Implementation of MFCC Features Extraction with Evaluation of Recurrent Neural Networks and Bidirectional LSTM for Speech to Text Transcription: A Case Study on the Lampung Language Dialect Api

M Ramadhan[1], Akmal Junaidi[2*], Aristoteles[3*], Favorisen R. Lumbanraja[4] and Ahmad Faisol[5]

[1,2,3,4,5] Faculty of Mathematics and Natural Science, Department of Computer Science University of Lampung, Bandar Lampung, Indonesia
ramadaninspira@gmail.com
favorisen.lumbanraja@fmipa.unila.ac.id
ahmadfaisol@fmipa.unila.ac.id
akmal.junaidi@fmipa.unila.ac.id
aristoteles.1981@fmipa.unila.ac.id

Abstract. Lampung language is a cultural heritage of Lampung province which is currently sought to be preserved and maintained its existence amid the declining trend in the use of Lampung language in the Lampung community itself. This research aims to encourage the need for preservation and widespread and inclusive use of Lampung language through speech-to-text conversion technology by utilizing MFCC feature extraction as well as RNN and BiLSTM models and evaluating the performance of acoustic models in text transcription from voice signals, using Word Error Rate (WER) as the main metric. The dataset consisted of regional language-based transcriptions, which were processed through the RNN and BiLSTM models to obtain text output from Lampung language voice conversion using dialect Api. This study used 800 audio sentences in wav format from 4 respondents of Lampung ethnic origin. During the prediction process, the accuracy of the RNN model produced a WER value of 0.5857 and BiLSTM showed a WER of 0.5589. The comparison of these two types of machine learning models shows that BiLSTM produces slightly better model results. Furthermore, the acoustic model was also implemented with 98.53% accuracy on the validation data after 20 epochs, indicating the model learned the data well. Despite the high accuracy, the main challenge faced in these models is generating optimal readable transcriptions, which is due to decoding and tokenization issues.

Keywords: MFCC, RNN, BiLSTM, Speech to Text, Lampung Language

S. Hadi et al. (eds.), *Proceedings of the 5th International Conference on Applied Sciences, Mathematics, and Informatics (ICASMI 2024)*, Advances in Physics Research 13,
https://doi.org/10.2991/978-94-6463-730-4_13

1 Introduction

The number of Lampung speakers is decreasing every year, so it is estimated that by the end of 2023, there will only be 6,250 people who still use Lampung language compared to the total population of Lampung today, so the number is very small [1]. Currently, there are seven regional languages in Lampung, including Lampung, Balinese, Javanese, Sundanese, Bugis, Ogan, and Basemah/Semende. Other regional languages besides Lampung developed through the process of transmigration of people to the Lampung region. Another thing is reinforced by statistics on the percentage of the total population in Lampung Province, currently the majority of non-Lampung tribes with the number of Javanese tribes in the first rank with 4,856,924 people and followed by other ethnicities [2]. Based on the composition data, Lampung language users may continue to decline. In addition to the lack of use, is because the number of users is also decreasing. Thus, Lampung language preservation efforts are a must to encourage the use of Lampung language to be more widespread and inclusive. The use of Lampung language tends to decline in the general community although some are still maintained in the family environment. In the study [3], it was found that 72% of respondents used Lampung language in the family environment and only 45% of respondents used Lampung language in the community environment. Many efforts have been made by the Government in 2024 to revitalize regional languages in 38 provinces with a target of revitalizing 92 regional languages. Efforts and strategies in this case are to cultivate speaking local languages in everyday life in the community and conduct several regional language teacher trainings to be able to teach local language learning in an innovative, creative and fun way.

Technological and information innovation is needed in this preservation effort. Natural Language Processing (NLP) is one method that can be used in language processing through machine learning techniques in it. The use of NLP is based on the application of *machine learning* and *deep learning* by training and analyzing text or audio datasets to solve real-life problems [4]. In the world of computing, text and audio signals are symbols that can be a medium of transaction in the machine learning process. In this case, voice to text (STT) conversion is an option in this research by utilizing the use of respondents' Lampung language voice recordings as the primary dataset. Although voice to text (STT) transformation offers complexity in voice identification as a *signal*. However, automatic speech recognition is currently a popular topic of analysis with several related studies including the use of RNN methods on voice to text conversion in identifying the probabilities of Bengali characters [5], research on Indonesian by applying the *Deep Bidirectional LSTM* algorithm by comparing two feature extraction techniques *spectogram* and *MFCC* [6].

Based on observation and analysis of several research publications related to voice to text (STT) conversion in NLP using *machine learning* or *deep learning* in several language objects, STT research on Lampung language objects has never been done. In this study, the researcher proposes the use of Lampung language audio with the Api dialect. There are two main dialects used in Lampung Language, i.e. Api dialect and Nyo dialect [7]. The distribution of dialect usage spans differently in each region. In

addition, based on research in [6] *using Deep Bidirectional LSTM*, the researcher proposes to compare the use of the deep learning method *Bidirectional LSTM* with RNN to find a comparison of the results of the two models that are characterized by information flow and data processing. The results of the learning model will be evaluated based on the *Word Error Rate.*

2 Materials and Methods

2.1 Data Collection

The data used in this study are audio recording data of Lampung language with the dialect of Api from 4 respondents consisting of 2 men and 2 women who come from areas with most Lampung-speaking people with the dialect of Api. The sound was recorded using with a sampling frequency of 16 kHz. Each respondent uttered 20 sentences in Lampung language consisting of subject, object, predicate and adverb which are shown in Table 1 and have been validated by linguists from the Lampung native community. Each sentence is spoken 10 times, so the total number of sentences recorded from all respondents is 800 sentences. The data is manually divided into 70% for training, 20% for validation, and 10% for testing, which is respectively 560 audios for training, 160 audios for validation and 80 for testing.

Table 1. Lampung language sentences spoken by the speaker

No.	Lampung Sentence	English Translation
1	Ana ngebeli kupi di warung	Ana bought coffee at the stall
2	Daing main bal di lapangan	Brother plays soccer on the field
3	Lita mupoh pakaian di duwai	Auntie washes clothes in the river
4	Buyah ngunut iwa di duwai makai jaring	Father fishes in the river with a net
5	Sikam ngebaca buku di ruang tamu	We read books in the living room
6	Dede nanom sayuran di huma	Dede grows vegetables in the garden
7	Dia ngeguai kan guring di dapor	She cooks fried rice in the kitchen
8	Adek ngisik kucing di mahan	My sister has a cat at home
9	Ibuk ngelajar sanak ngebaca di kamar	Mom teaches the child to read in the room
10	Sikam nyiram bunga di pengadapan	I watered the flowers in the yard
11	Ama nanom pari di sabah	Uncle planted rice in the field
12	Atu ngebabay umpu di pengadapan	Grandfather holding grandson in the yard
13	Kucing pedom di lammung kerasi	Cat sleeping on a chair
14	Ateh ngusung keranjang di pasar	Grandma carries a basket at the market
15	Mak wan ngeberasihko mahan jeno pagi	Auntie cleans the house in the morning
16	Ateh ngayam sulan di debah batang	Grandmother weaving a mat under a trunk
17	Ayah ngusung mubil baru	Father drives a new car
18	Umik mupoh pakaian kamak	Mom washes dirty clothes
19	Adek nganik buah manga.	Sister ate a mango
20	Daing main sepida di lapangan.	Brother plays bicycle in the field

2.2 Proposed Speech-to-Text Conversion Process

Speech-to-text conversion (STT) processes and identifies words and phrases in spoken language into text format by accurately analysing speech signals through a system with a trained dataset. STT implementation has been widely applied in various types of research such as in machine translation that facilitates user interaction. STT applications in many diverse research domains include voice command recognition in automobile systems, medical documentation, translation, powered virtual assistants and voice command systems for people with disabilities [8]. The speech-to-text conversion process in this study is shown in the block diagram in Figure 1.

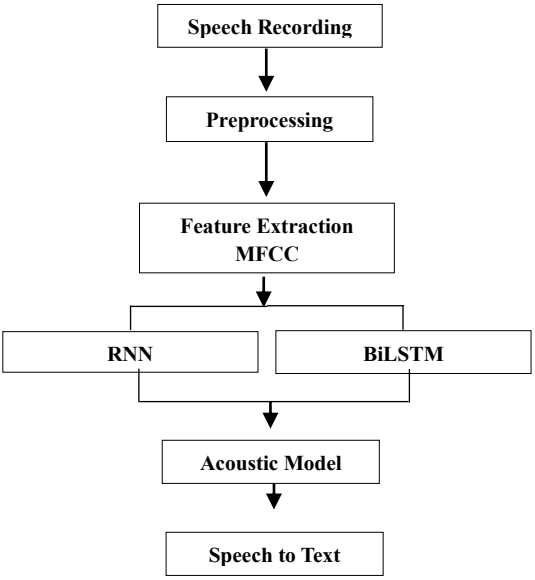


Fig. 1. Block Diagram of Speech to Text Conversion Process

2.3 Evaluation Metrics

The Word Error Rate (WER) evaluation is used in this study to determine the error ratio of the STT transcript output. In addition, this evaluation measures the success of the speech-to-text conversion process. The WER evaluation is based on the number of word substitutions, insertions, and subtractions used to convert speech to text. The calculation can be seen as follows:

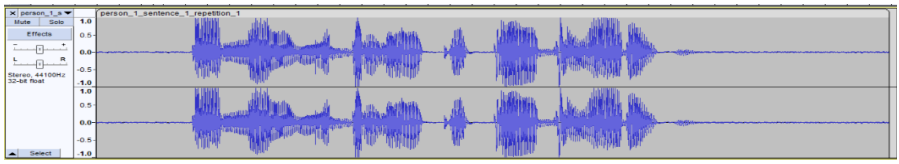
$$\text{WER} = \frac{1}{2} \sum \text{ED}\,(h(x)) \tag{1}$$

with known $ED\ (h(x))$ is the number of replacements, insertions and subtractions of words in the target sentence [9].

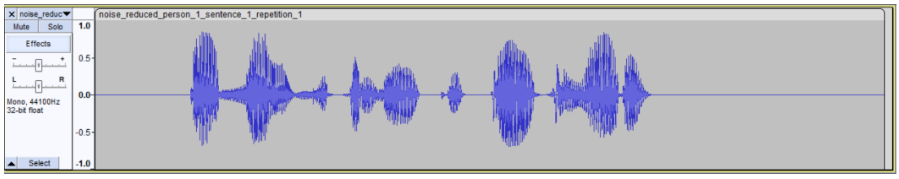
3 Results

3.1 Signal Preprocessing

In speech recognition experiments, a pre-processing stage is required by processing the recorded audio data in wav format followed by cleaning and noise removal. Proper noise reduction and removal will affect the performance of the STT conversion system. The preprocessing stage includes normalization, pre-emphasis, noise reduction and silence removal using the Python library Pyrus. Normalization is adjusting the audio amplitude to a standard range and ensuring all signals have the same amplitude level to make the analysis in feature extraction more accurate. Pre-emphasis suppresses high frequencies in the sound signal and removes noise in the signal. The removal of silence pauses and background noise in the sound environment is essential to focus on the required part of the speech signal and streamline the speech recognition algorithm.



(a)



(b)

Fig. 2. Speech Signal (a) before and (b) after Preprocessing

Figure 2 shows the difference in audio signals before (a) and after (b) the preprocessing phase is performed using Audacity software. From the figure, it can be seen that before preprocessing the audio amplitude signal is more than 0.5, and some areas of noise and silence at the beginning and end of the sound. In contrast, in Figure 2 (a) after preprocessing, the average signal amplitude is at a value of 0.5 which is the ideal sound signal [10] and the noise and silence areas are significantly reduced, and the duration of the sound signal is reduced from 3.6 to 2.6 seconds.

3.2 MFCC Extraction

In speech recognition experiments, the pre-processed speech signals go through a feature extraction process. In this process, the speech signal is analyzed to extract important features. Several popular techniques can be used including Mel Frequency Cepstrum Coefficient (MFCC), Linear Predicted Coefficient (LPC), and Perceptual Linear Prediction (PLP) [11]. Extraction techniques derive the main attributes of the

speech signal, so that spectrum-based parameters can be generated for each word. In many studies, it is found that MFCC features are considered the most widely used feature extraction technique by producing better results [12]. MFCC utilizes the critical bandwidth of the human ear which is known to change with frequency and is based on the perception of human hearing which is unable to detect frequencies above 1 kHz. The Mel-scale in MFCC characterizes the spectrogram frequencies to resemble and match how humans hear and respond to sound [13].

$$f_{mel} = 2595.LOG_{10}\left[1 + \frac{f_{lin}}{700}\right] \quad (2)$$

MFCC uses two types of filters arranged linearly at low frequencies below 1000 Hz and logarithmically above 1000 Hz. In addition, the Mel filter bank in MFCC provides a more perceptually appropriate representation of the speech signal that is less critical to frequency changes. Extraction of acoustic signal representation is an important stage to improve recognition performance and the efficiency at this stage significantly influences the effectiveness of subsequent stages. The next step is to identify the important components of the speech signal that are used to recognize the linguistic content and discard all other information such as noise, emotion, and others. The main components in MFCC are the feature extraction steps: Pre-emphasis is applied to flatten the spectrum of the input speech signal. Framing and Windowing serve to reduce the constantly changing fluctuations of the audio signal and divide the signal into segments. Fast Fourier Transform (FFT) computes the Discrete Fourier Transform (DFT), which converts the signal from the time domain to the frequency domain. Mel Filter simulates the way human hearing works in capturing sound frequency information, consisting of a series of filters distributed along the Mel frequency scale, which is computed using the formula (2). This formula transforms linear frequency in Hertz into the Mel scale, providing a more perceptually relevant representation of sound by compressing higher frequencies and preserving more detail at lower frequencies. The log is used to mimic the response of the human ear to changes in sound intensity. While the DCT converts the spectral representation of the log Mel filter bank output into the cepstrum domain which generates the MFCC coefficients.

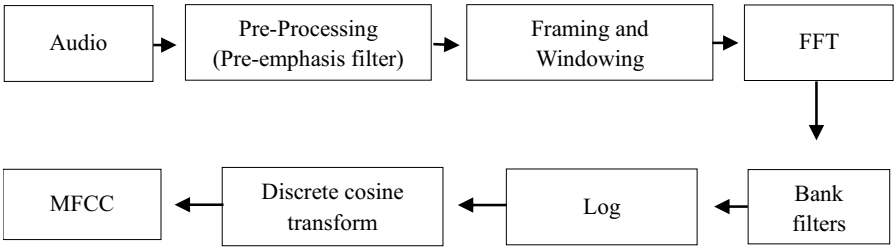


Fig. 3. Flow diagram of MFCC extraction process [14]

By using the librosa audio library in python. We extracted audio features that will be used in the implementation stage of Recurrent Neural Network (RNN) and Bidirectional LSTM models. As shown in Figure 4, the MFCC plot of one of the extracted

audio samples reveals a clear yellow spectrum, indicating that the key features of the vocal structure and formants are preserved despite noise reduction. The overall reduced variablity results in the middle and high MFCCs further reflects the effectiveness of the noise and background reduction, resulting in clean and focused audio.

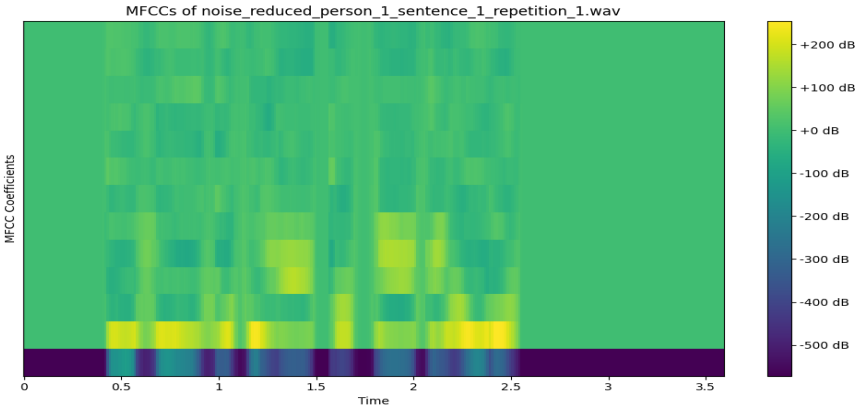


Fig. 4. Example of MFCC Plot

3.3 Training Dataset

The proposed approach is to implement RNN and BiLSTM model architectures to handle speech-to-text tasks and obtain a comparison of the results of both models that have two different characteristics of information flow and data processing. MFCC cepstral features from the extraction process are used as input to the RNN and BiLSTM models. RNNs are designed to process data such as text, voice signals and other time series data where the network has loops within it that allow information to be passed from one time step to the next depending on the current input and the result of the previous step. The BiLSTM model works by processing sequential data by capturing bidirectional information from the past (t - 1) and future (t + 1) and has multiple layers to create a high-level representation of the acoustic data [15].

The performance of the training model on the dataset is determined by the accuracy and loss value obtained from the results. The learning rate, dropout, and early stopping techniques were adjusted to prevent overfitting. The vocabulary size (number of recognized words) was limited to 29 tokens, including letters and some symbols. The model was trained using the commonly used Cross-Entropy Loss in multiclass classification with the Adam optimizer and the initial learning rate. Figure 5 shows the accuracy and validation loss curves of the RNN and BiLSTM models during the training process. Based on the plot, both models show a consistent decrease in loss during the initial epoch of training, indicating the model is not overfitting. On the training accuracy, both models show a gradually increasing trend with the validation accuracy of the RNN at 40% and the BiLSTM obtaining a higher value of 50%.

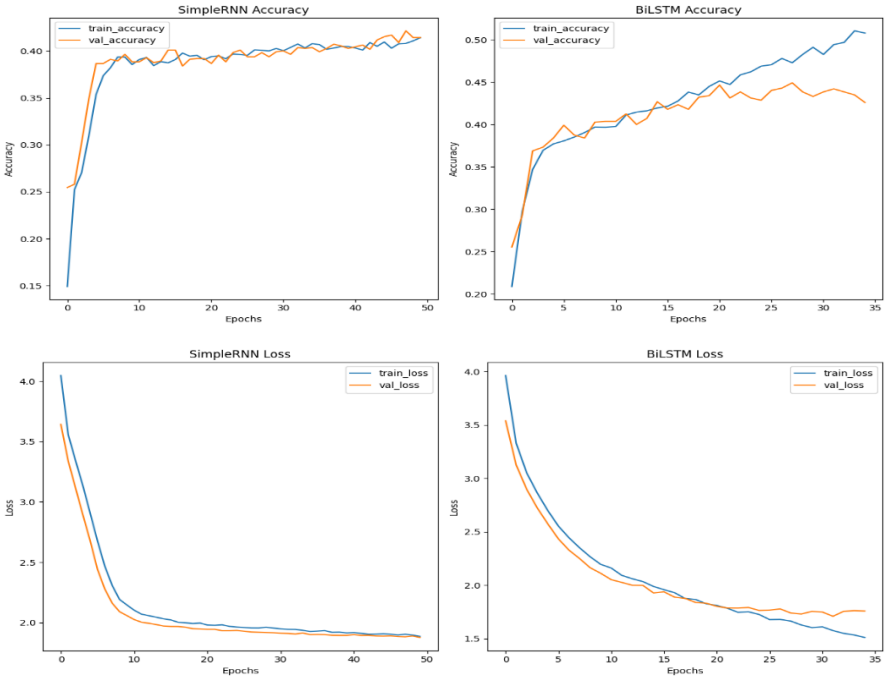


Fig. 5. Training and Validation Loss Values in RNN and BiLSTM

3.4 Testing Model

The sample data in the testing model is 80 audio samples that are not included in the training process. This number is based on the initial composition decided in the preparation stage. The accuracy is measured by the WER metric, where obtaining a low WER value indicates a high accuracy in testing. The expected word output prediction is derived from the decoding process of the acoustic model, based on the RNN and BiLSTM models developed during training. The WER value for RNN is 0.5857 which indicates a fairly high error rate and the WER value for BiLSTM is 0.5589 which shows a slight improvement compared to RNN. The high WER value for both models indicates that there is a problem in the quality of transcription produced by the models. Table 2 shows examples of transcription results in the RNN and BiLSTM model training process.

Table 2. Result of Decoding Predictions of Models

RNN			BiLSTM		
True Sentence	**Prediction Sentence**	**WER**	**True Sentence**	**Prediction Sentence**	**WER**
sikam nyiram bunga di pengadapan	ateh main bal di lapangan	0.571	sikam nyiram bunga di pengadapan	**sikam nyiram bunga di** lapangan	0.142
ateh ngusung keranjang di pasar	daing main pakaian di	0.571	ateh ngusung keranjang di pasar	ateh ngusung sayuran di pengadapan	0.285
lita mupoh pakaian di duwai	daing ngusung pakaian di duwai	0.285	lita mupoh pakaian di duwai	ateh nanom pakaian di lapangan	0.428
ayah ngusung mubil baru	ateh ngusung pakaian di duwai	0.571	ayah ngusung mubil baru	ayah ngusung mubil di	0.142
dede nanom sayuran di huma	adek main umpu di pengadapan	0.571	dede nanom sayuran di huma	ateh ngayam sulan di lapangan	0.571

This research does not employ a language model to optimize the transcription results, which indicates that the model is still not fully effective to recognizing words within the vocabulary. As a result, some words are marked as OOV (Out of Vocabulary). This arises due to the limited size of the vocabulary or the tokenization process. This challenge arises due to the limited vocabulary resources of Lampung language in Api dialect. Nevertheless, the RNN and BiLSTM models can show quite good results on pattern recognition achieving up to 50% with the prediction results illustrated in table 2. Based on the table 2, the lowest WER in the BiLSTM model is 0.142 where almost all words can be transcribed according to the ground truth sentence. While the highest average WER in sentence prediction is 0.5 to 0.8. This is influenced by some transcriptions that are not recognized well enough by the model.

In addition to training and testing with the RNN and BiLSTM models, this research further tried the use of acoustic models to see the representation of voice signals to produce text transcriptions. The results of the acoustic model process are shown in figure 6 with a very good accuracy of 98% accuracy, which shows that this model is very good at recognizing acoustic patterns as well as lower WER values. The limitation of both the previous model and this model is that it does not use a language model to optimize the transcription results so that the results of unreadable transcription or out of vocabulary become more dominant.

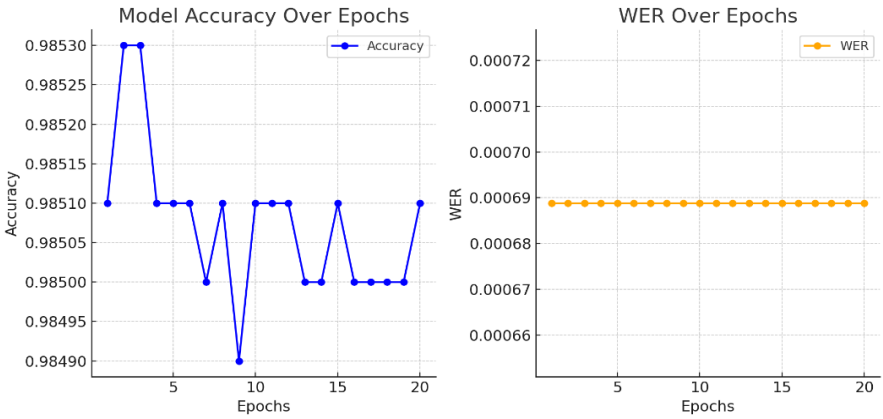


Fig. 6. Model Accuracy and WER over Epochs

4 Conclusion

The RNN and BiLSTM models have been successfully trained with an accuracy of around 50% with BiLSTM giving slightly better results as it has a better ability to understand sequential context, but still needs to be improved due to the high number of WER and OOV tokens that appear in the prediction. The results of this study are good enough with various limitations and challenges such as not using a language model to optimize the transcription results.

The high accuracy of the acoustic model obtained does not always correlate directly with the quality of the transcription produced. It can be seen from the transcription results that are not readable on the acoustic model because many words are not recognized by the model. In this study, a low WER on the acoustic model indicates that the model makes relatively few errors in recognizing and predicting word order, but this does not fully reflect that the transcription output is readable or linguistically meaningful. Suggestions for improving the vocabulary by increasing the size of the vocabulary and data augmentation as well as the use of language models can address the main challenges of the results.

References

1. Jumlah Penutur Bahasa Lampung, https://lampung.antaranews.com/berita/706452/kantor-bahasa-sebut-jumlah-penutur-bahasa-lampung, last accessed 2024/08/25
2. Ananta, A., Arifin, E. N., Hasbullah, M. S., Handayani, N. B., Pramono, A.: Demography of Indonesia's Ethnicity. Institute of Southeast Asian Studies, Singapore (2015). ISBN 978-981-4519-88-5. OCLC 1011165696
3. Zalmansyah, A.: Bahasa Lampung di Kalangan Anak Muda Lampung. Kelasa **14**(2), 145-158 (2019)
4. Wang, Z. J.: Putting Humans in the Natural Language Processing Loop: A Survey. Bridging Human-Computer Interaction and Natural Language Processing. In: HCINLP 2021 -

Proceedings of the 1st Workshop, pp. 47–52. Association for Computational Linguistics (2021)
5. Islam, J., Mubassira, M., Islam, M. R., Das A.K.: A speech recognition system for Bengali language using recurrent neural network. In: 2019 IEEE 4th International Conference on Computer and Communication System, pp. 73-76. IEEE (2019)
6. Dwijayanti, S., Tami, M. A., Suprapto, B. Y.: Speech-to-Text Conversion in Indonesian Language Using a Deep Bidirectional Long Short-Term Memory Algorithm. (IJACSA) International Journal of Advanced Computer Science and Applications **12**(3) (2021)
7. Permata, P., Abidin, Z.: Statistical Machine Translation Pada Bahasa Lampung Dialek Api Ke Bahasa Indonesia. Media Inform. Budidarma **4**(3), 519–528 (2021). doi: 10.30865/mib. v4i3.2116.
8. Sultana, S., Akhand, M., Das, P. K., Rahman, M. H.: Bangla speech-to-text conversion using sapi. In: 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 385–390. IEEE (2012)
9. Ahmed, A., Renals, S.: Word error rate estimation for speech recognition: e-WER. In: Proc. of the 56th Annual Meeting of the Association for Computational Lingustics (vol 2 short paper), pp. 20-24. Association for Computational Linguistics, Melbourne (2018)
10. Audacity, Guide to Using Audacity. (2018)
11. Dave, N.: Feature extraction methods LPC, PLP and MFCC in speech recognition. International Journal for Advance Research in Engineering and Technology **1**(6) (2013)
12. Benba, A., Jilbab, A., Hammouch, A.: Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease. IRBM **38**(6), 346-351 (2017). https://doi.org/10.1016/j.irbm.2017.10.002
13. Koeing, W.: A new frequency scale for acoustic measurements. Bell Telephone Laboratory Record **27**, 299-301 (1949)
14. Kaddoura, T., Vadlamudi, K., Kumar, S., Bobhate, P., Guo, L., Jain, S., Elgendi, M., Coe, JY., Kim, D., Taylor, D., Tymchak, W., Schuurmans, D., Zemp, R., Adatia, I.: Acoustic diagnosis of pulmonary hypertension: automated speech-recognition-inspired classification algorithm outperforms physicians. Scientific Reports **6**, 33182 (2016). https://doi.org/10.1038/srep33182
15. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proc. 23rd Int. Conf. Mach. Learn., pp. 369– 376. Association for Computing Machinery, New York (2006)