humanoid robots utilizing speech-to-text and text-to-speech technologies has predominantly focused on the English language. This study aims to develop a real-time two-way communication system between humans and a robot, with data collected from ten respondents, including eight males and two females. The sentences used adhere to the standard rules of the Indonesian language. The speech-to-text system employs a deep bidirectional long short-term memory algorithm, coupled with feature extraction via the Mel frequency cepstral coefficients, to convert spoken language into text. Conversely, the text-to-speech system utilizes the Python pyttsx3 module to translate text into spoken responses delivered by the robot. The results indicate that the speech-to-text model achieves a high level of accuracy under quiet-room conditions, with noise levels ranging from 57.5 to 60 dB, boasting an average word error rate (WER) of 24.99% and 25.31% for speakers within and outside the dataset, respectively. In settings with engine noise and crowds, where noise levels range from 62.4 to 86 dB, the measured WER is 36.36% and 36.96% for speakers within and outside the dataset, respectively. This study demonstrates the feasibility of implementing a two-way communication system between humans and a robot, enabling the robot to respond to various vocal inputs effectively.

Keywords: deep learning, humanoid robots, two-way communication, text-to-speech system, speech-to-text system.

I. INTRODUCTION

Science and technology are advancing rapidly, particularly in robotics, as evidenced by the increasing prevalence of robots that assist humans in various tasks. Robots are automated machines that can operate autonomously or with human assistance and are utilized across numerous industries to enhance manufacturing efficiency. In recent years, humanoid robots have garnered considerable interest.

Humanoid robots resemble humans and are capable of performing various human-like activities. They can assist with tasks such as cooking, acting as receptionists, providing guidance, and engaging in direct two-way communication. Common deployment locations for robots include hotels, hospitals, stores, and airports. Typically, robots possess the capability to respond to inquiries from users, thus facilitating two-way communication between humans and robots.

Two-way communication in humanoid robots is enabled by a speech-recognition system integrated within the robot. This system, crucial for information exchange, operates on acoustic signals [1] and falls under the domain of natural language processing (NLP). NLP, a programming technique enabling computers to understand and generate human language, serves as the bridge for communication between humans and machines [2], [3]. Speech recognition entails identifying spoken words by comparing sound pressure levels against existing database templates, necessitating an effective method for converting speech into text. Subsequently, the converted text is transformed into speech as a response from the humanoid robot.

Several studies have investigated voice recognition and its applications in humanoid robots. Alnuaim et al. [4] explored speech recognition for motion commands in humanoid robots using a multilayer perceptron (MLP). However, their approach is limited to issuing motion commands exclusively. Del Duchetto et al. [5] developed a museum robot for guiding visitors, but their research is constrained to interactions in the English language. Furthermore, research has been conducted on robots capable of two-way communication and Korean-to-English translation [6]. In subsequent research, Burewar [7] devised a voice-controlled robot, although the method relied on a probability neural network and has not been tested under real-world conditions. Patil et al. [8] implemented voice control for robots using LabView, while Andreas et al. investigated a self-learning humanoid robot employing NLP [9]. Meanwhile, Bingol and Aydogmus [10] proposed human-robot interaction for industrial robots using Turkish. Deuerlein et al. [11]

[*] Corresponding Author.
Email: sucidwijayanti@ft.unsri.ac.id

Received: October 29, 2023 ; Revised: January 27, 2024
Accepted: January 30, 2024 ; Published: August 31, 2024

developed a software interface that identifies commands through cloud-based speech processing and converts them into machine-readable code.

Previously, studies have explored speech recognition for the Indonesian language. Reference [12] utilized Indonesian to control a robot via Kinect 2.0, with commands spoken by individuals from diverse ethnic backgrounds. However, the commands were simplistic, comprising single words, and only employed Microsoft speech recognition, lacking two-way communication. Reference [13] employed a combination of Mel frequency cepstral coefficient (MFCC) and artificial neural network for commanding a 4-DoF manipulator robot, yet it also utilized single-word commands without two-way communication.

Based on these studies, research on two-way communication between humans and robots remains limited, often relying on conventional modules and methods with restricted accuracy, primarily utilizing simple commands. Thus, improvement is needed by implementing the long short-term memory (LSTM) method. LSTM has been demonstrated to enhance WER when combined with DNN-HMM [14], albeit tested on the TIMIT speech library, containing sentences in the American English dialect. LSTM has also been applied to the Sichuan dialect in combination with HMM [15], exhibiting superior performance compared to DNN. Moreover, [16] showcased the effectiveness of LSTM in recognizing English by integrating audio and visual inputs. Additionally, a study demonstrated LSTM's performance enhancement through regularized adaptation based on an acoustic model trained with connectionist temporal classification (CTC) [17], albeit only applied to Mandarin speech recognition. Furthermore, [18] utilized bidirectional LSTM (BiLSTM) and LSTM for sentiment analysis in Bahasa Indonesia, specifically handling negation words in written language, without implementing a communication method between humans and machines. Another study proposed an open-source tool for automatic speech recognition supporting various algorithms such as multilayer perceptron, LSTM, gated recurrent unit (GRU), and light gated recurrent unit [19], evaluated using English corpora including the DIRHA-English dataset, CHiME, and LibriSpeech.

Based on the studies mentioned earlier, in addition to the utilization of LSTM which is predominantly employed in English and Chinese languages, research on two-way communication between humans and robots remains limited. In response to these limitations, this study introduces the development of a two-way communication system for humanoid robots. The developed system includes speech-to-text conversion for transcribing speech into text and text-to-speech conversion for generating the robot's responses in the Indonesian language. The deep bidirectional LSTM method was employed to construct the speech-to-text system integrated into humanoid robots, with LSTM demonstrating satisfactory performance in speech-to-text conversion [20]. For text-to-speech functionality in humanoid robots, the pyttsx3 module was utilized to convert textual responses into speech, facilitating two-way communication between humanoid robots and humans.

Considering the limitations above, this study aims to develop a two-way communication system for implementation on humanoid robots, utilizing speech-to-text and text-to-speech based on LSTM and focusing on Bahasa Indonesia. The primary contributions of this study are outlined as follows:

1. Development of a speech-to-text and text-to-speech model for enabling two-way communication between robots and humans.
2. Creation of a dataset in the Indonesian language sourced from male and female speakers.
3. Implementation of real-time communication between robots and humans.

The remainder of this paper is structured as follows: Section II delineates the methodologies employed in this study. Section III presents the results and discussion, while Section IV concludes the study.

II. Methods

A. Humanoid Robot Design

In this study, various hardware components are utilized to facilitate the implementation of the humanoid robot, each serving a distinct purpose.

- FIFINE Microphone
- Active Speaker
- Arduino
- Raspberry Pi
- DOT Matrix

The FIFINE microphone is employed to capture sound within a frequency range of 20 Hz to 20 kHz, aligning with the 16 kHz frequency of the sound signal utilized in this study. An active speaker is utilized to produce audio output as the robot's response. The Raspberry Pi 3 Model B and Arduino serve as the core systems controlling the speech process. Additionally, a Dot Matrix 8 × 32 is incorporated to animate mouth movements when the humanoid robot generates output from the speaker. The arrangement of these components is illustrated in Figure 1 (A–C).

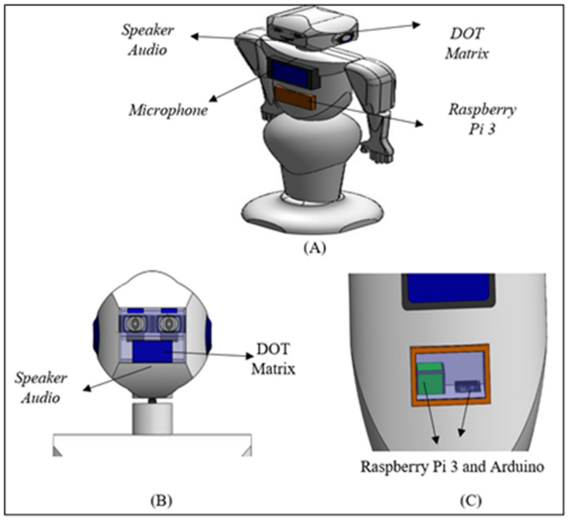

Figure 1. (A-C) Component design of the humanoid robot.

B. Algorithm Design for Two-way Communication System for a Humanoid Robot

1) Algorithm for Training a Two-Way Communication System in Humanoid Robot

In this study, the deep bidirectional LSTM, a fusion of bidirectional recurrent neural network and LSTM, is employed. This approach delves into long-range context dependencies both in the past (t-1) and the future (t+1) for modeling purposes. In the BiLSTM setup, the input is fed into both the forward and backward LSTM layers. Notably, there are no connections between the forward and backward layers in the hidden layer. The output layer consolidates information from both the forward and backward LSTM layers, as depicted in Figure 2.

The speech-to-text process flow utilized in this study is depicted in Figure 3. Initially, voice recording is conducted, followed by preprocessing of the recorded voice, which involves normalization, silence removal, and pre-emphasis. Subsequently, feature extraction is carried out using MFCCs as input for the deep BiLSTM. The output yields probabilities for each label. CTC is then employed to determine the loss of the deep bidirectional LSTM and decode the labeling process using the output of the deep BiLSTM.

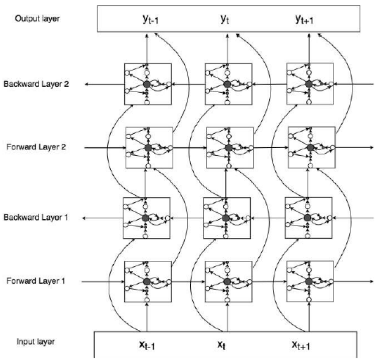

Figure 2. Architecture of deep bidrectional LSTM [21]

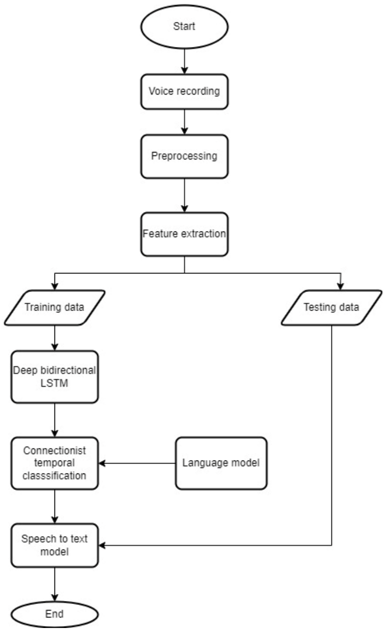

Figure 3. Flowchart of the speech-to-text process.

A language model plays a crucial role in predicting upcoming words during the speech-to-text process. To construct a language model, a text corpus, which is a compilation of electronically stored texts utilized for research purposes, particularly in NLP, is essential [22]. In this study, a text corpus was generated from a sketch engine using a corpus generator and saved in plain text (.txt) format. The corpus comprises over 300,000 words extracted from articles published in the Kompas online newspaper between 2001 and 2002. This corpus was selected due to its inclusion of 5000 articles written in Indonesian, encompassing a broad spectrum of words and phrases, thus rendering it well-suited for this investigation.

Before processing the text corpus into a language model, it must undergo a cleaning process. This involves eliminating sentences containing specific symbols to ensure that the sentences within the text corpus consist solely of alphabetical letters from A to Z and spaces.

2) Communication Algorithm in Humanoid Robots for Providing Responses

The developed algorithmic system is integrated into a humanoid robot to facilitate its engagement in two-way communication with humans, responding to inquiries and greetings. A flowchart outlining the humanoid-robot algorithm for generating responses is presented in Figure 4.

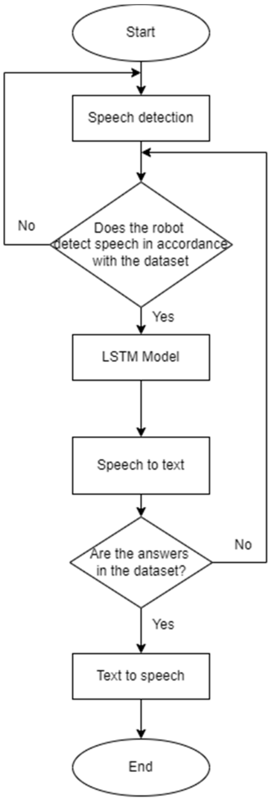
Start
Speech detection
No
Does the robot detect speech in accordance with the dataset
Yes
LSTM Model
Speech to text
Are the answers in the dataset?
No
Yes
Text to speech
End

Figure 4. Flowchart of the humanoid robot algorithm for providing a response.

C. Training and Testing Data

The voice dataset utilized in this study comprises speech recordings obtained from ten speakers who were students in the Electrical Engineering Class of 2018 at Universitas Sriwijaya. This group consisted of eight males and two females. Gender is a significant factor due to the differing fundamental frequencies between males and females; thus, ensuring variation in the data is crucial for achieving an optimal model capable of generalizing new data. Voice recordings were conducted using a FIFINE K669B microphone and Audacity software. To maintain consistency, various configurations were applied in Audacity during the recording process to ensure uniform parameters across all voice samples. These configurations included a sampling rate of 16,000 Hz, mono channel, and 16-bit PCM. To minimize noise interference, recordings and data collection were conducted in a quiet room at the Department of Electrical Engineering, Universitas Sriwijaya.

During the voice recording phase, each speaker was instructed to articulate 17 sentences. The sentences in Table 1 contain greetings and commands intended for communication with the humanoid robot. Each speaker repeated each sentence ten times with varying intonations to introduce data variation during training. Consequently, each recorded voice data point from the speakers was duplicated 25 times per sentence, resulting in a total voice dataset of 42,500 voice samples. This duplication was carried out to augment the training data necessary for deep learning.

The acquired voice data were partitioned into three segments: training (70%), validation (20%), and testing (10%). Subsequent to this division, the voice data underwent preprocessing prior to entering the training phase for the deep bidirectional LSTM. The sentences uttered by the speakers during the data collection process are detailed in Table 1.

D. Processing of Training and Testing Data

The collected training and test data underwent preprocessing, encompassing normalization, silence removal, and pre-emphasis, with a coefficient filter of 0.97. All preprocessing steps were executed utilizing Python libraries, specifically Pydub.

TABLE 1
SENTENCES SPOKEN BY SPEAKERS IN BAHASA INDONESIA

No	Sentence
1	Halo, selamat pagi
2	Halo, selamat siang
3	Halo, selamat sore
4	Hai, apa kabar?
5	Robot, ambilkan barang
6	Robot, bawakan barang
7	Maju
8	Mundur
9	Samping kanan
10	Samping kiri
11	Hadap kanan
12	Hadap kiri
13	Assalamualaikum
14	Halo robot
15	Siapa nama kamu?
16	Tolong berikan ke bu Suci
17	Tolong berikan ke pak Bhakti

Subsequently, the preprocessed audio dataset underwent feature extraction prior to the training process. Feature extraction from the preprocessed audio signals was accomplished using MFCCs due to its superior performance as an extraction method compared to spectrograms as input for the speech-to-text process in LSTM. MFCC feature extraction was conducted using the Librosa library.

The initial step in the feature extraction process for audio signals entails obtaining the log spectrum through short-term Fourier transform (STFT). Initially, the audio signal is segmented into windows of 32 ms length with a window step of 16 ms. Subsequently, the segmented audio signal is transformed from the time domain to the frequency domain utilizing fast Fourier transform (FFT). A frequency bin size of 512 was employed in the procedure to enhance frequency resolution. However, the values displayed in the spectrogram were $\frac{N}{2}+1$, resulting in 257 bins. The FFT outcomes were then converted into a log-power spectrum.

Following the spectrogram generation, the output underwent filtering using a filter bank and was presented on the Mel scale. The Mel filter bank employed 40 filters spanning from 0 to 8 kHz. This process yielded an output in the form of a Mel spectrum, which was then converted back to the time domain. Utilizing a coefficient of 13, a discrete cosine transform (DCT) was utilized to facilitate the conversion to the time domain. Figure 5 illustrates the MFCC feature extraction outcome for the sentence "Halo, selamat pagi." The output derived from the DCT process, referred to as the MFCCs, served as the input for the LSTM training process.

E. Evaluation System

An evaluation was subsequently conducted to ascertain the functionality of the designed two-way communication system between humans and humanoid robots. This evaluation aimed to determine the system's success rate or accuracy in converting speech into text and responding to inquiries and greetings from humans. In the event that the obtained results were suboptimal, a review of the system's algorithm was undertaken.

For speech-to-text testing, the word error rate (WER) was determined as (1),

$$E(h,S') = \frac{1}{z}\sum_{(x,z)\in S} ED(h(x)), \qquad (1)$$

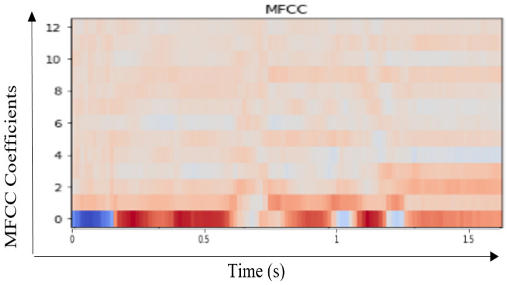

Figure 5. Plot of MFCC for “Halo, selamat pagi.”

where *h* denotes the target speech-to-text from *S'*, *Z* denotes the total target length, and the edit distance (*ED*) represents the changes required to reach the target. WER is a critical performance metric in speech recognition [23].

Subsequently, two-way communication was assessed through repeated experiments to ascertain the robot's capability to respond to provided questions and greetings.

III. Results and Discussion

A. Training Model

The training was performed using the deep bidirectional LSTM algorithm, a component of the recurrent neural network (RNN), for deep speech processing. The architecture of the deep BiLSTM employed in this study is depicted in Figure 6, with SoftMax serving as the activation function.

For the training process, the previously generated dataset needed to be organized in .csv format, featuring three columns with headers: wav_filename, size, and transcript of the audio signals. This organization facilitated the retrieval of audio samples during training.

The trained audio data corresponded to those obtained from the MFCC feature extraction process. The training was conducted using Google Colaboratory, with several parameters considered, including the number of epochs, batch size, learning rate, and hidden layers, as outlined in Table 2.

The training process continued until reaching a specified number of epochs, with each epoch comprising steps that involved iterating through all the data. Upon completing the training process, the trained model was saved in a.pb file format.

The training process was conducted three times, with different epochs and parameters as outlined in Table 2. This was done to observe variations in the obtained loss values. The changes in training and validation losses during the training for 75 epochs are illustrated in Figure 7.

Figure 7 depicts a training loss of 0.1083 and a validation loss of 0.9572 after 75 epochs of training. This process took a duration of 1 hour, 58 minutes, and 41 seconds to complete. The graph in Figure 7 indicates that the validation loss remains relatively high, suggesting that the model did not adequately generalize the validation data.

Additionally, when the number of epochs was increased to 150 (Figure 8), both the training and validation losses decreased to 0.0250 and 0.4638, respectively. Achieving these loss values required a training duration of 4 hours, 38 minutes, and 12 seconds. This outcome suggests that as the number of epochs used in the training process increased, the loss value decreased. However, it is worth noting that the training time also increased.

In the third training session, the batch size was reduced to 16 while maintaining the number of epochs at 150. The outcomes of this session revealed decreased training and validation losses, reaching 0.0178 and

TABLE 2
PARAMETERS FOR THE TRAINING PROCESS

Parameter	*1st Training*	*2nd Training*	*3rd Training*
Hidden Layer	100	100	100
Train Batch Size	32	32	16
Validation Batch Size	32	32	16
Test Batch Size	32	32	16
Learning Rate	10^{-4}	10^{-4}	10^{-4}
Epoch	75	150	150
Early Stopping	No	No	No

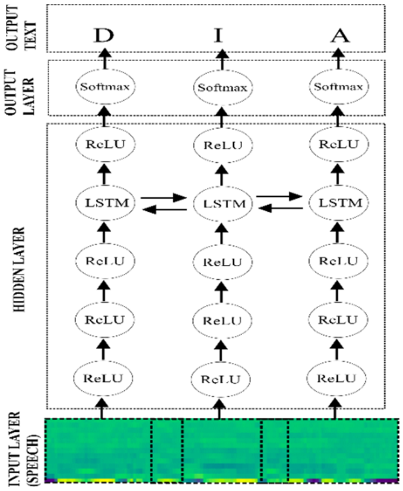

Figure 6. The architecture of Deep Bidirectional LSTM.

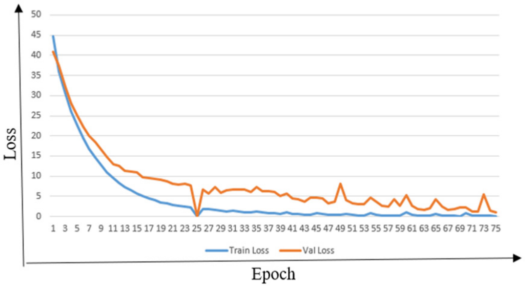

Figure 7. Training losses in the speech-to-text model for 75 epochs.

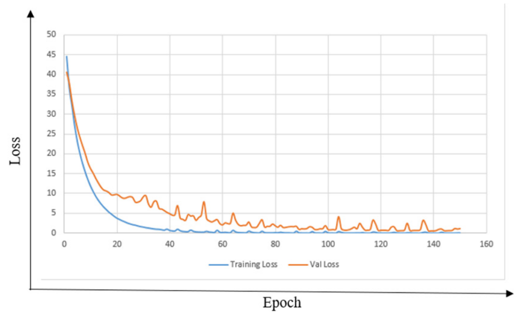

Figure 8. Training losses in the speech-to-text model for a batch size of 32 and 150 epochs.

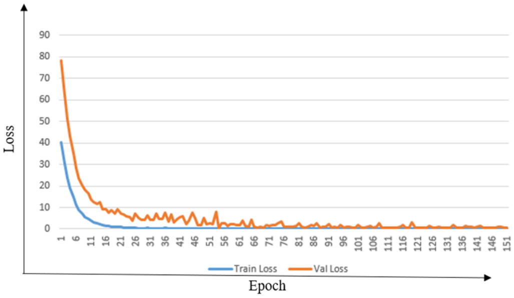

Figure 9. Training losses in the speech-to-text model for a batch size of 16 and 150 epochs.

0.4385, respectively, with a training duration of 5 hours, 59 minutes, and 3 seconds. The results of this session are illustrated in Figure 9, demonstrating that reducing the batch size can potentially lower the loss value during training and validation. However, this reduction is not statistically significant and necessitates a longer training time. Therefore, testing was conducted using the training results with a batch size of 32 and 150 epochs.

B. Model Evaluation

The testing process aimed to ascertain whether the developed model effectively converts speech into text. The test model was derived from the preceding training process, specifically utilizing a previously created test dataset comprising 10% of the available voice data. The accuracy of the model was evaluated using the WER method, which calculates the number of word insertions, substitutions, and deletions detected by the model. A smaller WER value indicates that the model has good accuracy.

The model underwent testing on each voice sample per sentence to evaluate its accuracy. Table 3 presents the WER values for a voice sample from a male speaker with the file name "aldi" per tested sentence, utilizing the model acquired from the training process.

The test results for one voice sample per sentence, as listed in Table 3, were reasonable, yielding an average WER of 11.76% across all sentences. Sentences that were easier to convert included "halo selamat pagi" due to the higher number of voice samples for the words "halo selamat," and the word "pagi" was relatively easier to recognize using the speech-to-text model. However, recognizing the female voice sample with the speech-to-text model during testing posed challenges due to the smaller number of female voice samples in the dataset during the training process.

In addition to utilizing the proposed method, this study also compares it with [24]. As depicted in Table 3, the proposed method proves suitable for the Indonesian language. While most sentences with the English model are not well-recognized, only certain words such as 'halo,' 'tolong,' 'kanan,' and 'bawakan,' can be identified.

As shown in Tables 3 and 4, the results exhibited good performance in terms of WER, compared to the English corpus model using MLP, LSTM, GRU, and Li-GRU. The DIRHA dataset [25] exhibited a higher WER than other datasets for all models. Compared to the proposed model using the Indonesian language, it outperformed the English model, with a WER of 11.76%. The DIRHA dataset shared similar characteristics with our dataset, as recorded in various rooms such as living rooms, kitchens, and corridors, with the subject sitting in front of a display in the recording room. Meanwhile, CHiME [26] had a noisy environment from buses, cafes, pedestrian areas, and street junctions, while LibriSpeech [27] had longer data considering the experiments used a 100 h dataset, compared to our dataset, which was 11 hours.

C. Testing of Two-Way Communication System with Real-Time Speech-to-Text and Text-to-Speech

In this study, the text-to-speech system utilized the Python module pyttsx3, an open-source library. This module facilitates the conversion of text to speech by

TABLE 4
WER (%) OF ENGLISH DATASET USING PYTORCH-KALDI TOOLKIT [19]

Algorithm	DIRHA	CHiME	LibriSpeech
MLP	26.1	18.7	6.5
LSTM	24.8	15.5	6.4
GRU	24.8	15.2	6.3
Li-GRU	23.9	14.6	6.2

TABLE 3
WER OF THE SENTENCES SPOKEN BY THE SPEAKER

No	**Target sentence**	**Detected sentence using proposed method**	**% WER**	**Detected sentence using** [24]	**%WER**
1	halo selamat pagi	halo selamat pagi	0%	Hallo salam mat paggy	100%
2	halo selamat siang	halo selamat siang	0%	Hallow is lamada	100%
3	halo selamat sore	halo selamat sore	0%	Hallow siamotory	100%
4	hai apa kabar	hai apa kabar	0%	Hi a **pakabar**	100%
5	robot ambilkan barang	robot ambil kan barang	33%	**Robot** a bill can barrow	66.67%
6	robot bawakan barang	robot bawakan barang	0%	Robert **bawakan** byrang	66.67%
7	Maju	Maju	0%	Madu	100%
8	Mundur	Mundur	0%	Monder	100%
9	samping kanan	samping kanan	0%	Somping gonnen	100%
10	samping kiri	samping kiri	0%	Fom the guite	100%
11.	hadap kanan	ada kanan	50%	How do o **kanan**	50%
12.	hadap kiri	ada kiri	50%	Id of futty	100%
13.	Assalamualaikum	assalamualaikum	0%	Asala molu coom	100%
14.	halo robot	halo robot	0%	**Halo** robat	50%
15.	siapa nama kamu	siapa namamu	67%	Sap an ma moor	100%
16.	tolong berikan ke bu suci	tolong berikan ke bu suci	0%	Salong briken ka **bu** serchi	80%
17.	tolong berikan ke pak bakti	tolong berikan ke pak bakti	0%	**Tolong** briken ke paback tea	60%

inputting the desired text. An advantageous feature of the pyttsx3 module is its ability to customize the voice to the desired language. For this study, the language utilized was Bahasa Indonesia, with language packages provided by Microsoft.

The two-way communication system, created by integrating speech-to-text and text-to-speech functionalities, underwent real-time testing. The testing was conducted in a classroom located in Building I of the Faculty of Engineering at the Palembang campus of Universitas Sriwijaya. Three main parameters were considered during testing: a quiet room, a room with machine noise, and a room with babble noise. The robot designed to provide system responses can be observed in Figure 10.

1) Test in a Quiet Room

The initial test was conducted in a quiet room, employing the parameters detailed in Table 5. Furthermore, testing in a quiet room was divided into two stages: evaluating voice data within and outside the dataset. The audio signal for the quiet room is depicted in Figure 11.

TABLE 5
PARAMETERS FOR TESTING A TWO-WAY COMMUNICATION SYSTEM IN REAL-TIME IN A QUIET ENVIRONMENT

Room condition	Quiet, noise level 57.6–60.2 dB
Microphone input volume	60%
Room dimension	12 × 6.6 × 2.8 m
Microphone distance	10 cm

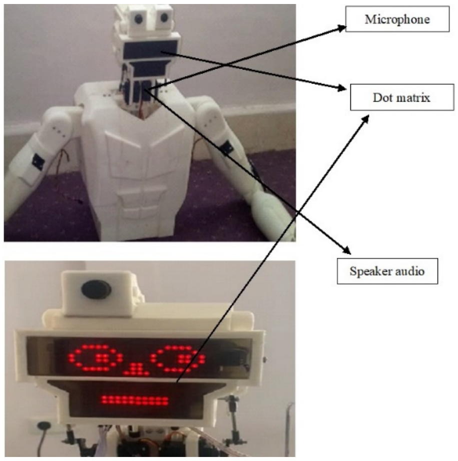

Figure 10. The designed robot.

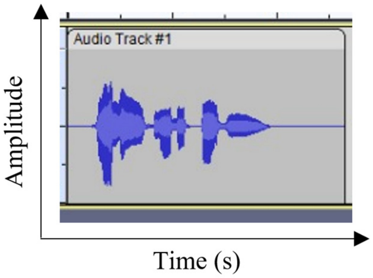

Figure 11. Sample of the audio signal in a quiet room.

The test results for several voice data samples within the dataset in a quiet room are presented in Table 6. These results indicate that the created model effectively converts speech to text, with the average WER value obtained from all voice samples in the dataset in a quiet room being 24.99%.

The WER value was derived from 850 tests, where each speaker pronounced each sentence five times. This illustrates that the two-way communication system utilizing the developed model can recognize spoken voices in a quiet room with a volume level ranging between 58 and 98 dB. However, the WER value for sample 3 exhibited the highest error rate, reaching 50%, attributed to the noise level reaching 98.7 dB, exceeding a normal conversational volume [28]. Overall, the two-way communication system effectively recognized speech-to-text, and the robot responded in the form of text-to-speech.

The test results for the voice data samples outside the dataset in a quiet room are listed in Table 7. These results demonstrate that the system can convert speech to text to enable two-way communication, even if the respondents' data are not included in the dataset. The average WER value obtained from real-time testing for all voice samples outside the dataset under quiet room conditions is 25.31%. The WER value was obtained from 510 tests, where each speaker uttered each sentence five times.

The real-time testing results, both for voice data samples inside and outside the dataset, indicate that the two-way communication system was effective. However, recognizing the female speech sample with the speech-to-text model posed challenges due to the fewer female voice samples present in the dataset during the training process.

TABLE 6
TEN SAMPLE AUDIO SIGNALS IN THE DATASET UNDER QUIET ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	Robot response	% WER
1	Sample 1	74.5–90.4	Male	Halo selamat pagi	Halo selamat pagi	Selamat pagi	0%
2	Sample 2	65.9–93.5	Male	Halo selamat siang	Halo selamat siang	Selamat siang	0%
3	Sample 3	59.3–98.7	Male	Hadap kanan	Adap kanan	-	50%
4	Sample 4	57.4–89.8	Male	Halo selamat sore	Halo selamat sore	Selamat sore	0%
5	Sample 5	58.6–92.3	Male	Hai apa kabar	Hai apa kabab	-	33%
6	Sample 6	59.6–91.3	Male	Siapa nama kamu	Siapa nama kamu	Saya roboto, robot asisten lab LKR	0%
7	Sample 7	62.8–96.3	Male	Halo robot	Halo robot	Halo juga	0%
8	Sample 8	58.2–94.1	Male	Robot bawakan barang	Robot awakan barang	-	33%
9	Sample 9	59.3–98.4	Female	Mundur	Mundur	Mundur	0%
10	Sample 10	65.4–87.6	Female	Maju	Maju	Maju	0%

TABLE 7
SIX SAMPLE AUDIO SIGNALS OUT OF THE DATASET UNDER QUIET ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	Robot response	% WER
1	Sample 11	54.6–82.2	Male	robot ambilkan barang	robot bambilkan barang	-	33%
2	Sample 12	58.9–90.2	Male	samping kanan	samping kanan	samping kanan	0%
3	Sample 13	55.3–86.5	Male	samping kiri	samping kiri	samping kiri	0%
4	Sample 14	63.6–91.6	Female	hadap kiri	hadap kiri	hadap kiri	0%
5	Sample 15	57.3–88.8	Female	assalamualaikum	Hassalamualaikum	-	100%
6	Sample 16	59.5–92.3	Female	tolong berikan ke bu suci	tolong berikan ke pak bakti	siap laksanakan	0%

TABLE 8
TEST PARAMETERS FOR A REAL-TIME TWO-WAY COMMUNICATION SYSTEM UNDER MACHINE NOISE CONDITIONS

Room condition	A bit noisy with the sound of a fan and noise level of 58.6–64.4 dB
Microphone input volume	60%
Room dimension	12 m × 6.6 m × 2.8 m
Microphone distance	10 cm

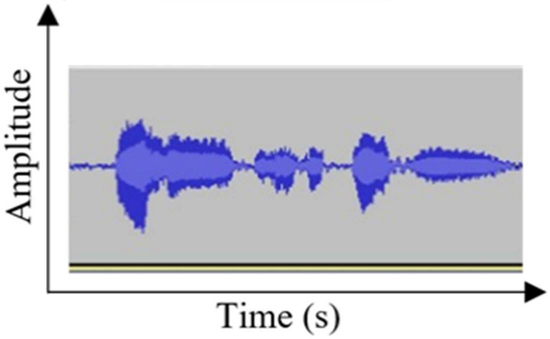

Figure 12. Audio signal for room conditions with machine noise.

2) Testing with Machine Noise

The second testing process was conducted in a noisy room due to the sound of a fan, utilizing the parameters listed in Table 8. The audio waveform for a room with machine noise is depicted in Figure 12.

Testing under noisy room conditions was divided into two stages: evaluating voice data within and outside the dataset. The test results for the voice data samples within the dataset under machine noise from the fan are listed in Table 9.

The results presented in Table 9 indicate that the proposed model can effectively convert speech to text. The average WER value obtained from all the voice data within the dataset, comprising 850 voice samples under fan machine noise, was 29.40%. As observed in Table 9, the robot is capable of responding to speech spoken by the speaker. For instance, when the speaker says 'Halo, selamat pagi,' the robot responds with 'Selamat pagi.' However, the robot is unable to provide a response if the WER is not 0%. This limitation exists because the robot is programmed to respond only to the detected sentence that it has learned.

The outcome is less favorable compared to tests conducted in a quiet room (without background noise), suggesting a significant impact of room noise on the system's speech-to-text conversion accuracy. In particular, samples 1 and 2 yielded a 33% WER, while sample 6 reached 50%, indicating higher error rates than those observed in a quiet environment.

Table 10 outlines the test results for voice data samples subjected to machine noise from outside the dataset. The overall average WER obtained from real-time testing in a quiet room stands at 29.77%. This figure

TABLE 9
TEN SAMPLE AUDIO SIGNALS IN THE DATASET UNDER MACHINE NOISE ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	Robot response	% WER
1	Sample 1	74.5–90.4	Male	halo selamat pagi	halo selamat pag	-	33%
2	Sample 2	65.9–93.5	Male	halo selamat siang	halo selamat siapag	-	33%
3	Sample 3	59.3–98.7	Male	hadap kanan	hadap kanan	-	50%
4	Sample 4	57.4–89.8	Male	halo selamat sore	halo selamat sore	selamat sore	0%
5	Sample 5	58.6–92.3	Male	hai apa kabar	hai pa kabar	-	33%
6	Sample 6	59.6–91.3	Male	siapa nama kamu	siapa na mu	-	50%
7	Sample 7	62.8–96.3	Male	halo robot	halo robot	halo juga	0%
8	Sample 8	58.2–94.1	Male	robot bawakan barang	robot bawakan barang	baik, saya akan membawa barang	0%
9	Sample 9	59.3–98.4	Female	mundur	mundur	Mundur	0%
10	Sample 10	65.4–87.6	Female	maju	maju	Maju	0%

TABLE 10
SIX SAMPLE AUDIO SIGNALS OUT OF THE DATASET UNDER MACHINE NOISE ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	Robot response	% WER
1	Sample 11	54.6-82.2	Male	robot ambilkan barang	robot ambilkan barang	baik, saya akan mengambil barang	0%
2	Sample 12	58.9-90.2	Male	samping kanan	hsamping kam	-	100%
3	Sample 13	55.3-86.5	Male	samping kiri	samping kiri	samping kiri	0%
4	Sample 14	63.6-91.6	Female	hadap kiri	hadap kiri	hadap kiri	0%
5	Sample 15	57.3-88.8	Female	assalamualaikum	hassalamualaikum	-	100%
6	Sample 16	59.5-92.3	Female	tolong berikan ke bu suci	htalo elrikat ke bu suci	-	40%

TABLE 11
TESTING PARAMETERS FOR REAL-TIME TWO-WAY COMMUNICATION SYSTEM UNDER MACHINE AND BABBLE NOISE CONDITIONS

Room condition	Noisy with the sound of a fan, people talking, and a noise level of 76–89 db
Microphone input volume	60%
Room dimension	12 m × 6.6 m × 2.8 m
Microphone distance	10 cm

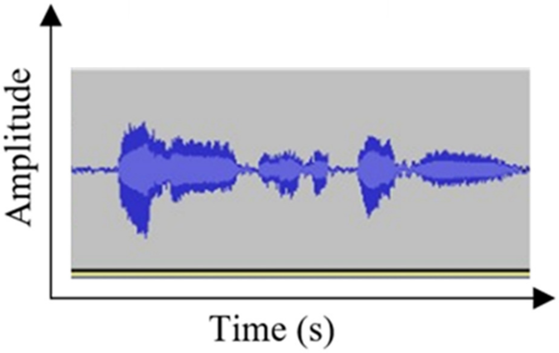

Figure 13. Audio signal for room conditions with machine and babble noise.

is derived from 510 tests, where each speaker vocalized each sentence five times.

The findings in Table 10 affirm the capability of the system to convert speech into text, facilitating two-way communication even when utilizing voice data beyond the dataset. This underscores the system's robustness in facilitating two-way interaction between robots and humans.

3) *Testing in a Room with Machine Noise and Babble Noise*

Subsequently, the third test was conducted in a noisy room with babble noise (conversations accompanied by machine noise), adhering to the parameters listed in Table 11.

The audio signal shape for the condition where both machine and crowd noise are present in the room is shown in Figure 13.

Similar to the two previous tests, the testing conducted under room noise with crowd noise and machine sounds comprised two stages: evaluating the voice data within and outside the dataset. The test results for voice data within the dataset under room noise conditions with crowd noise and machine sounds are listed in Table 12.

The results presented in Table 12 indicate that the developed model can effectively convert speech into text despite the higher noise level compared to the second test**.** The average WER obtained from the voice data included in the dataset is 36.36%. Notably, the highest WER value observed for samples 5 and 7 is 50%, suggesting that volume significantly impacts speech-to-text conversion. It is noteworthy that certain words, such as "robot" and "hai," could not be accurately converted.

Furthermore, the test results for voice data samples outside the dataset in the presence of conversational and machine noise are listed in Table 13. The average WER value obtained from real-time testing under noisy room conditions is 36.96%. This WER value was derived from 510 tests, where each speaker uttered each sentence five times.

The test results provided in Table 13 highlight a notable failure of the speech recognition system to convert speech to text accurately for sample 5, which is consistent with the outcomes observed in the first (quiet room) and second (machine noise) experiments. These findings suggest that the model encountered difficulty in recognizing the word "assalamualaikum" compared to other words. Additionally, this word is infrequently represented in the language model of the text corpus, appearing only three times. Consequently, the

TABLE 12
TEN SAMPLE AUDIO SIGNALS IN THE DATASET UNDER MACHINE AND BABBLE NOISE ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	System response	%WER
1	Sample 1	74.5–90.4	Male	halo selamat pagi	halo selamatgi	-	67%
2	Sample 2	65.9–93.5	Male	halo selamat siang	halo selamat siapang	-	33%
3	Sample 3	59.3–98.7	Male	hadap kanan	hadap kanan	hadap kanan	0%
4	Sample 4	57.4–89.8	Male	halo selamat sore	halo selamat sore	selamat sore	0%
5	Sample 5	58.6–92.3	Male	hai apa kabar	hadap kanabar	-	50%
6	Sample 6	59.6–91.3	Male	siapa nama kamu	siapa nama kamu	saya roboto, robot asistem lab LKR	0%
7	Sample 7	62.8–96.3	Male	halo robot	halo selaot	-	50%
8	Sample 8	58.2–94.1	Male	robot bawakan barang	robot awakan barang	-	33%
9	Sample 9	59.3–98.4	Female	mundur	mundur	Mundur	0%
10	Sample 10	65.4–87.6	Female	maju	maju	Maju	0%

TABLE 13
SIX SAMPLE AUDIO SIGNALS FROM THE DATASET UNDER MACHINE AND BABBLE NOISE ROOM CONDITIONS

No	Sample	Noise level (dB)	Gender	Target sentence	Detected sentence	System response	%WER
1	Sample 11	54.6–82.2	Male	robot ambilkan barang	robot ambilkan barang	baik, saya akan mengambil barang	0%
2	Sample 12	58.9–90.2	Male	samping kanan	samping ka	-	50%
3	Sample 13	55.3–86.5	Male	samping kiri	sampng kiri	-	50%
4	Sample 14	63.6–91.6	Female	hadap kiri	hadap kan	-	50%
5	Sample 15	57.3–88.8	Female	assalamualaikum	assalamialaikuc	-	100%
6	Sample 16	59.5–92.3	Female	tolong berikan ke bu suci	tolong berikan ke bu suci	siap laksanakan	0%

recognition system struggles to accurately interpret it due to insufficient exposure. Nonetheless, despite these challenges, the two-way communication system effectively functioned under noisy conditions, as evidenced by the WER for samples 11 and 16 being 0%, allowing the robot to provide real-time responses.

IV. Conclusion

A successful two-way communication system was established between humans and robots, enabling the humanoid robot to respond appropriately to specific sound inputs. The LSTM method employed for converting speech to text in Bahasa Indonesia demonstrated accuracy in real-time tests conducted under three distinct conditions. The speech-to-text model performed well in a quiet room with noise levels ranging between 57.5 and 60.2 dB, achieving a commendable average WER of 24.99% and 25.31% for the speaker inside and outside the dataset, respectively. This represents an improvement compared to the performance of the models in conditions involving machine noise and crowds, where noise levels ranged between 62.4 and 86.8 dB, resulting in a WER of 36.36% and 36.96% for the speaker inside and outside the dataset, respectively. Future endeavors in this field could explore incorporating various types of noise into the dataset. Additionally, expanding the number of sentences within the dataset and enhancing the robot's response capabilities to accommodate instances where the WER is less than 50% are avenues for further improvement.

Declarations

Conflict of Interest

The authors have declared that no competing interests exist.

CRediT Authorship Contribution

Suci Dwijayanti: Conceptualization, Methodology, Writing-Original draft, Funding Acquisition, Supervision; Ahmad Reinaldi Akbar: Data curation, Investigation, Software; Bhakti Yudho Suprapto: Writing-Reviewing and Editing, Validation.

Funding

This study was funded by the Directorate of Research, Technology and Communication Service, Directorate General of Higher Education, Research and Technology (Research Contract Number: 164/E5/PG.02.00). PL/2023

Data Availability

The data that support the findings of this study can be acquired from the corresponding author upon request.

References

[1] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimed. Tools Appl.*, vol. 80, pp. 9411–9457, 2021, doi: 10.1007/s11042-020-10073-7.

[2] K. R. Chowdhary, *Fundamentals of artificial intelligence*. Springer India, 2020. doi: 10.1007/978-81-322-3972-7.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[4] A. A. Alnuaim *et al.*, “Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier,” *J. Healthc. Eng.*, vol. 2022, Art. no. 6005446, 2022, doi: 10.1155/2022/6005446.

[5] F. Del Duchetto, P. Baxter, and M. Hanheide, “Lindsey the tour guide robot - usage patterns in a museum long-term deployment,” in *Proc. 2019 28th IEEE Int. Conf. Robot Hum. Interact. Commun.*, 2019, doi: 10.1109/RO-MAN46459.2019.8956329.

[6] E. J. Hwang, B. A. MacDonald, and H. S. Ahn, “End-to-end dialogue system with multi languages for hospital receptionist robot,” in *Proc. 2019 16th Int. Conf. Ubiquitous Robot*, 2019, pp. 278–283, doi: 10.1109/URAI.2019.8768694.

[7] S. L. Burewar, “Voice controlled robotic system by using FFT,” in *Proc. 2018 4th Int. Conf. Converg. Technol.*, 2018, pp. 2018–2021, doi: 10.1109/I2CT42659.2018.9058098.

[8] S. Patil, A. Abhigna, Arpitha, Deepthi, and Priyanka, “Voice controlled robot using labview,” in *Proc. 2018 Int. Conf. Des. Innov. 3Cs Comput. Commun. Control,* 2018, pp. 80–83, doi: 10.1109/ICDI3C.2018.00025.

[9] V. Andreas, A. A. S. Gunawan, and W. Budiharto, “Anita: intelligent humanoid robot with self-learning capability using indonesian language,” in *Proc. 2019 4th Asia-Pacific Conf. Intell. Robot Syst.,* 2019, pp. 144–147, doi: 10.1109/ACIRS.2019.8935964.

[10] M. C. Bingol and O. Aydogmus, “Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot,” *Eng. Appl. Artif. Intell.*, vol. 95, Art. no. 103903, 2020, doi: 10.1016/j.engappai.2020.103903.

[11] C. Deuerlein, M. Langer, J. Seßner, P. Heß, and J. Franke, “Human-robot-interaction using cloud-based speech recognition systems,” *Procedia CIRP*, vol. 97, 2021, pp. 130–135, doi: 10.1016/j.procir.2020.05.214.

[12] M. H. Tambunan, Martin, H. Fakhruroja, Riyanto, and C. Machbub, “Indonesian speech recognition grammar using kinect 2.0 for controlling humanoid robot,” in *Proc. 2018 Int. Conf. Signals Syst.*, 2018, pp. 59–63, doi: 10.1109/ICSIGSYS.2018.8373568.

[13] R. Rendyansyah, A. P. P. Prasetyo, and S. Sembiring, “Voice command recognition for movement control of a 4-DoF robot arm,” *ELKHA Jurnal Tek. Elektro*, vol. 14, no. 2, pp. 118–124, 2022, doi: 10.26418/elkha.v14i2.57556.

[14] Z. He, “Improving LSTM based acoustic model with dropout method,” in *Proc. 2019 Int. Conf. Artif. Intell. Adv. Manuf.,* 2019, pp. 27–30, doi: 10.1109/AIAM48774.2019.00012.

[15] W. Ying, L. Zhang, and H. Deng, “Sichuan dialect speech recognition with deep LSTM network,” *Front. Comput. Sci.*, vol. 14, pp. 378–387, 2020, doi: 10.1007/s11704-018-8030-z.

[16] R. Shashidhar, S. Patilkulkarni, and S. B. Puneeth, “Combining audio and visual speech recognition using LSTM and deep convolutional neural network,” *Int. J. Inf. Technol.*, vol. 14, pp. 3425–3436, 2022, doi: 10.1007/s41870-022-00907-y.

[17] J. Yi, H. Ni, Z. Wen, B. Liu, and J. Tao, “CTC regularized model adaptation for improving LSTM RNN based multi-accent mandarin speech recognition,” in *Proc. 2016 10th Int. Symp. Chinese Spok. Lang. Process.,* 2016, doi: 10.1109/ISCSLP.2016.7918420.

[18] R. W. Pratiwi, Y. Sari, and Y. Suyanto, “Attention-based BiLSTM for negation handling in sentimen analysis,” *Indonesian Jurnal Comput. Cybern. Syst.*, vol. 14, no. 4, pp. 397–406, 2020, doi: 10.22146/ijccs.60733.

[19] M. Ravanelli, T. Parcollet, and Y. Bengio, “The Pytorch-Kaldi speech recognition toolkit,” in Proc. *IEEE Int. Conf. Acoust. Speech Signal Process.,* 2019, pp. 6465–6469, doi: 10.1109/ICASSP.2019.8683713.

[20] S. Dwijayanti, M. A. Tami, and B. Y. Suprapto, “Speech-to-text conversion in Indonesian language using a deep bidirectional long short-term memory algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, pp. 225–230, 2021, doi: 10.14569/IJACSA.2021.0120327.

[21] A. Graves, N. Jaitly, and A. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM," in Proc. *2013 IEEE Work. Autom. Speech Recognit. Underst.*, 2013, pp. 273–278. doi: 10.1109/ASRU.2013.6707742.

[22] K. Kurniawan. “Indonesian NLP resources.” Accessed: Oct 15, 2021. [Online]. Available: https://github.com/kmkurn/id-nlp-resource

[23] Y. Y. Wang, A. Acero, and C. Chelba, “Is word error rate a good indicator for spoken language understanding accuracy,” in *Proc.*

2003 IEEE Work. Autom. Speech Recognit. Understanding, 2003, pp. 577–582, doi: 10.1109/ASRU.2003.1318504.

[24] J. Grosman. “HuggingSound: a toolkit for speech-related tasks based on hugging face’s tools.” Accessed: Dec. 15, 2023. [Online]. Availble: https://github.com/jonatasgrosman/huggingsound

[25] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmuller, and P. Maragos, “The DIRHA simulated corpus,” in *Proc. 9th Int. Conf. Lang. Resour. Eval.,* 2014, pp. 2629–2634.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge : dataset, task, and baselines," in *Proc. 2015 IEEE Workshop Automatic Speech Recognition and Understanding*, 2015, pp. 504–511, doi: 10.1109/ASRU.2015.7404837.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[28] Debbie Clason, “Understanding the degrees of hearing loss.” Accessed: Apr. 7, 2023. [Online]. Available: https://www.healthyhearing.com/report/41775-Degrees-of-hearing-loss