# Enhancing Low-Resource Lampung Speech Recognition through Cross-Lingual XLSR-Wav2Vec 2.0 Pretraining

Hendra Kurniawan[1], Akmal Junaidi[2,*], Favorisen Rosyking Lumbanraja[3], Wamiliana[4]

[1]*Faculty of Computer Science, Institute of Informatics and Business Darmajaya, Indonesia*

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

[2,3,4]*Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia*

(Received: February 28, 2026; Revised: May 1, 2026; Accepted: June 12, 2026; Available online: June 28, 2026)

**Abstract**

This study investigates the application of Wav2Vec 2.0 (W2V2) and Cross-Lingual Speech Representation (XLSR) models to Lampung language speech recognition. LampungNyow v1.0 is introduced, a speech corpus designed to provide a baseline for training and evaluating Automatic Speech Recognition (ASR) for this low-resource regional language of Indonesia. The dataset enables supervised fine-tuning and standardized evaluation, addressing the lack of publicly available linguistic resources for Lampung. Several pre-trained W2V2 models on Lampung speech recognition using Word Error Rate (WER) as the evaluation metric. The evaluated models include W2V2-Base, W2V2-Large, W2V2-Large-XLSR-Indonesian, W2V2-Large-XLSR-Sundanese, W2V2-Large-XLSR-53, and the multilingual W2V2-Large-XLSR-Indonesia-Javanese-Sundanese model. Monolingual models have higher WER values, according to experimental results: W2V2-Base achieved 36,23%, while W2V2-Large achieved 36,30%. XLSR models, such as XLSR-53 (33,88%), Sundanese (33,99%), and Indonesian (33,70%), demonstrated modest improvements. The W2V2-Large-XLSR-Indonesian-Javanese-Sundanese model, which was the foundation for the Lampung automatic speech recognition system in this study, achieved lower WER of 17,39%. These findings suggest that, in contrast to more comprehensive multilingual or monolingual pretraining models, multilingual pretraining utilizing a number of Indonesian regional languages can produce acoustic and contextual speech representations that are better suited for the resource-constrained Lampung automatic speech recognition task. When compared to the baseline W2V2-Large model, the obtained WER of 17,39% indicates a relative improvement of more than 50%.

*Keywords:* Low-Resource Language, Cross-Lingual Learning, Automatic Speech Recognition, Wav2Vec 2.0, XLSR

## 1. Introduction

Communication between humans by speech is the most natural way, and the process of understanding and converting spoken language into text. ASR is an important technology for facilitating and improving human-human (HHC) and human-machine (HMC) communication interactions. Voice technology can eliminate barriers in HHC interactions. In the past, people speaking different languages needed human translators to communicate. This significantly limited who could communicate and when. Voice technology can also significantly improve HMC [1]. ASR systems can automatically transcribe speech and have been widely applied in various domains, including virtual assistants, customer service systems, voice search, mobile voice input, and online speech-to-text services [2], [3], [4], [5]. Recent advances in deep learning have significantly improved ASR performance and expanded its applicability to multilingual and low-resource language scenarios [6], [7], [8]. For example, the use of speech recognition for stroke patients through a finger training system that helps rehabilitation doctors [9], and speaker verification to overcome the problem of voice similarity using MFCC and fuzzy [10].

Indonesia is a country of a thousand islands, each region has its own unique culture and regional language. This cultural and linguistic diversity is spread throughout Indonesia, from Aceh to Papua. According to data from the Central Statistics Agency, Indonesia's population was about 284,438.8 million people [11] in mid-June 2025. Indonesia is also one of the most linguistically diverse countries, with about 10% of the world's languages spoken there. Indonesia is estimated to have regional languages other than Indonesian as the national language, with 700+ [12]. This linguistic

*Corresponding author: Akmal Junaidi (akmal.junaidi@fmipa.unila.ac.id)
DOI: https://doi.org/10.47738/jads.v7i3.1388

diversity represents an important cultural asset but also introduces challenges for the development of language technologies, particularly for regional languages with limited digital and speech resources. Many Indonesian regional languages remain underrepresented in Natural Language Processing (NLP) and speech technology research, including the Lampung language.

Lampung language is one of the 700+ regional languages in Indonesia and is the mother tongue for the people of Lampung Province [13], [14]. According to Sanusi et al. [15], The Lampung language comprises two dialects: O/Nyow (Pepadun) and A/Api (Saibatin). The differences in the Lampung language stem from geographical differences. Lampung language with dialect O is the language used by the Lampung people in non-coastal areas. Meanwhile, the Lampung language with the dialect A is used by coastal communities. Examples of the Lampung dialect A can be seen in the pronunciation of certain words. For example, in Lampung dialect A, the word "kota" (“city”) is pronounced [kuta], while in Lampung dialect O it is pronounced [kuto]. As time goes by and the era of globalization changes, the Lampung language itself undergoes a speech shift, one of which is driven by multicultural, multilingual immigrant communities, social change, and the dominance of the Indonesian language [13]. This phenomenon results in a decline in the use of Lampung by native speakers and a preference for speaking Indonesian [16]. In fact, the number of active Lampung language speakers has continued to decline due to language shift, urbanization, and the increasing dominance of Indonesian in daily communication [16].

Language preservationists and NLP researchers can be motivated to support efforts to document and revitalize endangered languages like Lampung by acknowledging the value of linguistic diversity. This is consistent with Moseley's assertion in the UNESCO book "Atlas of the World's Languages in Danger" that languages found all over the world may be in danger of going extinct [17]. Researchers have shown a great deal of interest in ASR research for indigenous or regional languages. In a similar vein, Lampung's language faces extinction due to this gap. Creating ASR for regional languages can be extremely important for maintaining identity and cultural legacy. People in Lampung will have greater access to ASR technology, which will facilitate their ability to comprehend and value their language and use it in daily interactions and communications. [18]. Recent studies have shown increased interest in languages with limited resources, as linguistic diversity and inclusivity are important in NLP. While most progress has focused on major languages, it is important to bring these advances to less-studied languages as well. For example, a study project on the Lampung dialect A/Api uses RNN and BiLSTM models [19] and applies perceptual linear prediction (PLP) [20].

Meanwhile, research on the Lampung dialect O has not yet been conducted. It involves models developed with the Transformer architecture, such as the XLSR model trained on a dataset consisting of 128 languages. However, the XLSR model design allows cross-linguistic resource transfer from resource-rich languages to improve the representation of languages excluded during model training by also adding Indonesian, Sundanese, and Javanese, which are related to the language family, along with Lampung. These three languages are the most widely spoken in Indonesia. Sundanese and Javanese can be categorized as languages that remain popular despite the linguistic heterogeneity that can influence them. Indonesian, Sundanese, and Javanese can also be considered high-resource languages compared to Lampung as a low-resource language. Models trained in this language can benefit the development of more robust ASR.

This research is part of our effort to build an ASR model that can help preserve Lampung and other similarly low-resource languages. This work presents a new corpus, Lampung Nyow v1.0, consisting of harmonized Lampung audio and text. This software is designed to accommodate a variety of tasks, including ASR, Text-To-Speech (TTS), and parts-of-speech (POS). This improvement makes the model more accurate by using the Transformer architecture, such as W2V2 and XLSR. This research examines the challenges posed by different text preprocessing methods, with a focus on the Lampung dialect O. These models often struggle to accurately predict or understand this dialect.

This research typically involves fine-tuning a pre-trained model by adjusting its parameters to optimize performance. This study was conducted by retraining the pre-trained model using a new dataset (in this case, the Lampung language dataset) so that the pre-trained model can learn and understand Lampung potentially more suitable than other languages, considering that the pre-trained W2V2/XLSR models have been trained on various languages, both monolingual and multilingual. Two monolingual and four multilingual pre-trained models were chosen for testing.

This paper is structured into several major sections, first: introduction, second: presenting research relevant to the research being conducted, third: methodology containing the W2V2/XLSR model architecture and how the W2V2/XLSR model is used, followed by an explanation of data collection, data preprocessing, pre-trained model

selection, hyperparameter setting, and the use of WER evaluation metrics. Fourth: presenting experimental results and discussions of the experimental results; fifth: Conclusion.

## 2. Related Works

Traditionally, ASR systems have been modular, with several models consisting of three components (acoustic model, pronunciation model, and language model) that operate separately [21]. However, over time, ASR has experienced rapid progress thanks to the development of deep learning models, such as DNN [22], [23], CNN [24], [25], RNN [26], model *end-to-end* [27], [28], model *sequence-to-sequence* [29], [30] which have demonstrated superior performance compared to conventional hybrid models [31]. The performance of ASR on high-resource languages , such as English, Chinese, and Hindi, demonstrates the advantages of using neural models [32], [33], [34]. These models work well when there is plenty of labeled data, but their performance drops when labeled data is limited. Wav2vec 2.0, created by Facebook AI, uses self-supervised learning to train on large amounts of unlabeled audio, then fine-tunes on smaller labeled datasets [35]. This two-step approach has been shown to improve ASR performance, especially when labeled data is scarce.

Low-resource speech recognition is difficult to learn because of limited training data. Therefore, through pre-training, the model can learn general information about speech in a large-resource source language, and then transfer the model to target-language speech recognition to improve its accuracy, also known as transfer learning. Several studies have utilized the W2V2 model and languages with large resources. One study [36] combines the Speech-Transformer architecture with Connectionist Temporal Classification (CTC) and Attention mechanisms in an end-to-end approach. The main innovation of this study is the adoption of the Wav2vec 2.0 (W2V2) model for pre-training on unlabeled data. Using FBank and MFCC as acoustic features, the model is trained in Chinese and English and tested in Hakka. The results show that the baseline performance with FBank achieves a Character Error Rate (CER) of 15.2% on the development data and 41.0% on the test data. The relative accuracy rate increased when using the Wav2vec 2.0 model, with a CER of 14.8% on the development data (Chinese and English) and 37.2% on the test data (Hakka) with drills on Chinese. A transfer learning study of the Czech pre-trained W2V2 model into Slovak [37].

This study explored three Slovak datasets, namely CommonVoice, VoxPopuli, and MALACH. Based on the evaluation results, the multilingual W2V2-XLS-R-300M achieved the best WER of 6.90% among other models, such as W2V2-cs, W2V2-ck, W2V2-cs-sk, and Whisper-Large, on the CommonVoice dataset. This study shows that a model trained on a wide, multilingual dataset with a large number of parameters (312 million) captures better acoustic representations, especially on datasets with high variation. Another study utilized the W2V2-XLS-R-300M model to transcribe Telugu [38]. Performance metrics demonstrated the proposed ASR system's performance, with a WER of 42.98%, indicating that the model still has room for improvement in accurately transcribing speech, especially in resource-constrained languages. Both the CER and PER were 9.65%, indicating that errors were less frequent at the character and phoneme levels. Most recognition errors, therefore, occurred at the word level. In another study, researchers working on Sundanese ASR with the OpenSLR dataset achieved a WER of 23.5% by fine-tuning the W2V2 model [39]. Similarly, another study on Javanese ASR, leveraging the XLSR variant of W2V2, achieved a WER of 17.95% [40]. In line with the utilization of XLSR, this study used XLSR-53 for Indonesian speech recognition [41].

This study demonstrates the model's performance, achieving an average WER of 20% without a language model. The WER can be increased to 12% with a language model. Another study on Mizo language ASR, leveraging the XLSR, reported a WER of 16.59% for the Wav2vec-Base-Mizo-Lus model, The XLSR-300M-Mizo-Lus model performed much better, reaching a WER of 11.84% and setting a new baseline for accuracy in Mizo [42]. Another study on Brazilian Portuguese [43], using 7 datasets and the W2V2-XLSR-53 model, reported an average WER of 12.4% and an increase in accuracy of 10.5% when applying a language model. A recent study [44] used W2V2 and conducted experiments across several languages, including English, French, German, Portuguese, Spanish, Italian, and Polish. The experimental results showed that the average WER ranged from 13.1% (in Spanish) to 34.8% (in Portuguese). Another study [45] reported W2V2 performance of 7.51% WER on the MyST dataset, 3.48% WER on the PFSTAR dataset, and 14.18% WER on the CMU dataset.

Speech recognition research on the Kazakh language [46] using the W2V2 model achieved a WER of 9.8% on a mixture of the ISSAI, KSC1, and language model datasets. The performance of Bengali speech recognition using the W2V2 model on the Bengali CommonVoice dataset also showed a high WER level of 25.24%. The XLSR-53 and XLSR-0.3B models were tested on the Lingala Read Speech Corpus, which contains 4 hours of labeled audio. The XLSR-53

 ISSN 2723-6471


model, which is smaller, achieved a word error rate (WER) of 6.8%, better than the larger XLSR-0.3B model, which achieved 7.0% of WER [47].

W2V2 is an unsupervised learning framework for learning speech representations from raw audio. As illustrated in figure 1, the model uses a convolutional feature encoder to transform the audio signal into a latent representation. 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 [35]. The Transformer architecture in W2V2/XLSR utilizes a self-attention mechanism to capture contextual relationships between speech representations over time. The attention mechanism allows the model to focus on the most relevant parts of the input sequence when generating contextual embeddings. Compared to traditional recurrent architectures, the model is better able to capture long-range temporal dependencies in speech signals thanks to self-attention. Three vectors are used by the attention function to generate the output: query, key, and value. A weighted sum of the values is the outcome of self-attention [48]. Similarly, XLSR learns robust representations across languages and performs better in low-resource languages by using self-attention to process cross-lingual speech data. This mechanism is important in both models because it helps them generalize to large, varied datasets and perform well across different speech processing tasks [49]. The mathematical details of the scaled dot-product attention mechanism used in the Transformer are shown in the eq. (1):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (1)$$

Note:

*Q (Query),* representation of the query (e.g., the search token).

*K (Key),* representation of the key (e.g., the feature term).

*V (Value),* representation of actual information associated with each key.

$d_k$ (Dimension), the dimension of the key vectors and query.

$QK^T$, representation of the similarity score between each query and each key.

Q, K, and V are vector representations in a transformer used in the attention mechanism. Because it processes the same input, Q, K, and V are referred to as the self-attention mechanism. The Query is the token for which information is sought. The Key is a feature or identifier of the token in the query, while the Value is the value that is retrieved if the token is relevant to the query.

W2V2 models are trained in two steps: self-supervised pretraining and supervised fine-tuning. During the self-supervised pretraining phase, the model learns speech representations by predicting masked frames from a large set of unlabeled speech data. Additionally, it learns how to map input frames to discrete speech units and choose masked frames from a set of distractors by using quantized speech data to complete a contrastive task. The model learns to comprehend and encode the meaning of each audio frame based on its context because at this point, it only has access to raw audio and no written data [37].

The model learns from 10 or 100,000 hours of unlabeled speech data during the pre-training stage. This provides a stronger foundation than models based solely on labeled data. Here, the W2V2 model's pre-trained weights provide a useful foundation for supervised training. This paper investigated whether the pre-training phase can benefit from the use of pre-trained weights from a model trained on a related language. By doing this, the model may be able to retain cross-linguistic knowledge and apply it to tasks in the other language. The W2V2 model is language-independent, but instead learns patterns common to all spoken languages. Therefore, pre-training across multiple languages is beneficial even when the target language is not included [37].

Once pre-training is complete, the model uses what it has learned for the target ASR task during fine-tuning. This step is supervised and needs labeled speech data. To find the most likely grapheme sequences, the model uses a final Connectionist Temporal Classification (CTC) layer [50]. By grouping audio frames that share the same output token, CTC reduces the number of frame-level predictions to a shorter output token sequence. ASR performance has been greatly enhanced by the W2V2 model, particularly when labeled data is scarce.

ISSN 2723-6471


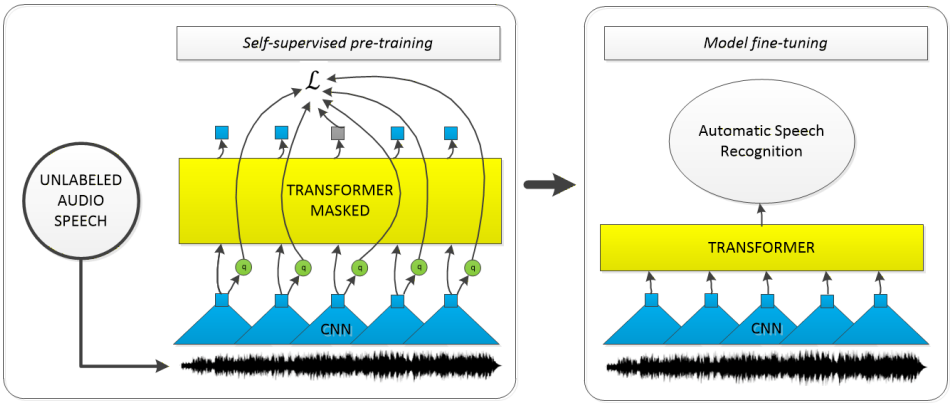


**Figure 1.** The Structure of W2V2/XLSR Model [51]

A novel approach to speech recognition is the W2V2/XLSR model. Its ASR system is built to handle multiple languages at once. Because it was trained using data from several languages, the W2V2/XLSR model is unique. In this manner, the W2V2/XLSR model gains a better understanding of speech in any language by learning patterns that are shared by numerous languages. Languages with limited resources, such as those with little training data, benefit greatly from the W2V2/XLSR model. Because the W2V2/XLSR model makes use of patterns that are common to all languages, it can function well even in cases where one language has limited data. Languages that aren't as well-represented benefit from this [51].

## 3. Methodology

Experiment began using pre-trained W2V2 models. Each model was fine-tuned using the ASR Lampung Nyow v1.0 training dataset, performance was evaluated using a separate testing dataset. Afterwards, all models on a relevant test dataset. The test dataset was used throughout the fine-tuning process, and no speakers overlapped with speakers in the training or development datasets. Figure 2 shows the experiment step-by-step: Data were obtained by independently recording Lampung-language sentences. Once dataset was collected, pre-processing was performed, special characters, were remove, converted all text to lowercase, removed punctuation, defined the vocabulary, and converted audio files to a 16 kHz sample rate with mono channels. The files were then into training and test sets. These steps made sure everything met the W2V2 model's requirements. The next step was to prepare the pre-trained models for processing the Lampung language dataset.

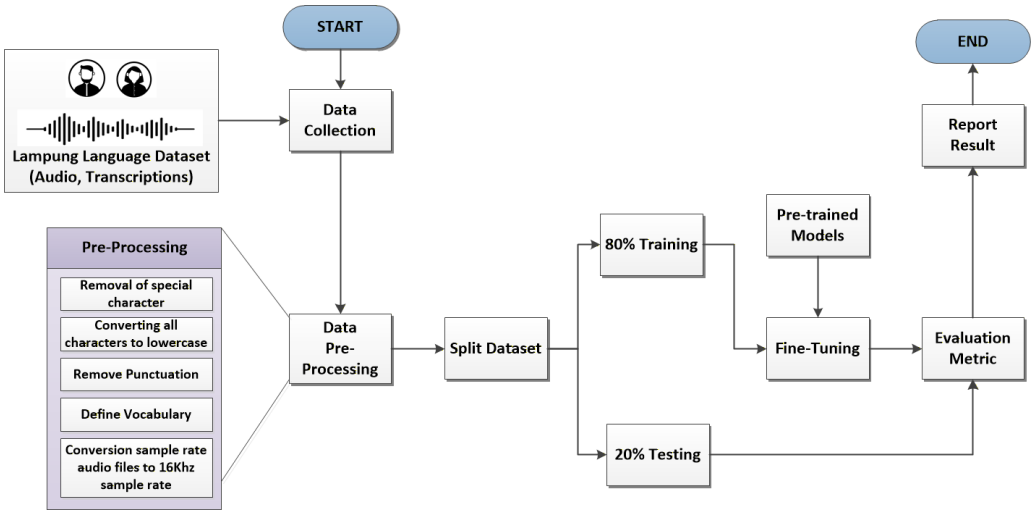


**Figure 2.** Experimental Steps

Once the pre-trained models were completed, the next phase involved fine-tuning the W2V2 models. Several W2V2 models were tested to determine their adaptation to the Lampung language dataset. W2V2 was fine-tuned on LampungNyow v1.0 and updated the model weights to improve Lampung recognition accuracy. To

monitor transcription accuracy, WER was used to periodically assess the model's performance during training. After the objective was achieved, training was stop.

## 3.1. Data Collection

This work presents a speech dataset for the Lampung language that was created independently. A wireless lavalier microphone (UORRIS 3-in-1 3.5 mm) was used to record the audio. The dataset is made up of audio and transcription files for paired speech that are saved in *. xlsx file type. Five male and female speakers provided a total of 12,180 utterances, yielding roughly 6.76 hours of speech data with an average utterance duration of about 2 seconds. Prior to the recording process, informed consent was freely given by each participant. Participants were told that the recordings would only be used for academic study and the advancement of ASR. No personally identifiable information was included in the dataset that was made public in order to protect confidentiality. The audio files were labeled in a methodical manner based on the order of the speakers and matched the corresponding transcription files. The dataset was divided into training and testing subsets using an 80:20 split ratio, consisting of approximately 5.41 hours of training data and 1.35 hours of testing data. As mentioned in table 1 and table 2, the transcript comprises 6,76 hours of audio and 12180 speech files. The original audio recordings in the LampungNyow v1.0 corpus were collected at a sampling rate of 48 kHz, and each file contains mono or stereo channels.

**Table 1.** Lampung Language Distribution Dataset

| Category | Hours | Number of Utterances |
|---|---|---|
| Training | 5,41 | 9,744 |
| Testing | 1,35 | 2,436 |

**Table 2.** Audio Properties

| Audio Properties | Values |
|---|---|
| Format | wav |
| Bitrate | 128Kbps |
| Sample Rate | 48 Khz |
| Channel | Mono/Stereo |

## 3.2. Data Pre-Processing

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. The CTC model used 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.

## 3.3. Pre-Trained Models

The pre-trained models used in this study were selected to investigate the influence of monolingual and multilingual speech representations on automatic speech recognition performance in low-resource Lampung language settings. The selected models differ in terms of training language coverage, dataset scale, and multilingual representation capability. Monolingual models were included to evaluate the effectiveness of language-specific acoustic representations, while multilingual XLSR-based models were selected to explore the benefits of cross-lingual transfer learning for low-

resource ASR tasks. The XLSR-53 model was selected because it was pre-trained on speech data from 53 languages and has demonstrated strong generalization capability in multilingual speech recognition tasks. Its broad multilingual training enables the model to learn generalized acoustic and phonetic representations that can be transferred effectively to low-resource languages such as Lampung. Conversely, language-specific multilingual variants such as XLSR-Indonesian, XLSR-Sundanese, and XLSR-Javanese were selected to investigate whether multilingual pre-training on Indonesian regional languages could provide more relevant acoustic and contextual speech representations for low-resource Lampung ASR tasks. In addition, the multilingual Indonesian-Javanese-Sundanese model was included to evaluate whether combining several Indonesian regional languages could produce richer multilingual speech representations and improve recognition performance for Lampung speech data. This section describes the pre-trained models tested. Two monolingual pre-trained Wav2Vec 2.0 models were used:

W2V2-Base, W2V2-Large, and four model multilingual (XLSR): W2V2-Large-XLSR-53, W2V2-Large-XLSR-Indonesian, W2V2-Large-XLSR-Sundanese, W2V2-Indonesian-Javanese-Sundanese. The models listed will be explained in detail in the following paragraphs.

**W2V2-Base.** The W2V2-Base model is a model pre-trained on only English speech or a monolingual model. The publicly available LibriSpeech dataset was used. It was trained on approximately 960 hours of read English speech.

**W2V2-Large.** The W2V2-Large is a monolingual model pre-trained on English speech. The publicly available LibriSpeech (LS) and LibriVox (LV) datasets was used. It was trained from 960 hours (LS-960) and 60000 hours (LV-60k) of English speech.

**W2V2-Large-XLSR-53.** The W2V2-Large-XLSR-53 is a multilingual pre-trained model that learns cross-lingual speech representation from 53 languages. W2V2-Large-XLSR-53 was trained on 53 languages using the Common Voice, BABEL, and Multilingual LibriSpeech (MLS) datasets.

**W2V2-Large-XLSR-Indonesian.** This is the model for W2V2-Large-XLSR-Indonesian, a fine-tuned W2V2-Large-XLSR-53 model on the Indonesian Common Voice dataset.

**W2V2-Large-XLSR-Sundanese.** This is the model for Wav2Vec2-Large-XLSR-Sundanese, a fine-tuned W2V2-Large-XLSR-53 model on the OpenSLR dataset with identifier SLR44. OpenSLR identifier SLR44 is a high-quality dataset specifically designed for research and development of TTS technology or voice synthesis in Sundanese. The data is split by gender: su_id_female.zip (approximately 861 MB) and su_id_male.zip (approximately 610 MB). Google collected this dataset in collaboration with the Indonesian University of Education.

**W2V2-Large-XLSR-Indonesian-Javanese-Sundanese.** This is the model for W2V2-Indonesian-Javanese-Sundanese, a fine-tuned W2V2-Large-XLSR-53 model on the OpenSLR dataset with identifiers SLR44 for Sundanese, SLR41 for Javanese, and Indonesia Common Voice for Indonesia. OpenSLR identifier SLR44 is a high-quality dataset specifically designed for research and development of TTS technology or voice synthesis in Sundanese. The data is split by gender: su_id_female.zip (approximately 861MB) and su_id_male.zip (approximately 610MB). Google collected this dataset in collaboration with the Universitas Pendidikan Indonesia. Similar to SLR41 is a high-quality dataset specifically designed for research and development of TTS technology or voice synthesis in Javanese. There are two data divisions based on gender: su_id_female.zip (approximately 967MB) and su_id_male.zip (approximately 923MB). Google collected this dataset in collaboration with the Universitas Gadjah Mada.

## 3.4. Fine-Tuning

The experiment started by opening Hugging Face Hub to load a W2V2/XLSR model trained on a large dataset. Hugging Face Hub is a popular source for pre-trained models, especially for ASR tasks. This model had already been trained on large monolingual and multilingual datasets. Next, it was fine-tuned it using Lampung language data from the LampungNyow v1.0 dataset. Experiments were conducted to determine which pre-trained model produced the best transcription results for Lampung. The W2V2/XLSR model had already been trained on massive monolingual, and multilingual datasets, which were loaded from Hugging Face Hub for fine-tuning. Subsequently, it was finetuned on the LampungNyow v1. Python was used on an NVIDIA GPU (8 GB memory) with 32 GB RAM. Training batch size and evaluation batch size were set to 8, under which memory problems were not observed. A 500-step learning rate warmup was used, evaluated and saved the model every 500 steps, and logged every 50. Disk space was spared by keeping only two checkpoints. When a GPU was available, training was performed in mixed-precision (fp16), otherwise 32-bit precision was used. To ensure reproducibility, the random seed was fixed to 42 and uploads to the

 ISSN 2723-6471


Hugging Face Hub and external logging were disabled. Table 3 summarizes the hyperparameter settings and training configurations used during the model fine-tuning experiments.

**Table 3.** Training Configuration and Hyperparameter Settings

| Parameter | Value | Description |
|---|---|---|
| per_device_train_batch_size | 8 | Number of training samples processed per batch on each device. A batch size of 8 helps prevent out-of-memory errors on GPUs with limited memory while maintaining stable training performance. |
| per_device_eval_batch_size | 8 | Number of evaluation samples processed per batch on each device. Using the same batch size as training ensures consistent memory usage during evaluation. |
| evaluation_strategy | "steps" | Specifies that the model is evaluated at fixed step intervals rather than only after each epoch, allowing more frequent monitoring of training performance. |
| num_train_epochs | 20–30 | Number of complete passes through the training dataset. Training for 20–30 epochs allows the model to converge while balancing learning performance and the risk of overfitting. |
| save_steps | 500 | Saves a model checkpoint every 500 training steps, enabling recovery from interruptions and preserving intermediate model states. |
| eval_steps | 500 | Performs model evaluation every 500 training steps to monitor validation performance throughout training. |
| logging_steps | 50 | Logs training metrics such as loss every 50 steps, facilitating close monitoring of the training process. |
| learning_rate | 1e-4 | Initial learning rate used by the optimizer. A value of $1\times10^{-4}$ is commonly adopted for fine-tuning Wav2Vec2 models, providing stable convergence. |
| warmup_steps | 500 | Gradually increases the learning rate during the first 500 steps to stabilize optimization and reduce training instability at the beginning of fine-tuning. |
| save_total_limit | 2 | Limits the number of saved checkpoints to two. Older checkpoints are automatically removed to reduce disk usage. |
| fp16 | torch.cuda.is_available() | Enables mixed-precision (16-bit floating point) training when a CUDA-enabled GPU is available, reducing memory consumption and accelerating training. Otherwise, training uses 32-bit precision. |
| push_to_hub | False | Prevents the trained model from being uploaded to the Hugging Face Hub. The model is stored only in the local output directory. |
| report_to | None | Disables integration with external experiment tracking platforms such as TensorBoard or Weights & Biases. Training logs are displayed only in the console. |
| seed | 42 | Sets a fixed random seed to ensure reproducibility by keeping data shuffling, parameter initialization, and training results consistent across repeated experiments. |

## 3.5. Evaluation

In this study, Lampung ASR was tested on W2V2 models using the WER. WER is a standard measure of how accurate ASR systems are, and it is used to assess system performance across both simple and complex tasks [52]. Instead of CER, the use of WER is sufficient because it allows a strong evaluation of the accuracy of machine-generated transcriptions compared to the reference text as ground truth [53]. This metric compares the system's transcription to the correct text by calculating the ratio of word additions, substitutions, and deletions to the total number of words in the reference. Each type of error addition (extra words), substitution (wrong words), and deletion (missing words), counts equally in the calculation. According to [54], [55], the usual formula for WER is shown in the following eq. (2):

$$\text{Word Error Rate} = \frac{S + D + I}{N} \tag{2}$$

S (Substitutions), number of substitutions incorrectly. D (Deletions), number of deletions from the transcript. I (Insertions), number of insertions word. N (Total Reference Words), number of words in the reference.

The Lampung language text transcription has been aligned with each audio file. As explained in the preprocessing stage, the text transcription was cleaned of punctuation, converted from uppercase to lowercase, and had special characters removed before refinement. The refined models only predict words and cannot predict punctuation or uppercase characters. Therefore, the reference and prediction models produce lowercase words in sentences.

ISSN 2723-6471


## 4. Results and Discussion

### 4.1. Result

In this section, experimental results from several models are presented, and baseline performance results for the Lampung speech recognition development are provided. The performance results are divided into three parts: WER for all models; visualization of training and validation per epoch for the best WER model; and visualization of training and validation per epoch for the best WER model.

Word Error Rate (WER) Comparison: table 4 presents a comparison of the WER values of several pre-trained W2V2 and XLSR models evaluated on the LampungNyow v1.0 dataset. The recognition performance of the W2V2 Base and W2V2 Large models was comparable, with WER values of 36,23% and 36,30%, respectively. The WER values of the multilingual XLSR models, XLSR-53 (33,87%), XLSR-Indonesia (33,70%), and XLSR-Sundanese (33,99%) were lower than those of the baseline models. With a WER of 17,39%, the W2V2-Large-XLSR-Indonesia-Javanese-Sundanese model outperformed the other models assessed in terms of recognition error. These findings imply that in Lampung province, an area with limited resources, multilingual pre-training utilizing a variety of regional Indonesian languages can produce more effective speech representations for automatic speech recognition tasks.

**Table 4.** Differences WER

| No | Models | Language Type | Epoch | WER (%) |
|---|---|---|---|---|
| 1 | W2V2 Base | Monolingual | 30 | 36.23% |
| 2 | W2V2 Large | Monolingual | 30 | 36.30% |
| 3 | W2V2-Large-XLSR-53 | Multilingual | 30 | 33.87% |
| 4 | W2V2-Large-XLSR-Indonesian | Multilingual | 30 | 33.70% |
| 5 | W2V2-Large-XLSR-Sundanese | Multilingual | 30 | 33.99% |
| 6 | W2V2-Large-XLSR-Indonesian-Javanese-Sundanese | Multilingual | 20 | 17.39% |

Training and Validation Loss W2V2-Large-XLSR-Indonesian-Javanese-Sundanese Model: figure 3(a) shows the training and validation loss for the W2V2-Large-XLSR-Indonesian-Javanese-Sundanese model over 20 epochs. Pre-trained on Indonesian-Javanese-Sundanese speech data, the refined W2V2/XLSR model exhibits a comparatively high initial training loss that steadily declines over epochs and converges faster than the other assessed models. This behavior suggests that the multilingual pre-training approach offers useful contextual and acoustic speech representations that help with adaptation when the Lampung dataset is being fine-tuned. Additionally, figure 3(a) demonstrates that during the training process, the evaluation loss exhibits a similar trend to the training loss. The validation loss falls in tandem with the training loss in the early training phases. In later epochs, there is still a discrepancy between training and validation loss.

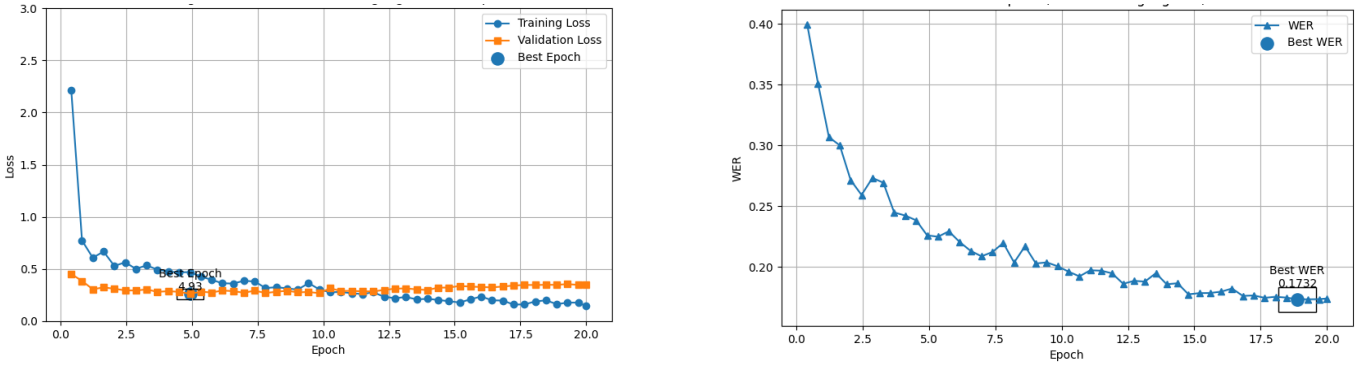


**Figure 3.** (a) Training Loss vs Validation Loss (b) Evaluation WER W2V2-Large-XLSR-Indonesian-Javanese-Sundanese Model

## 4.2. Discussion

The pre-trained Indonesian-Javanese-Sundanese model outperformed all other evaluated models in Lampung speech recognition, according to the experimental results [56], [57]. In comparison to the other models, the training process demonstrated comparatively stable convergence and consistently lower loss values. Furthermore, the multilingual Indonesian-Javanese-Sundanese model performed better on the assessed Lampung dataset, achieving lower WER values than both monolingual and broader multilingual models [56], [57], [58]. The findings imply that for low-resource Lampung ASR tasks, multilingual pre-training utilizing a number of regional Indonesian languages can produce more pertinent acoustic and contextual speech representations. The Indonesian-Javanese-Sundanese model showed more consistent recognition performance and greater adaptation during fine-tuning when compared to models trained on larger multilingual datasets [57]. These results suggest that choosing a suitable multilingual. Limited speaker diversity and recording variations can increase the risk of speaker-specific bias and reduce the model's ability to generalize to unseen accents, speaking styles, and acoustic environments. In the future, these limitations can be minimized, in part, through the use of audio data augmentation. Strategies that can be implemented include adding specific sound effects (eg. abstract noise, light music, animal chirping, etc), pitch shifting, time stretching, and so on. In low-resource ASR settings, small datasets may also increase sensitivity to overfitting and reduce the stability of model evaluation results [58]. Therefore, the reported performance should be interpreted within the scope of the current experimental setting. Nevertheless, this study primarily aims to establish an initial baseline for Lampung ASR and to investigate the effectiveness of multilingual transfer learning under limited-resource conditions.

## 5. Conclusion

In this paper, monolingual and multilingual W2V2 models are compared to evaluate ASR performance on the Lampung dataset. Based on the experimental results, the W2V2-Large-XLSR-Indonesian-Javanese-Sundanese model achieved the best performance among all evaluated models. This model obtained the lowest WER of 17.39%, outperforming the monolingual W2V2-Large-XLSR-Indonesian model (33.70%) and the W2V2-Large-XLSR-Sundanese model (33.99%). The lower WER achieved by the W2V2-Large-XLSR-Indonesian-Javanese-Sundanese model indicates that multilingual pre-training using several Indonesian regional languages can potentially provide more suitable speech representations for the Lampung ASR task compared to models trained using only a single additional language. On the assessed dataset, however, more comprehensive multilingual models like XLSR-53 and W2V2 Base/Large obtained comparatively higher WER values. Overall, the findings show that ASR performance in low-resource Lampung speech recognition tasks can be enhanced by multilingual transfer learning. Additionally, this study offers a starting point for further research on Lampung ASR and the creation of pre-trained speech recognition models for regional languages that are underrepresented. In order to increase recognition accuracy and model robustness, future work will concentrate on growing the Lampung speech corpus, refining it using more speech data, and data augmentation.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization: H.K., A.J., F.R.L., and W.; Methodology: H.K., A.J., F.R.L., and W.; Software: H.K.; Validation: A.J., F.R.L., and W.; Formal Analysis: H.K., A.J.; Investigation: H.K., A.J.; Resources: H.K.; Data Curation: H.K., A.J; Writing Original Draft Preparation: H.K., A.J., F.R.L., and W.; Writing Review and Editing: H.K., A.J.; Visualization: H.K., A.J.; All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The corresponding author is available upon request for the data used in this study.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 6.4. Institutional Review Board Statement

Not applicable.

## 6.5. Informed Consent Statement

Not applicable.

## 6.6. Declaration of Competing Interest

All authors declare that there are no conflict of interest.

**References**

[1] Y. Agiomyrgiannakis and Y. Stylianou, "The harmonic model codec (HMC) framework for voIP," in *Proc. Interspeech 2007*, vol. 2007, no. August, pp. 1681–1684, 2007, doi: 10.21437/Interspeech.2007-473.

[2] A.-L. Georgescu, A. Pappalardo, H. Cucu, and M. Blott, "Performance vs. hardware requirements in state-of-the-art automatic speech recognition," *J Audio Speech Music Proc*, vol. 2021, no. 1, pp. 1–30, 2021, doi: 10.1186/s13636-021-00217-4.

[3] S. Karpagavalli and E. Chandra, "A Review on Automatic Speech Recognition Architecture and Approaches," *IJSIP*, vol. 9, no. 4, pp. 393–404, 2016, doi: 10.14257/ijsip.2016.9.4.34.

[4] Y. He, "Streaming End-to-end Speech Recognition for Mobile Devices," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2019, no. May, pp. 6381–6385, 2019, doi: 10.1109/ICASSP.2019.8682336.

[5] M. Schuster, "Speech Recognition for Mobile Devices at Google," in *PRICAI 2010: Trends in Artificial Intelligence, vol. 2010, no. August,* pp. 8–10, 2010, doi: 10.1007/978-3-642-15246-7_3.

[6] H. H. O. Nasereddin and A. A. R. Omari, "Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation," *in 2017 Computing Conference*, vol. 2017, no. Jul, pp. 200–207, Jul. 2017, doi: 10.1109/SAI.2017.8252104.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc*., vol. 29, no. Oct, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

[8] D. O'Shaughnessy, "Trends and developments in automatic speech recognition research," *Computer Speech & Language*, vol. 83, no. Jan, p. 101538, 2024, doi: 10.1016/j.csl.2023.101538.

[9] X. Wei, C. Dong, and Y. Xu, "The Mechanical and System Design of Finger Training Rehabilitation Device Based on Speech Recognition," *Journal of Applied Data Sciences*, vol. 3, no. 2, pp. 60–65, 2022, doi: 10.47738/jads.v3i2.58.

[10] H. I. Pratiwi, I. H. Kartowisastro, B. Soewito, and W. Budiharto, "Dispute on Security Framework Model of MFCC Mixed Methods in Speech Recognition System," *Journal of Applied Data Sciences*, vol. 6, no. 3, pp. 1542–1550, 2025, doi: 10.47738/jads.v6i3.689.

[11] Statistics Indonesia (BPS), "Mid-year population" Accessed: Feb. 14, 2026. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun--ribu-jiwa-.html

[12] A. F. Aji, "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),* vol. 2022, no. May, pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.

[13] N. W. Putri, "Pergeseran Bahasa Daerah Lampung pada Masyarakat Kota Bandar Lampung," *PRASASTI: Journal of Linguistics,* vol. 3, no. 1, pp. 83–97, 2018, doi: 10.20961/prasasti.v3i1.16550.

[14] H. Nasution, R. Rahayu, E. Wibowo, D. M. Harum, and F. Moses, "Distribution of languages in Lampung Province," Lampung Province Language Office, 2008.

[15] A. E. Sanusi, S. Zamzanah, M. Widodo, I. Sunarti, and S. Samhati, "Function words in the Abung dialect of the Lampung language," Indonesian Center for Language Development and Cultivation, 1997. [Online]. Available: https://repositori.kemendikdasmen.go.id/1987/1/Kata%20Tugas%20Bahasa%20Lampung%20Dialek%20Abung%201997.pdf

[16] N. E. Rusminto, F. Ariyani, A. B. Setiyadi, and G. E. Putrawan, "Local language vs. national language: The Lampung language maintenance in the Indonesian context," *Kervan. International Journal of African and Asian Studies*, vol. 25, no. 1, 2021, doi: 10.13135/1825-263X/5787.

[17] C. Moseley, Atlas of the World's Languages in Danger. UNESCO Digital Library, 2012. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000187026

[18] A. Raharjo and A. Zahra, "Javanese and Sundanese speech recognition using Whisper," *Computer Science and Information Technologies*, vol. 6, no. 3, pp. 253–261, 2025, doi: 10.11591/csit.v6i3.p253-261.

[19] A. Ramadhan, A. Junaidi, Aristoteles, F. R. Lumbanraja, and A. Faisol, "Implementation of MFCC Features Extraction of Recurrent Neural Network and Bidirectional LSTM for Speech to Text Transcription: A Case Study on the Lampung Language Dialect Api," in P*roceedings of the 5th International Conference on Applied Sciences, Mathematics, and Informatics (ICASMI)*, vol. 2025, no. May, pp. 139–150, 2025, doi: https//doi.org/10.2991/978-94-6463-730-4_10.

[20] M. Zainudin, A. Junaidi, F. R. Lumbanraja, Aristoteles, and Tristiyanto, "Perceptual Linear Prediction Features Extraction for Lampung Voice to Text Transcription (Case Study: Lampung Pepadun Dialect A)," in *Proceedings of the 5th International Conference on Applied Sciences, Mathematics, and Informatics (ICASMI)*, vol. 2025, no. May, pp. 108–117, 2025, doi: https//doi.org/10.2991/978-94-6463-730-4_10.

[21] M. Khudhair and A. Talib, "Improving Low Resources Arabic Speech Recognition using Data Augmentation," in *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, vol. 2022, no. Nov, pp. 60–65, 2022, doi: 10.1109/CSCTIT56299.2022.10145613.

[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012, doi: 10.1109/TASL.2011.2134090.

[23] N. Kaur and P. Singh, "Modelling of Speech Parameters of Punjabi by Pre-trained Deep Neural Network Using Stacked Denoising Autoencoders," *ACM Trans. Asian Low-Resour. Lang. Inf. Process*., vol. 22, no. 3, pp. 1–17, 2023, doi: 10.1145/3568308.

[24] T. G. Fantaye, J. Yu, and T. T. Hailu, "Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition," *Computers*, vol. 9, no. 2, pp. 1–27, 2020, doi: 10.3390/computers9020036.

[25] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2013, no. May, pp. 8614–8618, 2013, doi: 10.1109/ICASSP.2013.6639347.

[26] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2013, no. May, pp. 6645–6649, 2013, doi: 10.1109/ICASSP.2013.6638947.

[27] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2016, no. Mar, pp. 4945–4949, 2016, doi: 10.1109/ICASSP.2016.7472618.

[28] E. Battenberg, "Exploring neural transducers for end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 2017, no. Dec, pp. 206–213, 2017, doi: 10.1109/ASRU.2017.8268937.

[29] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A Comparison of Sequence-to-Sequence Models for Speech Recognition," in *Interspeech 2017*, vol. 2017, no. August, pp. 939–943, 2017, doi: 10.21437/Interspeech.2017-233.

[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, vol. 2, no. Dec, pp. 3104–3112, Dec. 2014.

[31] E. Trentin and M. Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1, pp. 91–126, 2001, doi: 10.1016/S0925-2312(00)00308-8.

[32] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "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)*, vol. 2016, no. Mar, pp. 4960–4964, 2016, doi: 10.1109/ICASSP.2016.7472621.

[33] A. Gulati, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech 2020*, vol. 2020, no. Oct, pp. 5036–5040, 2020, doi: 10.21437/Interspeech.2020-3015.

[34] A. Hannun, "Deep Speech: Scaling up end-to-end speech recognition," vol. 2014, no. Dec, pp. 1–12, 2014, doi: 10.48550/arXiv.1412.5567.

[35] A. Baevski, H.Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Proceedings of the 34th International Conference on Neural Information Processing Systems, vol. 2020, no. Dec, pp. 12449–12460, 2020.

[36] W.-H. Tsai, P. L. Thi, T.-C. Tai, C.-L. Huang, and J.-C. Wang, "Low-Resource Speech Recognition Based on Transfer Learning," *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, vol. 2022, no. Dec, pp. 145–149, 2022, doi: 10.1109/RIVF55975.2022.10013881.

[37] J. Lehečka, J. V. Psutka, and J. Psutka, "Transfer Learning of Transformer-Based Speech Recognition Models from Czech to Slovak," in *Text, Speech, and Dialogue*, vol. 2023, no. August, pp. 328–338, 2023, doi: 10.1007/978-3-031-40498-6_29.

[38] A. R. Bharadwaj, K. L. Srina, K. Snuhith, R. Bhavani, and M. A. Jabbar, "Telugu Language Low-resource ASR Fine-Tuning with Wav2Vec2-XLS-R-300 and Interface Development," *2025 Global Conference in Emerging Technology (GINOTECH)*, vol. 2025, no. May, pp. 1–6, 2025, doi: 10.1109/GINOTECH63460.2025.11077079.

[39] A. Cryssiover and A. Zahra, "Speech recognition model design for Sundanese language using WAV2VEC 2.0," *Int J Speech Technol*, vol. 27, no. 1, pp. 171–177, 2024, doi: 10.1007/s10772-023-10066-5.

[40] P. Arisaputra, A. T. Handoyo, and A. Zahra, "XLS-R Deep Learning Model for Multilingual ASR on Low-Resource Languages: Indonesian, Javanese, and Sundanese," *ICIC Express Letters, Part B: Applications*, vol. 15, no. 6, pp. 551–559, 2024, doi: 10.24507/icicelb.15.06.551.

[41] P. Arisaputra and A. Zahra, "Indonesian Automatic Speech Recognition with XLSR-53," *ISI*, vol. 27, no. 6, pp. 973–982, Dec. 2022, doi: 10.18280/isi.270614.

[42] A. Bawitlung, S. K. Dash, and R. M. Pattanayak, "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.*, vol. 24, no. 7, p. 72:1-72:15, Jul. 2025, doi: 10.1145/3746063.

[43] L. R. Stefanel Gris, E. Casanova, F. S. De Oliveira, A. Da Silva Soares, and A. Candido Junior, "Brazilian Portuguese Speech Recognition Using Wav2vec 2.0," in *Computational Processing of the Portuguese Language*, vol. 13208, no. March, pp. 333–343, 2022, doi: 10.1007/978-3-030-98305-5_31.

[44] J. C. Vásquez-Correa and A. Álvarez Muniain, "Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper," *Sensors*, vol. 23, no. 4, p. 1843, Jan. 2023, doi: 10.3390/s23041843.

[45] A. Barcovschi, R. Jain, and P. Corcoran, "A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, vol. 2023, no. Oct, pp. 42–47, 2023, doi: 10.1109/SpeD59241.2023.10314867.

[46] K. Bauyrzhan, M. Madina, and O. Assel, "Fine-Tuning the Wav2vec2 Model for Kazakh Speech: A Study on a Limited Corpus," *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, vol. 2023, no. May, pp. 124–128, 2023, doi: 10.1109/SIST58284.2023.10223504.

[47] U. Kimanuka, C. wa Maina, and O. Büyük, "Speech recognition datasets for low-resource Congolese languages," Data in Brief, vol. 52, no. Feb, p. 109796, 2024, doi: 10.1016/j.dib.2023.109796.

[48] A. Vaswani, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 2017, no. Dec, pp. 6000–6010, 2017.

[49] L. Chen, M. Asgari, and H. H. Dodge, "Optimize Wav2vec2s Architecture for Small Training Set Through Analyzing its Pre-Trained Models Attention Pattern," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2022, no. May, pp. 7112–7116, 2022, doi: 10.1109/ICASSP43922.2022.9747831.

[50] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd international conference on Machine learning, vol. 2006*, no. June, pp. 369–376, 2006, doi: 10.1145/1143844.1143891.

[51] A. Babu, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," *Interspeech 2022,* vol. 2022, no. Sept, pp. 2278–2282, 2022, doi: 10.21437/Interspeech.2022-143.

ISSN 2723-6471


[52] L. Rabiner, S. Levinson, A. Rosenberg, and J. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 27, no. 4, pp. 336–349, 1979, doi: 10.1109/TASSP.1979.1163259.

[53] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1, pp. 19–28, 2002, doi: 10.1016/S0167-6393(01)00041-3.

[54] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised Automatic Speech Recognition: A review," *Speech Communication*, vol. 139, no. April, pp. 76–91, 2022, doi: 10.1016/j.specom.2022.02.005.

[55] S. Jothilakshmi and V. N. Gudivada, "Chapter 10 - Large Scale Data Enabled Evolution of Spoken Language Research and Applications," *Handbook of Statistics*, vol. 35, no. August, pp. 301–340, 2016, doi: 10.1016/bs.host.2016.07.005.

[56] V. Pratap, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Interspeech 2020*, vol. 2020, no. Oct, pp. 4751–4755, 2020, doi: 10.21437/Interspeech.2020-2831.

[57] A. Tjandra, "Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2023, no. June, pp. 1–5, 2023, doi: 10.1109/ICASSP49357.2023.10094667.

[58] R. Amooie, W. De Vries, Y. Hao, J. Dijkstra, M. Coler, and M. Wieling, "Enhancing Standard and Dialectal Frisian ASR: Multilingual Fine-tuning and Language Identification for Improved Low-resource Performance," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2025, no. April, pp. 1–5, 2025, doi: 10.1109/ICASSP49660.2025.10889692.