

Speech Recognition for Lampung Language Using Wav2Vec 2.0

Hendra Kurniawan¹, Akmal Junaidi², Favorisen R. Lumbanraja³, Wamilian⁴

¹Doctoral Program, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

¹Faculty of Computer Science, Institute of Informatics and Business (IIB) Darmajaya, Lampung, Indonesia

^{2,3,4}Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

*Corresponding author: 2437061015@students.unila.ac.id

ABSTRACT

Indonesia is home to more than 700 regional languages, many of which are threatened by declining numbers of speakers. One of these languages is Lampung, a low-resource language that is experiencing language shift and faces the risk of extinction. Automatic Speech Recognition (ASR) can support language preservation by enabling speech-to-text conversion; however, developing ASR systems for Lampung remains challenging due to the limited availability of annotated speech data. This study investigates the adaptation of self-supervised speech models for low-resource Lampung ASR through fine-tuning of the Wav2Vec 2.0 Base (W2V2-B) and Wav2Vec 2.0 Large (W2V2-L) architectures. A Lampung speech corpus consisting of approximately 10 hours of transcribed speech and 17,682 audio recordings was used for training and evaluation. The dataset was divided into training and testing subsets, and model performance was assessed using Word Error Rate (WER). The main contribution of this work is the establishment of a benchmark for Lampung ASR using state-of-the-art self-supervised learning models and an empirical comparison of W2V2 architectures in a low-resource language setting. Experimental results show that W2V2-B achieved a WER of 36.23%, while W2V2-L achieved a WER of 36.30%. These findings demonstrate the feasibility of applying self-supervised speech representation learning to low-resource Lampung languages and provide a foundation for future research on multilingual transfer learning and data augmentation for Lampung ASR.

Keywords: Lampung Language, Speech Recognition, Fine-tuning, W2V2, Low-resource Language.

1. INTRODUCTION

Indonesia is a country of a thousand islands, each region with its own unique culture and language. This cultural and linguistic diversity is spread throughout Indonesia, from Aceh to Papua. Indonesia's unique regional languages are estimated to have over 700 (Aji et al., 2022). Indonesia ranks second among countries with the most regional languages, after Papua New Guinea. As one of Indonesia's over 700 regional languages, Lampung is the mother tongue of the inhabitants of Lampung Province (Nasution et al., 2008).

Lampung language is divided into 2, which Pepadun and Saibatin. The differences in Lampung language stem from geographical differences. Lampung language, with the O/nyow dialect (Pepadun), is used by Lampung people in non-coastal areas. Meanwhile, Lampung, with the A/Api dialect (Saibatin), is the language used by coastal communities. Lampung language itself is experiencing a speech shift, one of which is caused by multicultural and multilingual immigrant communities (Putri, 2018). This phenomenon has led to a decline in Lampung use among native speakers and a preference for Indonesian (Dewi et al., 2018). This is in accordance with Moseley's statement in the UNESCO-published book "Atlas of the World's

Languages in Danger," which states that languages spread throughout the world have the potential to become extinct (Moseley, 2012).

Traditionally, ASR systems are modular, with several models consisting of three components (acoustic model, pronunciation model, and language model) that operate independently (Khudhair & Talib, 2022). However, over time, automatic speech recognition has progressed rapidly due to the development of deep learning models, such as DNN (Dahl et al., 2012; Sobti et al., 2024), CNN (Fantaye et al., 2020; Sainath et al., 2013), RNN (Graves et al., 2013), *end-to-end* models (Amodei et al., 2016; Bahdanau et al., 2016; Battenberg et al., 2017), *sequence-to-sequence* models (Prabhavalkar et al., 2017; Sutskever et al., 2014) which show superior performance compared to conventional hybrid models (Trentin & Gori, 2001). These models can directly leverage neural networks to learn the mapping from speech to text, thereby eliminating the need for complex modeling. There are three key success factors of deep learning in ASR: 1) large datasets, 2) advances in graphics processing units, and 3) improvements in learning algorithms and model architectures (Hannun, 2021).

ASR has become a fundamental problem in artificial intelligence. ASR performance on high-resource languages demonstrates superiority over neural models (Amodei et al., 2016; Chan et al., 2016; Gulati et al., 2020; Hannun et al., 2014). However, most of the world's languages, even those spoken by tens of millions of people, lack sufficient transcribed speech and are considered low-resource languages. Low-resource languages account for a significant proportion of the world's languages, with 94% spoken by fewer than 1,000,000 people (Eberhard et al., 2021). Developing ASR for low-resource languages is an urgent need and a concern for preserving language extinction and preserving cultural heritage. ASR for low-resource languages presents a unique challenge, and a series of studies have focused on it (Chen & Mak, 2015; Du et al., 2023; Miao et al., 2013; Sutskever et al., 2014; Thomas et al., 2013).

Many studies have looked at how speech recognition can be used with local languages. One such study is the application of speech recognition to local languages, such as Mvsvoke, using Meta's Massively Multilingual Speech (MMS) model (Mainzinger & Levow, 2024) which is a representation of the W2V2 architecture (Baevski et al., 2020a; Pratap et al., 2024) pre-trained on unlabeled data from 1,406 languages.

The approach adopted in this research involves fine-tuning the W2V2 model pre-trained on the labeled Lampung language dataset. W2V2 is a leading model developed by Facebook AI for tasks such as ASR and speech representation learning (Bawitlung et al., 2025), and it has established its architecture and been trained across various languages. It can improve the model's speech recognition performance through fine-tuning. Furthermore, fine-tuning can adapt the W2V2 model to specific datasets, enabling it to be used for the desired language (Becerra et al., 2022). Previous studies have shown that W2V2 is highly effective for speech recognition tasks and holds significant potential for enhancing the performance of existing systems. Its primary advantage lies in its ability to operate with limited labeled data while achieving optimal speech representations.

2. METHODS

2.1 Method Used

This research fine-tunes two W2V2 models: W2V2B and W2V2L. The main differences are as follows. W2V2B has 12 transformer blocks, a model dimension of 768, an inner dimension (FFN) of 3,072, and 8 attention heads. For this model, batches are created by cropping 250,000 audio samples (15.6 seconds) from each example. In contrast, W2V2L has 24 transformer blocks, a model dimension of 1,024, an inner dimension of 4,096, and 16 attention heads. For W2V2L, 320,000 audio samples (20 seconds) are cropped from each example (Baevski et al., 2020b).

2.1.1 W2V2

W2V2 is an unsupervised system that learns to represent speech directly from raw audio. It uses a convolutional feature encoder to turn audio signals into simpler, hidden forms. This representation is then randomly hidden in the latent space, allowing the model to learn context from previously unseen audio data. The next step involves a Transformer network that generates contextual representations via self-attention, allowing the model to capture relationships between audio features over time (Baevski et al., 2020b).

Figure 1 explains to address the problem of unlabelled data, W2V2 uses a quantization module to transform the latent representation into discrete units that serve as training targets. The model is trained using a contrastive task, in which it must distinguish the correct hidden representation from a set of distortions. After pre-training, the model is fine-tuned on labelled data using Connectionist Temporal Classification (CTC) as the loss function to map the audio representation to the transcribed text. This combination of unsupervised learning, fine-tuning, quantization, and refinement allows W2V2 to perform well even with limited datasets, making it an excellent solution for ASR across multiple languages and dialects (Baevski et al., 2020b). The architecture of W2V2 is shown in Fig 1.

The W2V2 architecture consists of a multi-layer convolutional feature encoder that processes raw audio and generates latent speech representations for each time step. These representations are then passed to a Transformer, which captures information from the entire sequence. The feature encoder's output is discretized using a quantization module to define targets for the self-supervised objective. W2V2 builds contextual representations from continuous speech, and self-attention models dependencies across the full sequence of latent representations (Baevski et al., 2020b).

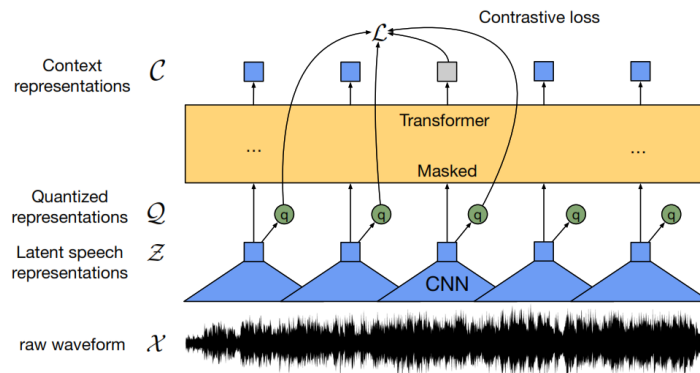


Figure 1. W2V2 Architecture (Baevski et al., 2020b)

The study introduces a Lampung language dataset that was built and recorded independently. The audio dataset was recorded using the UORRIS 3-in-1 Mic 3.5mm Wireless Lavalier Portable Audio Video microphone. The recorded audio data is speech with an average duration of 2 seconds. The dataset consists of audio files and transcriptions in the form of *.xlsx files. The constructed dataset contains transcriptions of audio data for the Lampung language. The Lampung language audio data consists of 12180 utterances, totaling approximately 6.76 hours. 5 speakers spoke the utterances. The audio files are named according to the order of the speakers and are adjusted to the transcription file. The speakers are male and female. The dataset will be split into two parts at an 80:20 ratio: 80% for training (approximately 5,41 hours) and 20% for testing (approximately 1,35 hours).

Table 1. Lampung Language Distribution Dataset

Category	Hours	Number of Utterances
Training	5,41	9744
Testing	1,35	2436

Table 2. Audio Properties

Audio Properties	Values
Format	wav
Bitrate	128Kbps
Sample Rate	48 Khz
Channel	Mono & Stereo

As mentioned in Table 1 and Table 2, the transcript comprises 6,76 hours of audio and 12180 speech files. The audio files are converted to the WAV format with a bitrate of 128 Kbps and a sample rate of 48 kHz, and each file contains mono or stereo channels.

2.1.3 Proposed Experimental

The experiments in this research will use code with Python and publicly available libraries. Two different W2V2 models will be tested: W2V2B and W2V2L as illustrated in Fig 2.

The Automatic Speech Recognition (ASR) experiment for the Lampung language was conducted using a fine-tuning approach using the Wav2Vec2 self-supervised learning model using the Connectionist Temporal Classification (CTC) scheme. The entire experimental pipeline included data preparation, character tokenizer development, audio signal processing, construction of the processed dataset, model training, and evaluation based on a transcription error metric.

The initial stage involved loading dataset metadata from a spreadsheet containing pairs of audio file names and Lampung-language transcriptions. Each metadata entry was mapped to a .wav audio file in the corpus folder. The system then validated the file's existence to ensure only available audio data was used in the experiment. The transcribed text was cleaned through a normalization process that included converting letters to lowercase, removing non-alphabetic characters, and consolidating double spaces into single spaces.

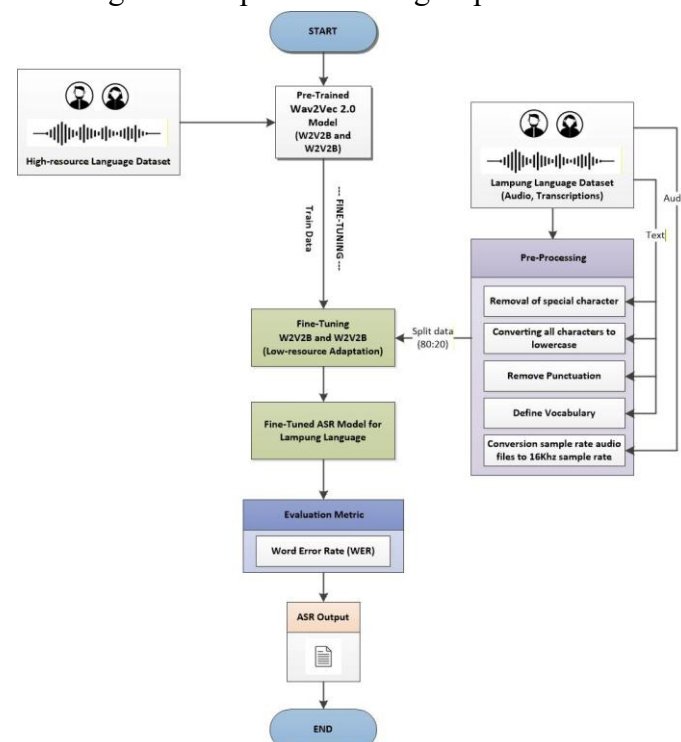


Figure 2. Proposed Experimental

The CTC model applies a character-based tokenizer, so the vocabulary is built solely from the training data text to avoid information leakage by extracting all unique characters, replacing spaces with special word-separator tokens, and adding special tokens for padding [PAD] and unknown characters [UNK]. Then the vocabulary is stored in JSON format and used to form the Lampung language CTC tokenizer. Next, the Wav2Vec2Processor is built, which combines a feature extractor and a tokenizer. Where the feature extractor is configured to receive an audio signal with a sampling rate of 16 kHz, perform amplitude normalization, and generate an attention mask. At the same time, the tokenizer uses a predefined character vocabulary and special word-separator tokens to represent spaces. The training and test data are then converted to the HuggingFace Dataset format to be compatible with the Trainer-based training pipeline, with each entry containing an audio path and a transcribed text.

In the signal processing stage, each audio file is read using the torchaudio library, converted to mono if it has more than one channel by averaging the channels, then resampled to 16 kHz if the initial sampling rate is different, and converted to a float32 numeric representation so that it is ready for processing by the processor. The processor then extracts waveform features and normalizes them into input values for the model. At the same time, the text is transcribed into a sequence of character tokens as CTC labels, with spaces automatically converted to word-separator tokens. Because the length of the audio signal and the length of the labels vary, a CTC-specific data collator is used to pad the input and labels dynamically, and to set the label padding value to -100 to be ignored in the CTC loss calculation.

Next, the experiment proceeds to training and evaluation, with the dataset randomly split into 80% for training and 20% for testing, using a fixed seed for reproducibility. Before starting the training and evaluation, two things are done: creating a data collator and setting up the training arguments. Training arguments define the settings for the training process, such as the maximum number of training steps, batch size, learning rate, and other model hyperparameters. In this paper, the model trained using python programming and open-source libraries on an GPU with 8 GB and RAM with 32 GB. The training used parameter settings designed for limited resources, as documented in the source. The output directory was set to `"/wav2vec2-finetune-lampung"`. Both the training and evaluation batch sizes were set to 8 to avoid memory problems. We trained the model for 30 epochs, allowing it to process the full dataset 30 times, which helped the algorithm improve while reducing the risk of overfitting. The evaluation and saving steps occurred every 500 steps, and we logged metrics every 50 steps to track progress. We used a 500-step warmup with a learning rate schedule to stabilize training. Only two checkpoints were saved to save disk space. If a GPU was available, 16-bit precision was used to speed up training; otherwise, 32-bit precision was used. We set the random seed to 42 to ensure reproducibility. The options to push the model to Hugging Face Hub and to report logs externally were both turned off.

3. RESULTS AND DISCUSSIONS

The experiments in this paper show that the W2V2B and W2V2L models were successfully fine-tuned for Lampung. W2V2 models were also evaluated using the WER metric.

This study compares two W2V2 models and highlights their performance differences. Table 2 presents the word error rate (WER) for each model. The W2V2B model outperforms the W2V2L model, likely because the fine-tuning dataset is too small. The W2V2B and W2V2L models have WER values of 36.23% and 36.30%, respectively, indicating that both make errors in about 36% of words compared to the reference transcription. Table 1 demonstrates that the W2V2 model maintains low WERs, even with a small dataset and multiple languages. This suggests the model can create effective speech representations even with limited training data. The results also show that the W2V2B model performs about as well

as the W2V2L model when trained on a small dataset as illustrated in Fig 2. We evaluate the model's transcription of Lampung speech by measuring the word error rate, defined as:

$$\text{Word Error Rate} = \frac{S + D + I}{N}$$

Here, S stands for substitutions, D for deletions, I for insertions, and N for the total number of words in the reference. Word Error Rate (WER) measures overall transcription errors by counting incorrect, missing, and extra words in the output.

Table 1. Word Error Rate (WER) Results

Model	WER (%)
W2V2B	36,23%
W2V2L	36,30%

Table 2. Differences WER Results between Other Research

Languages	Model	Duration audio (hours)	WER (%)
Brazilian Portuguese	Fine-tune W2V2B & W2V2L	470 h	10.5% & 12.4%
Sundanese	Fine-tune W2V2B & W2V2L	53 h	23.5% & 24.0%
This research	Fine-tune W2V2B & W2V2L	6,76 h	36,2% & 36,3%

Based on the comparison Table 2 of WER results between studies, several conclusions can be drawn. First, the length of the training data appears to influence the performance of W2V2-based ASR strongly. The study on Brazilian Portuguese, which used a very large dataset (470 hours), reported the lowest WER of 10–12%, indicating that the W2V2 fine-tuning model can achieve high performance when supported by a large corpus. Second, in a medium-resource language such as Sundanese with a duration of 53 hours, the WER value increased to around 23–24%. However, this performance is still considered good for low- to medium-resource scenarios. It indicates that the pretrained model can adapt quite effectively even with data that is not as large as hundreds of hours. Third, in this study, with a data duration of only around 6.76 hours, the WER was around 36%, which is indeed higher than the two comparative studies, but still shows that W2V2 fine-tuning is still capable of producing a functional ASR system under very low-resource conditions. The small difference in WER between the Base and Large models across all rows of the table also indicates that increasing model size does not always provide significant improvements when training data is very limited. Overall, the table supports the conclusion that the W2V2 fine-tuning approach remains feasible for resource-constrained languages, but performance is significantly affected by the dataset's duration and scope. In other words, increasing the quantity and diversity of training data can significantly reduce WER.

Table 3. Differences W2V2B & W2V2L Reference and Predicted Comparisons for Lampung Dataset

No	Reference	Predicted (W2V2B)	Predicted (W2V2L)
1	satpam merikso kendaraan di gerbang	ulam merikso kendaraan di gerbang	dpem merikso kendaraan di gerbang
2	kemaman matcing ughang ditulung	kemaman matcing uhang ditulung	kemaman matcing kughang ditulung
3	nikew buguh binatang peliharaan	nikew buguh binatam peharaan	nikew buguh binatam periharaan
4	ikamjo ngehias kelas adino nyiapke acara perpisahan	ikamjo ngehias kelas adino ngenyiapke acar perpisahan	ikamjo ngehias kelas adino ngenyiapke acar perpisahan
5	ayah ngebo keluarga dak restoran untuk mengan debingei	ayah ngebo keluarga dak restoran untuk mengan debingei	ayah ngebo keluarga dak restoran untuk mengan debingei

Based on Table 3, a comparison of the transcription results between the reference text and the model outputs for five speech samples shows that W2V2B and W2V2L achieve relatively comparable performance with similar error patterns. In the first sample, both models make errors on the initial word, but still correctly recognize the core phrase of the sentence, with the Base variant producing a form that is slightly closer phonetically than the large variant. In the second sample, both models maintain the main sentence structure, but Base shows a smaller deviation from the words than Large. In the third sample, both models produce predictions that are nearly identical to the reference, demonstrating excellent recognition performance on sentences with words that are more common or occur more frequently in the training data. In the fourth sample, both models again successfully recognize most of the sentence structure but make errors on the final verb, with the large model producing a form closer to the reference than the Base model. In the fifth sample, the predictions of both models are nearly identical, with most words recognized correctly, with only minor differences that do not change the main meaning of the sentence. Overall, the performance difference between the W2V2B and W2V2L models on this limited test data is insignificant, with errors most frequently occurring on initial and low-frequency words. This finding corroborates the quantitative WER evaluation results, which showed very little difference between the two model variants, suggesting that the Base model is a more computationally efficient option with competitive prediction quality.

4. CONCLUSION

The results of this experiment show that the W2V2B model has a slightly lower WER than the W2V2BL model. This shows that models with more or larger parameters do not necessarily yield better WER values, as in the two studies presented in Table 2: Brazilian Portuguese

(Stefanel Gris et al., 2022) and Sundanese (Cryssiover & Zahra, 2024). Looking at the word error rate (WER) results for the two models, the W2V2 models are very promising, despite limited data availability, underscoring their robustness and adaptability. This study is expected to make a significant contribution to speech recognition in Lampung, the target language, by expanding and refining the Lampung language dataset to achieve better WER.

5. REFERENCES

- Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasajo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022(May), 7226–7249. <https://doi.org/10.18653/v1/2022.acl-long.500>
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Damos, G., Ding, K., Du, N., Elsen, E., ... Zhu, Z. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *Proceedings of The 33rd International Conference on Machine Learning*, 173–182. <https://proceedings.mlr.press/v48/amodei16.html>
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020a). Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, 2020(June), 12449–12460.
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020b). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *In Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020(Dec), 12449–12460.
- Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016(Mar), 4945–4949. <https://doi.org/10.1109/ICASSP.2016.7472618>
- Battenberg, E., Chen, J., Child, R., Coates, A., Li, Y. G. Y., Liu, H., Satheesh, S., Sriram, A., & Zhu, Z. (2017). Exploring neural transducers for end-to-end speech recognition. *In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017(Dec), 206–213. <https://doi.org/10.1109/ASRU.2017.8268937>
- Bawitlung, A., Dash, S. K., & Pattanayak, R. M. (2025). Mizo Automatic Speech Recognition: Leveraging Wav2vec 2.0 and XLS-R for Enhanced Accuracy in Low-Resource Language Processing. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(7), 72:1-72:15. <https://doi.org/10.1145/3746063>
- Becerra, H., Ragano, A., & Hines, A. (2022). *Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction*. 4088–4092. <https://doi.org/10.21437/Interspeech.2022-10766>
- Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *In 2016 IEEE International*

- Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016(Mar), 4960–4964. <https://doi.org/10.1109/ICASSP.2016.7472621>
- Chen, D., & Mak, B. K.-W. (2015). Multitask Learning of Deep Neural Networks for Low-Resource Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7), 1172–1183. <https://doi.org/10.1109/TASLP.2015.2422573>
- Cryssiover, A., & Zahra, A. (2024). Speech recognition model design for Sundanese language using WAV2VEC 2.0. *International Journal of Speech Technology*, 27(1), 171–177. <https://doi.org/10.1007/s10772-023-10066-5>
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 30–42. <https://doi.org/10.1109/TASL.2011.2134090>
- Dewi, R., Ariyani, F., & Rusminto, N. E. (2018). Pemertahanan Bahasa Lampung dalam Ranah Pendidikan. *Jurnal Tiyuh Lampung: Pendidikan Bahasa dan Kebudayaan*, 3(1), 48–57.
- Du, Y.-Q., Zhang, J., Fang, X., Wu, M.-H., & Yang, Z.-W. (2023). A Semi-Supervised Complementary Joint Training Approach for Low-Resource Speech Recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31, 3908–3921. <https://doi.org/10.1109/TASLP.2023.3313434>
- Eberhard, D. M., Simon, G. F., & Fennig, C. (2021). *Ethnologue: Languages of the world* (24th ed.). SIL International. <http://www.ethnologue.com>
- Fantaye, T. G., Yu, J., & Hailu, T. T. (2020). Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition. *Computers*, 9(2), 1–27. <https://doi.org/10.3390/computers9020036>
- Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013(May)*, 6645–6649. <https://doi.org/10.1109/ICASSP.2013.6638947>
- Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, 2020(Oct), 5036–5040. <https://doi.org/10.21437/Interspeech.2020-3015>
- Hannun, A. (2021). *The History of Speech Recognition to the Year 2030* (arXiv:2108.00084). arXiv. <https://doi.org/10.48550/arXiv.2108.00084>
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). *Deep Speech: Scaling up end-to-end speech recognition*. 2014(Dec), 1–12. <https://doi.org/10.48550/arXiv.1412.5567>
- Khudhair, M., & Talib, A. (2022). Improving Low Resources Arabic Speech Recognition using Data Augmentation. *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, 60–65. <https://doi.org/10.1109/CSCTIT56299.2022.10145613>
- Mainzinger, J., & Levow, G.-A. (2024). Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke. In X. Fu & E. Fleisig (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 170–176). Association for Computational Linguistics. <https://aclanthology.org/2024.acl-srw.16>

- Miao, Y., Metze, F., & Rawat, S. (2013). Deep maxout networks for low-resource speech recognition. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 398–403. <https://doi.org/10.1109/ASRU.2013.6707763>
- Moseley, C. (2012). *Atlas of the World's Languages in Danger*. UNESCO Digital Library. <https://unesdoc.unesco.org/ark:/48223/pf0000187026>
- Nasution, H., Rahayu, R., Wibowo, E., Harum, D. M., & Moses, F. (2008). *Persebaran Bahasa-Bahasa di Provinsi Lampung*. Kantor Bahasa Provinsi Lampung.
- Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017). A Comparison of Sequence-to-Sequence Models for Speech Recognition. *In Interspeech 2017, 2017(August)*, 939–943. <https://doi.org/10.21437/Interspeech.2017-233>
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(1), 4798–4849.
- Putri, N. W. (2018). Pergeseran Bahasa Daerah Lampung Pada Masyarakat Kota Bandar Lampung. *PRASASTI: Journal of Linguistics*, 3(1), Article 1. <https://doi.org/10.20961/prasasti.v3i1.16550>
- Sainath, T. N., Mohamed, A., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013(May), 8614–8618. <https://doi.org/10.1109/ICASSP.2013.6639347>
- Sobti, R., Guleria, K., & Kadyan, V. (2024). Automatic Speech Recognition System for Low Resource Punjabi Language using Deep Neural Network-Hidden Markov Model (DNN-HMM). *International Journal of Intelligent Systems and Applications in Engineering*, 12(19s), Article 19s.
- Stefanel Gris, L. R., Casanova, E., De Oliveira, F. S., Da Silva Soares, A., & Candido Junior, A. (2022). Brazilian Portuguese Speech Recognition Using Wav2vec 2.0. *In Computational Processing of the Portuguese Language, Lecture Notes in Computer Science*, 13208(March), 333–343. https://doi.org/10.1007/978-3-030-98305-5_31
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*, 2(Dec), 3104–3112.
- Thomas, S., Seltzer, M. L., Church, K., & Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6704–6708. <https://doi.org/10.1109/ICASSP.2013.6638959>
- Trentin, E., & Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1), 91–126. [https://doi.org/10.1016/S0925-2312\(00\)00308-8](https://doi.org/10.1016/S0925-2312(00)00308-8)