# UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

**Colegio de Ciencias e Ingenierías**

# Fine-tuning Wav2Vec2 for Automatic Speech Recognition in Kichwa

.

**Christian Eduardo Santamaria López**

**Ingeniería en Ciencias de la Computación**

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 19 de diciembre de 2024

# UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

## Colegio de Ciencias e Ingenierías

## HOJA DE CALIFICACIÓN
## DE TRABAJO DE FIN DE CARRERA

# Fine-tuning Wav2Vec2 for Automatic Speech Recognition in Kichwa

# Christian Eduardo Santamaria López

| | |
|---|---|
| **Nombre del profesor, Título académico** | **Felipe Grijalva, PhD.** |

Quito, 19 de diciembre de 2024

# © DERECHOS DE AUTOR

Nombres y apellidos: Christian Eduardo Santamaria López

Código: 00215605

Cédula de identidad: 1600578551

Lugar y fecha: Quito, 19 de diciembre de 2024

# RESUMEN

En los últimos años, los avances en inteligencia artificial han impulsado el desarrollo de modelos de procesamiento del lenguaje natural y reconocimiento automático del habla (ASR) para idiomas mayoritarios, generando preocupaciones sobre la marginación de lenguas ancestrales y subrepresentadas. Este trabajo propone el ajuste fino del modelo Wav2Vec 2.0, desarrollado por Meta AI, para ASR en Kichwa, un idioma hablado en los Andes ecuatorianos. Se utilizaron dos conjuntos de datos, que suman 8 horas de grabaciones de audio segmentadas en fragmentos de 1.5 a 5 segundos con transcripciones detalladas utilizando el software ELAN, para el entrenamiento. El proceso de ajuste fino empleó el algoritmo CTC. Después de varios experimentos, realizamos una prueba de Wilcoxon de dos colas que indicó que no hubo una mejora significativa con SpecAugment. El mejor modelo, entrenado sin aumentación de datos, logró resultados prometedores en el conjunto de prueba, con una Tasa de Error de Palabra (WER) de 0.262, una Tasa de Error de Caracteres (CER) de 0.120 y una Tasa de Error de Coincidencia (MER) de 0.401. Estos resultados destacan el potencial de los modelos ASR ajustados para generalizar de manera efectiva en contextos de bajos recursos, incluso con disponibilidad limitada de datos, ofreciendo un camino hacia una mayor inclusividad lingüística en inteligencia artificial.

**Palabras clave:** Kichwa, Reconocimiento Automático del Habla, Ajuste Fino, Clasificación Temporal Conexionista, Aprendizaje Profundo, Audio.

# ABSTRACT

In recent years, advancements in artificial intelligence have driven the development of natural language processing and automatic speech recognition (ASR) models for majority languages, raising concerns about the marginalization of ancestral and underrepresented languages. This work proposes the fine-tuning of the Wav2Vec 2.0 model, developed by Meta AI, for ASR in Kichwa, a language spoken in the Ecuadorian Andes. Two datasets, totaling 8 hours of audio recordings segmented into 1.5 to 5-second clips with detailed transcriptions using the ELAN software, were used for training. The fine-tuning process employed the CTC algorithm. After several experiments, we performed a two-tailed Wilcoxon test that indicated no significant improvement with SpecAugment. The best model, trained without data augmentation, achieved promising results on the test set, with a Word Error Rate (WER) of 0.262, a Character Error Rate (CER) of 0.120, and a Match Error Rate (MER) of 0.401. These results highlight the potential of fine-tuned ASR models to effectively generalize in low-resource settings, even with limited data availability, offering a pathway toward greater linguistic inclusivity in artificial intelligence.

**Key words:** Kichwa, Automatic Speech Recognition, Fine-Tuning, Connectionist Temporal Classification, Deep Learning, Audio.

# CONTENT INDEX

# TABLE INDEX

## ÍNDICE DE FIGURAS

# INTRODUCTION

While major languages such as English, Mandarin, and Spanish have benefited from vast datasets, resources, and advances in large languages models, many low-resource languages remain underrepresented. Those often lack the extensive corpus of transcriptions, audio recordings, and linguistic tools necessary to effectively train state-of-the-art models. This disparity places minority languages, such as Kichwa, at risk of further marginalization and eventual extinction.

In Ecuador, the situation of language shift has become increasingly evident over the past several decades. Kichwa, spoken by indigenous communities in the Ecuadorian Andes, has gradually lost its prominence due to various socio-political, economic, and cultural factors. The expansion of Spanish in public education, media, and administrative domains has led to a decline in Kichwa use, particularly among younger generations. According to data from the National Institute of Statistics and Census of Ecuador (INEC), the percentage of Ecuadorians who speak Kichwa as their main language was only 5% in 2010 [1]. This decline reflects a broader trend of Indigenous tongues being displaced by dominant national ones, a phenomenon exacerbated by migration to urban areas where Spanish serves as the primary means of communication.

This problematic is not merely a linguistic concern, but also a cultural one, as Kichwa embodies centuries of tradition, oral history, and identity for its speakers. The weakening of generational language transmission has further accelerated this decline. In many indigenous households, Spanish has become the preferred language due to its perceived economic and social advantages, leaving Kichwa relegated to ceremonial or informal uses. The lack of digital tools, linguistic technologies, and educational resources in Kichwa further contributes to its marginalization, limiting opportunities for the language to adapt and thrive in modern contexts.

Addressing this pressing issue requires innovative solutions to bridge the technological and data gaps faced by low-resource languages. Automatic Speech Recognition (ASR), which has advanced considerably in recent years, offers a potential tool to revitalize Kichwa by creating opportunities for its integration into modern technologies. ASR systems, which convert speech into text, have evolved significantly since their early reliance on Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) [2][3]. These traditional approaches, while effective, required complex design processes and large volumes of labeled data to model acoustic, pronunciation, and language components. However, the emergence of end-to-end models has simplified this process by integrating all these components into a single architecture optimized jointly. These models have demonstrated superior performance, particularly for major languages where labeled data is abundant.

Fortunately, advancements in deep neural networks (DNNs) and, more recently, transformers have provided more flexible and accurate ways to model the acoustic features of speech. The self-attention mechanisms introduced in Vaswani et al.'s Attention Is All You Need [4] allow models to assign relevance to different parts of an input sequence, enabling them to learn patterns more robustly and dynamically. This has led to models that approach speech learning in ways that resemble human intuition.

In particular, the Wav2Vec 2.0 model, proposed by Baevski et al. [5], introduces a self-supervised pre-training approach using Contrastive Predictive Coding (CPC) to learn speech representations from unlabeled audio. The model’s two-stage process begins with extracting acoustic features using a convolutional network, followed by a transformer trained to predict masked segments of the discrete representations. This contrastive loss enables the model to capture temporal relationships and general linguistic patterns, which can then be fine-tuned for downstream tasks with limited labeled data.

Wav2Vec 2.0 has demonstrated its effectiveness in drastically reducing the need for labeled data. For instance, with only 10 minutes of transcriptions, it achieved a Word Error Rate (WER) of 5.7/10.1 on the noisy/clean Librispeech test sets [6]. This makes Wav2Vec 2.0 an ideal candidate for addressing the challenges of low-resource languages like Kichwa. By leveraging self-supervised learning, it becomes possible to develop speech recognition tools that do not rely on large annotated datasets, thus offering a practical solution for languages at risk of marginalization.

This work focuses on adapting Wav2Vec 2.0 for the ASR task in Kichwa, using approximately 4 hours of audio recordings segmented into short clips of 1.5 to 5 seconds, and 1 hour for validation. The fine-tuning process employs Connectionist Temporal Classification (CTC) [7], a loss function well-suited for sequential problems like speech recognition. By demonstrating the effectiveness of Wav2Vec 2.0 in a low-resource setting, this study not only contributes to the technological development of Kichwa ASR systems but also supports the broader efforts to preserve and revitalize low resource languages.

The following section will present works with similar objectives.

## RELATED WORKS

The use of pretrained models has proven to be an effective strategy for improving ASR performance in low-resource languages. In the study Transfer Ability of Monolingual Wav2vec2.0 for Low-resource Speech Recognition [8], the model's ability to efficiently transfer learning from a single language to various low-resource languages is evaluated. This work highlights the importance of considering linguistic variability within a multilingual context, revealing that the model can adapt well in such scenarios. However, it is noted that the effectiveness of this transfer largely depends on the linguistic proximity between the languages involved, suggesting that the multilingual approach has certain limitations when working with highly divergent languages.

On the other hand, in Applying Wav2Vec2.0 to Speech Recognition in Various Low-resource Languages [9], a different approach is taken by evaluating Wav2Vec in several languages independently. Through experiments with databases of 15 hours each in Mandarin, English, Arabic, Japanese, German, and Spanish, the study reaffirms the robustness and effectiveness of the model when faced with data scarcity. Additionally, this article introduces a comparison between two fine-tuning techniques: CTC (Connectionist Temporal Classification) and the LM-decoder approach, concluding that the latter offers better performance due to the support of a language model; nevertheless, CTC stands out for achieving comparable results at a lower computational cost.

Finally, [10] analyzes the applicability of the proposed model to Bengali speech recognition, a language that, although having a considerable number of speakers, has not been widely considered for ASR tasks. Using the "Bengali Common Voice Speech" dataset, the study demonstrated a WER of 0.2524 for this non-Western language with significant dialectal variation. Additionally, this work emphasizes that, although Wav2Vec 2.0's self-supervised pre-training reduces data barriers, the fine-tuning process remains crucial for achieving optimal results.

# METHODOLOGY

## MODEL SELECTION

For this ASR task, facebook/wav2vec2-xls-r-300m was selected, a pretrained model from Facebook AI's XLS-R series. XLS-R (Cross-Lingual Speech Representations) is a large-scale multilingual model built for speech representation learning, and it is particularly suited for low-resource languages due to its extensive pretraining across a vast array of languages. The model is based on the Wav2Vec 2.0 architecture, which is optimized to extract robust features from raw audio using unlabeled data. This makes it an ideal choice for fine-tuning on tasks like Automatic Speech Recognition (ASR), especially in scenarios with limited labeled data.

The architecture of Wav2Vec 2.0 consists of a convolutional feature extractor that processes raw audio signals, generating latent representations, Z, which are then passed through a Transformer model. By masking parts of the input sequence during training and using Contrastive Predictive Coding (CPC), the model learns to predict the missing segments based on the surrounding context. This approach forces the model to learn highly contextualized representations that generalize well across unseen data, which is particularly beneficial in speech recognition tasks where labeled data is scarce. The block diagram is at Fig. 1

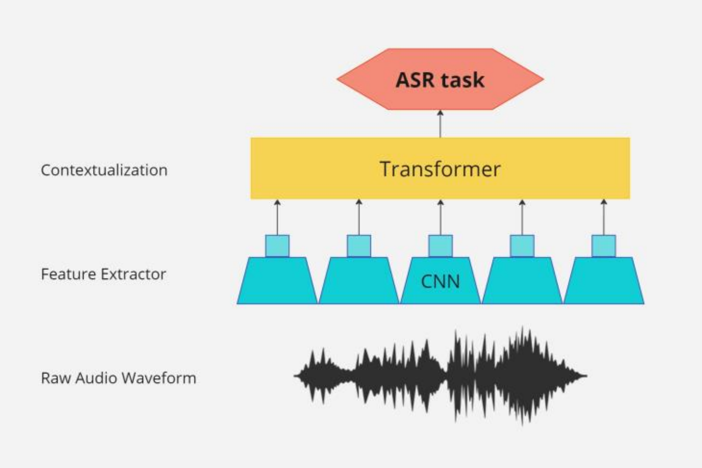

**Fig 1. Block Diagram of Wav2Vec2.0 model architecture based on [5]**

One of the key reasons for choosing the facebook/wav2vec2-xls-r-300m model is its multilingual pretraining on 436,000 hours of publicly available speech data in 128 languages. This extensive pretraining equips the model with the ability to capture diverse acoustic and linguistic patterns across a wide range of languages, making it highly effective for low-resource languages like Kichwa. This multilingual knowledge helps the model extract meaningful audio features even from languages it wasn't explicitly trained on, thereby improving its generalization capabilities.

In addition to its multilingual pretraining, the model's 300 million parameters give it the capacity to retain a wealth of information about speech patterns, phonetics, and temporal relationships in audio. This makes it particularly well-suited for fine-tuning with Connectionist Temporal Classification (CTC), where it can handle the variability in speech length and pauses efficiently.

Moreover, XLS-R has been shown to achieve state-of-the-art performance across a wide range of speech processing tasks, including automatic speech recognition, speech translation, and language identification. In particular, XLS-R lowers word error rates (WER) by 20-33% on multilingual benchmarks like BABEL and CommonVoice, and sets new standards in language identification on VoxLingua107 [11] This strong track record, combined with its ability to work well with low-resource languages, makes facebook/wav2vec2-xls-r-300m a powerful choice for Kichwa ASR fine-tuning, even with a relatively small labeled dataset.

## DATASETS

For this experiment, two datasets from different sources were used: one for training and validation, and the other for testing. The primary data for the first dataset was obtained from the KILLKAN repository [12], from which around 5 hours of master audio recordings and their transcriptions were extracted. On the other hand, the test dataset was created with the help of a paid volunteer who read excerpts from two Kichwa texts: Ecuador Watapak Mamakamachiy (Constitution of Ecuador 2008, official translation into Kichwa) and Taruka: La Venada (stories from Kichwa oral tradition). In total, approximately 3 hours of audio were collected, along with their corresponding master transcriptions. This process resulted in the raw datasets.

With those, the next step was to join the audio files with their transcriptions using the ELAN software, developed by the Max Planck Institute for Psycholinguistics. This step was only necessary for the dataset we created, as KILLKAN had already labeled their data using the same software. Once both datasets were prepared, the audio was divided into shorter fragments, ranging from 1.5 to 5 seconds, to accommodate the available computational resources and the model's capabilities.

In this way, we obtained a database ready for use. The full diagram is at Fig. 2.

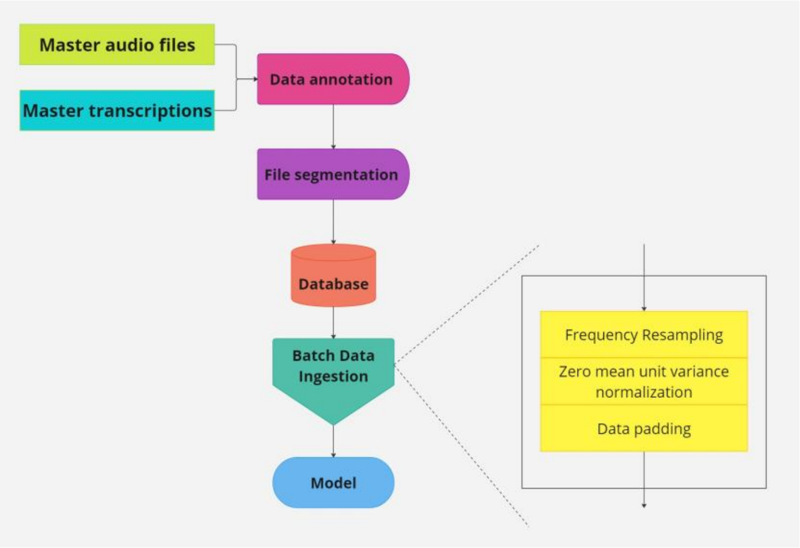

**Fig 2. Block diagram of data flow**

## DATA AUGMENTATION

Given that the database is relatively small, with only 4 hours of labeled data for training, the use of Spec Augment was considered—a data augmentation technique recognized for its effectiveness in ASR tasks [13] To implement it, an augmentation policy is applied that focuses on three key aspects: time warping, time masking, and frequency masking.

Time warping involves stretching or compressing the spectrogram along the time axis to simulate natural variations in speaking speed. By distorting the time axis around a randomly selected point, variations in speech tempo are generated, helping the model better generalize to different speaking styles without altering the frequency content.

Time masking covers random sections of the time axis in the spectrogram, effectively silencing parts of the input. By randomly masking consecutive time steps, the model is forced to rely on the broader context of the input rather than specific time slices. This technique simulates missing or occluded audio, making the model more resilient to temporal noise or dropouts in the signal.

Finally, frequency masking targets the frequency axis by masking random frequency bands. This simulates the loss of certain frequency components, such as pitch variations between speakers or noise affecting certain frequencies. By covering random frequency ranges, the model learns to generalize across different vocal ranges and background noise, focusing more on patterns that span across multiple frequencies rather than overfitting to specific frequency information.

## EXPERIMENTAL SETUP

For the implementation, PyTorch, Lightning and Hugging Faces Transformers libraries were primarily used for model management, data handling, and the integration of data augmentation technique.

The hyper-parameters selected to fine-tune the Wav2Vec2 model are as follows: For feature extraction, the model uses 7 convolutional layers, each with 512 filters. The strides are set to [5, 2, 2, 2, 2, 2, 2], meaning the first layer applies a larger stride to quickly reduce the input's temporal resolution, while subsequent layers apply smaller strides for more refined feature extraction. The kernel sizes are [10, 3, 3, 3, 3, 2, 2], allowing the first layer to capture broad, general audio patterns, while the smaller kernels in later layers progressively focus on finer, more detailed features. In the Transformer model, the architecture consists of 768-dimensional hidden states, with 12 hidden layers and 12 attention heads, giving the model robust capacity to learn complex patterns and capture long-range dependencies in speech. Regularization is achieved through hidden dropout, attention dropout, and activation dropout, all with probabilities set to 0.1, as well as a layer drop of 0.1, which randomly drops entire layers during training to prevent overfitting and improve generalization. The learning rate is set at 1e-4, a standard value for fine-tuning large models, ensuring stable weight updates, and a batch size of 4 is chosen to balance computational efficiency and model performance. When SpecAugment is applied we use a 0.05 probability of masking out 10 time steps, simulating temporal distortions and further boosting the model's ability to generalize.

## EVALUATION METRICS

The evaluation metrics used to assess the performance of the Wav2Vec2 model fine-tuned for Kichwa Automatic Speech Recognition (ASR) are critical for measuring the quality of the model's transcriptions. The primary metrics are Word Error Rate (WER), Character Error Rate (CER), and Match Error Rate (MER). These metrics quantify how well the model predicts words, characters, and overall matches between predicted transcriptions and ground truth. In this section, we provide an explanation and formula for each metric.

### Word Error Rate (WER)

The Word Error Rate (WER) is a common metric for evaluating the performance of speech recognition systems. It measures the percentage of incorrectly predicted words by comparing the predicted transcription to the ground truth. The WER is calculated as:

```
\begin{equation}
  \text{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}
\end{equation}
```

Where:

```
\begin{itemize}
  \item \(S\) is the number of substitutions (incorrect words),
  \item \(D\) is the number of deletions (missing words),
  \item \(I\) is the number of insertions (extra words),
  \item \(N\) is the total number of words in the ground truth transcription,
```

\item \(C\) is the number of correct words.

\end{itemize}

WER ranges from 0 to 1, where a lower WER indicates better transcription accuracy.

### \Character Error Rate (CER)

The Character Error Rate (CER) is similar to WER but operates at the character level; it measures the percentage of incorrectly predicted characters and is computed as:

\begin{equation}

\text{CER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}

\end{equation}

Where:

\begin{itemize}

\item \(S\) is the number of character substitutions,

\item \(D\) is the number of character deletions,

\item \(I\) is the number of character insertions,

\item \(N\) is the total number of characters in the ground truth transcription,

\item \(C\) is the number of correct characters.

\end{itemize}

As with WER, lower CER values indicate more accurate character-level transcription.

### Match Error Rate (MER)

The Match Error Rate (MER) evaluates the transcription quality by measuring how well the predicted and ground truth transcriptions match in a more holistic way. Unlike WER and CER, which focus on substitutions, deletions, and insertions, MER is often computed by examining exact matches or mismatches in the sequence, and it is typically expressed as a percentage.

The formula for MER is:

\begin{equation}
    \text{MER} = \frac{S + D + I}{N + I} = \frac{S + D + I}{S + D + C + I}
\end{equation}

Where:
\begin{itemize}
    \item \(S\) is the number of substitutions,
    \item \(D\) is the number of deletions,
    \item \(I\) is the number of insertions,
    \item \(N\) is the number of words in the reference,
    \item \(C\) is the number of correct words.
\end{itemize}

A lower MER reflects better agreement between the predicted and actual sequences, indicating a more accurate model.

These metrics collectively provide a comprehensive assessment of the Wav2Vec2 model's ability to recognize both individual characters and entire words, as well as its overall accuracy in matching predictions to ground truth transcriptions.

## EXPERIMENTS AND RESULTS

TABLE 1. COMPARISON OF MODEL PERFORMANCE METRICS WITH AND WITHOUT DATA AUGMENTATION USING SPECAUGMENT

| Technique | LOSS | WER | CER | MER |
|---|---|---|---|---|
| **Data-augmentation** | 0.0884 | 0.1184 | 0.0591 | 0.2554 |
| **No data-augmentation** | 0.0896 | 0.1224 | 0.0595 | 0.2510 |

To evaluate the impact of the data augmentation technique on the model’s training performance, multiple comparative experiments were conducted, applying the technique or not, while keeping the hyperparameters and other configurations constant. The results were analyzed using the two-tailed Wilcoxon statistical test with a 95% confidence level, applied to each metric and the loss function. The p-values obtained were 0.922, 0.770, 0.888, and 0.695 for LOSS, WER, CER, and MER, respectively. These values indicate that there is no sufficient statistical evidence to claim that the data augmentation technique used, in this case, SpecAugment, significantly improved the model's performance during training. Table 1 summarizes the averages obtained in the experiments, differentiating between those conducted with and without data augmentation.

Among all the experiments, the model with the lowest WER value was selected as the best-trained model. Fig. 3 shows its loss and WER curves achieved in the validation dataset. Finally, when evaluating its performance on the test set, the following metrics were obtained: CER = 0.120, MER = 0.401, and WER = 0.262.

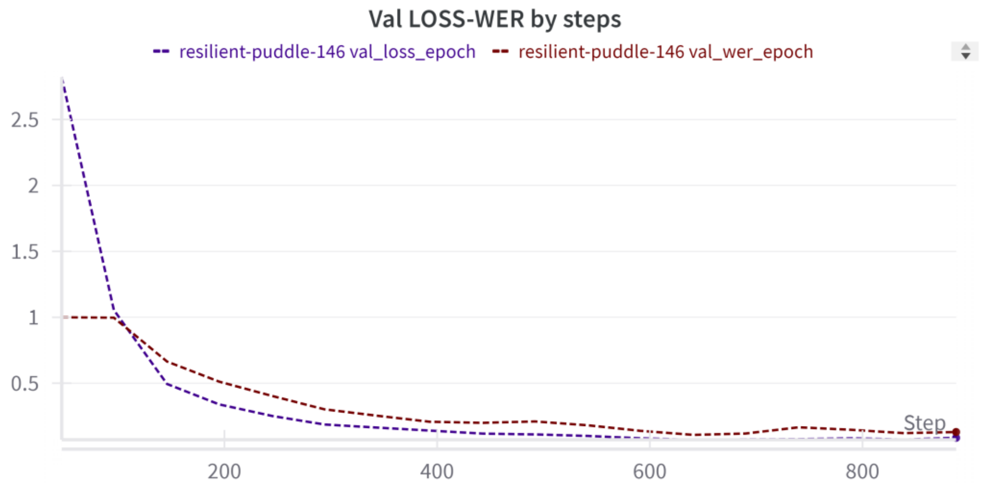

**Fig 3. Validation loss function and WER curves of best fine-tuned model.**

# CONCLUSIONS

This work presents an effective approach for fine-tuning the pre-trained Wav2Vec2.0 model for ASR tasks in Kichwa, a language that, due to advancements in globalization and technology, has been pushed to the background, putting its preservation at risk. The model's ability to generalize and extract features from audio is impressive, as with only 5 hours of transcribed recordings, outstanding metrics were achieved, such as a WER of 0.262. This result demonstrates the potential of deep learning models to make technology more inclusive from a linguistic perspective, without requiring large amounts of data. In this way, new opportunities arise for creating more accessible and useful tools for a greater number of languages and communities.

As a future direction, it would be ideal to work with a higher-quality dataset, including recordings of everyday conversations on more common topics, such as daily life or social issues, rather than focusing solely on stories or political speeches. Additionally, having more speakers (participants) in the recordings would contribute to a better model generalization. This approach would also help improve transcription accuracy, particularly by avoiding “stumbles” in the speech of volunteers. In this case, stumbles occurred when the volunteer used filler words during the reading, which were captured in the audio but not in the transcriptions, potentially distorting the metrics slightly by not considering them. A better dataset and a greater variety of situations would help minimize these effects and achieve even more accurate results.

## REFERENCES

[1] Language data for Ecuador – Translators without Borders — translatorswithoutborders.org. https://translatorswithoutborders.org/language-data-for-ecuador. [Accessed 18-12-2024].

[2] K-F Lee, H-W Hon, and Raj Reddy. An overview of the sphinx speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45, 1990.

[3] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[4] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[6] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[7] Julian Salazar, Katrin Kirchhoff, and Zhiheng Huang. Self-attention networks for connectionist temporal classification in speech recognition. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7115–7119. IEEE, 2019.

[8] Cheng Yi, Jianrong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. Transfer ability of multilingual wav2vec 2.0 for low-resource speech recognition. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–6. IEEE, 2021.

[9] C Yi, J Wang, N Cheng, S Zhou, and B Xu. Applying wav2vec2.0 to speech recognition in various low-resource languages. arxiv 2020. *arXiv preprint arXiv:2012.12121*.

[10] HAZ Shagir, Khondker Salman Sayeed, and Tanjeem Azwad Zaman. Applying wav2vec2 for speech recognition on bengali common voices dataset. arXiv preprint arXiv:2209.06581, 2022.

[11] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296, 2021*.

[12] Chiriro Taguchi. Killkan. https://github.com/ctaguchi/killkan, 2024.

[13] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.