Speech Recognition for Inventory Management in Small Businesses

Bruno Tiglla-Arrascue, Junior Huerta-Pahuacho and Luis Canaval
Universidad Peruana de Ciencias Aplicadas, Lima, Peru

Keywords: Speech-to-Text, Machine Learning, Deep Learning.

Abstract: In recent years, we have seen an increase in independent businesses working primarily focused on online sales, where they offer products through ads and manage the business with electronic tools. This could leave behind some traditional businesses, especially those that are managed by a single family, where the adaption of new technologies is slower than new business. That's why we want to give them a tool that it's easy to control, a virtual assistant where they can manage the inventory even if they don't know about databases. For this work, we propose to create a speech-to-text platform with machine learning so those users who have difficulties adapting to these new tools can use their voice to command the database and have first contact with these new technologies. Through a fine-tuning process to a pre-trained speech-to-text model in Spanish, we managed to obtain a percentage error result lower than the model used, this being 14.3%, this means that our model has a better accuracy in the context of a Peruvian convenience store.

1 INTRODUCTION

After the lockdown, there was a considerable increase in the number of small businesses, more people wanted to offer their products whether virtually or in physical shops and the commercial sector started to grow. As a result, the supply and the demand were gradually increasing as the population began to leave their homes.

In recent years, technology has played a crucial role in the business sector. And now, most retail businesses turn to digital solutions to improve their sales and bring new tools to employees and customers.

However, there is still a sector where tech adoption seems to have stopped. In these types of businesses, the implementation of technological solutions tends to be a challenge, as many of these merchants tend to have a more traditional approach (Peng and Bao, 2023).

These traditional business practices are mainly based on manual statistics, inefficient analysis, and error-prone decision-making. Additionally, some internal data of these businesses tends to be fragmented and decentralized, making it difficult to compile it into a database.

Although the implementation of digital transformation has driven the development of technology solutions focused on streamlining processes, the development of artificial intelligence (AI) solutions stands out.

Ghobakhloo et al. (Ghobakhloo et al., 2023) claim that in the context of digital transformation, AI has become a key term. Furthermore, it points out that AI is revolutionary due to its unique features, such as the ability to simulate human intelligence and interact in real-time with people through voice recognition.

These features enable it to adapt to new business circumstances and predict potential outcomes, as mentioned in (Peng and Bao, 2023). In Peru, many types of traditional businesses can be found in every neighborhood, with the most common being those that sell high-turnover products. Some of these establishments still follow traditional management methods and are often run by individuals who are not familiar with virtual tools.

However, due to the necessity brought about by digital solutions such as virtual wallets or online commerce through social networks, they have had to learn and use these technological tools, including the use of mobile devices and specialized software.

In these businesses where they handle a lot of different products, it's common to face problems related to inventory disorganization, which can result n in delays in the registration and checkout processes of products.

This is often the result of manual control work, in addition to being slow, it tends to fail, and increasing the risk of information loss by relying on physi-

Tiglla-Arrascue, B., Huerta-Pahuacho, J. and Canaval, L.
Speech Recognition for Inventory Management in Small Businesses.
DOI: 10.5220/0012813600003764
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 21st International Conference on Smart Business Technologies (ICSBT 2024)*, pages 95-102
ISBN: 978-989-758-710-8; ISSN: 2184-772X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

cal reports. A technological solution for this type of problem can be found in the use of databases. By maintaining a virtual inventory through platforms and thanks to their accessibility, it becomes easier to keep track of product movements.

However, there is little information about how these tools could be applied to their businesses. Many of these inventory management solutions for traditional businesses are not focused on those who are just learning to handle technological devices or are unfamiliar with database terms.

To solve this problem there are different methods to facilitate the use of these tools for new users. For example, an application where the user can manage the inventory. Our goal is to provide support to the technological transformation movement by facilitating work performance in those businesses.

Likewise, with the high use and development of artificial intelligence in recent years, a solution that involves this technology could facilitate the use of technological tools within a store, especially to users that, especially for those users who do not handle electronic devices. The objective of this project is to provide essential support to small business owners by offering them tools to create and utilize a database.

The idea is to provide an easy-to-use solution. For example, a mobile application where the user can control the inventory of their business, but with a speech-to-text function with which the user can manage the application more easily.

By listing their products, the goal is to simplify the creation, basic organization, and management of a virtual inventory for their business. Through a web platform, the intention is to support the user by providing them with more effective control over their inventory, enabling them to make more appropriate and accurate decisions when acquiring new products and have knowledge about their inventory.

Now, we will review some related works on speech-to-text solutions and solutions applied to a non-common technological area in Section 2. Then, we mention the terms and tools that we used in developing our proposal in Section 3. Likewise, we review how our proposal using the speech-to-text function is performing in Section 4. In addition, the setup, experiments, and results. And finally the conclusions and discussions of our proposal in Section 5.

2 RELATED WORK

In this section, we will briefly discuss the implementation of speech-to-text in video games, and how the implementation of a technology solution as speech-to-text can improve the efficiency of a specific area.

In (Aguirre-Peralta et al., 2023), the authors focus on the use of convolutional neural networks for speech-to-text recognition as the control of a video game The objective was to develop a medical tool for individuals with upper limb motor disabilities. The paper presents a convolutional neural network architecture focused on speech-to-text recognition, and three turn-based mini-games were developed within the video game to process the data provided by the convolutional neural network. Additionally, it provides an analysis of the most promising results that demonstrate the project's feasibility. It explains the definitions of technologies related to convolutional neural networks, the chosen architecture for the solution, and terms such as NLP, CNN, speech-to-text recognition, and gamification. Likewise, it details the experiments, the types of data they used, and their results.

In (Waqar et al., 2021), the authors propose a real-time voice command recognition system using Convolutional Neural Networks (CNN) to control the Snake game. The authors prepared a dataset for voice commands: up, down, left, and right, for training, validation, and testing. They proposed an optimal voice command recognition system based on MFCC (Mel Frequency Cepstral Coefficients) and CNN to recognize the four voice commands. The proposed algorithm achieved a high recognition accuracy of 96.5% and successfully detected all four commands. Finally, the proposed algorithm was integrated into a Python-based Snake game.

In (Mallikarjuna Rao et al., 2022), the authors focus on creating a chatbot for information queries in a university context. It outlines the issue where students and parents need to navigate the university's website or make phone inquiries to obtain information, and how a chatbot can address this problem by providing automated responses. Different types of chatbots, such as rule-based and machine learning-based ones, are also mentioned, along with a discussion of the advantages and limitations of each approach. The text provides details about the structure of the proposed chatbot and its implementation using AIML and natural language processing.

2.1 Machine Learning

In recent years, technology has experienced exponential advancement.

In particular, chatbots and other types of artificial intelligence solutions, such as Machine Learning algorithms and process automation, can significantly reduce administrative burden (Androutsopoulou et al., 2019).

Their ability to enable machines to learn and adapt from data has revolutionized a wide range of applications in various areas.

Machine Learning is transforming our approach to complex problems and opening up new possibilities, such as empowering machines to interact with users through voice.

Here, we will explore some fundamentals and applications for the field of speech-to-text recognition.

2.2 Convolutional Neural Networks

In the field of computer vision and deep learning, Convolutional Neural Networks (CNN) have emerged as a fundamental architecture, inspired by the natural mechanism of visual perception in living beings.

In the context of these networks, their potential lies in the ability to extract and process local information through convolutions applied to input data, using sets of filters with a fixed size (Apicella et al., 2023).

2.3 Speech-to-Text

Peralta et al. (Aguirre-Peralta et al., 2023) mention that speech-to-text is the capacity for the machine to recognize human speech and convert it into a written text.

Also, they mention that it can be useful for people who don't know how to work with technological tools.

2.4 Wav2Vec 2.0

Wav2Vec 2.0 is a model for Automatic Speech Recognition (ASR) with self-supervised training.

This model contains different elements where the audio will pass for different processes to covert in a textual representation.

First, it receives a raw audio where the model use an encoder based on a Convolutional Neural Network (CNN), that extracts the features from the audio, this network is using to represent audio to be more compact and significant

After passing through the encoder, the audio features are quantized using multiple codebooks.

This means that it reduces the amount of data necessary to represent a signal while maintaining an acceptable level of fidelity.

The quantization is done using a function called "Gumbel softmax". This quantized representation of the audio is used in the decoding process.

3 METHOD

In this section, we will explain some preliminary phases for the development of the project, where the speech-to-text recognition model that we plan to implement will be mentioned.

Also, we will explain the main workflow of the proposed application.

In addition, we mention the data used for retraining the model and as well its preparation to receive this new information.

Finally, we will explain the expected final product.

3.1 The ML Model Structure

As we previously mentioned, the speech-to-text recognition process will be explained using the Wav2Vec2 model for ASR.

In this model, there are several processes in which the audio will go through different parts of the model to become a textual representation of the audio.

It begins by receiving raw audio signals where the model uses an encoder based on a convolutional neural network (CNN), which will be responsible for extracting audio features, which will be used to represent the audio in a more compact and meaningful way. (Baevski et al., 2020) The model is composed of a multi-layer feature encoder with convolutional characteristics. This model learns the basic units of speech to perform a self-supervised task, enabling learning from unlabeled training data to facilitate speech recognition systems for multiple languages. (Baevski et al., 2020)

$$P_{g,v} = \frac{\exp(lg_v + n_v)/\tau}{\sum_{k=1}^{V} \exp(lg_k + n_k)/\tau}$$

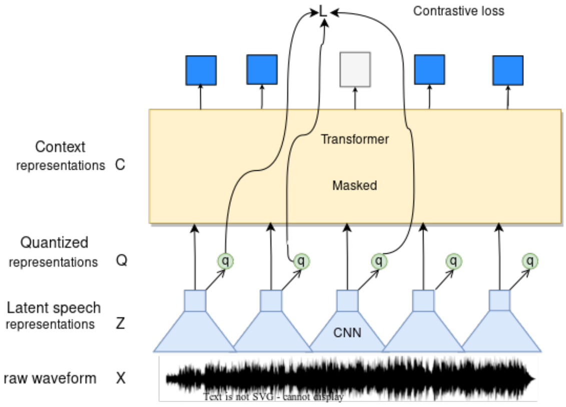
L
Contrastive loss
Context representations C
Transformer
Masked
Quantized representations Q
q
Latent speech representations Z
CNN
raw waveform X
Text is not SVG - cannot display

Figure 1: The framework which jointly learns contextualized speech representations and an inventory of discretized speech units. (Baevski et al., 2020).

After the quantized representation, the audio features are passed through a ”Transformer” neural network. This network captures relationships between the audio features and generates a textual representation.

This representation consists of sequences of code words that correspond to the sounds of the audio. Finally, the model employs decoders situated in the last output layer that perform the final classifications.

This layer is regarded as the ”Decoder,” as it decodes the context representations into a text transcription. The model is depicted in Figure 1.

3.2 Workflow of the Project

Thanks to the previously mentioned model, we can transform audio into text. This will serve as the required input to ensure that the proposed platform can follow the intended workflow to assist in inventory management.

However, the model in question is trained in a general mode with a wide range of Spanish words, and its accuracy can vary depending on how users pronounce them. Therefore, the decision was made to retrain the model within the context of inventory management.

To undertake this task, new training data is needed. As a result, new audio data will be collected and divided into three groups.

The platform operates by allowing the user to access various functionalities using their mobile phone's microphone.

The platform is to be hosted at first stance in our local computers but for better research, the plan is to deploy it as an AWS Cloud service, where both the web platform's logic and the pre-trained model are running.

The information captured by the front end will be sent to the back end, where the model will interpret these audio signals and automatically perform learned actions based on the audio input.

The logic architecture is depicted in Figure 2

3.3 Data Pre-Processing

To retrain the model, the proposal involves using a new set of audio recordings focused on the inventory management context.

From the collected audio data, it was planned to divide them into three groups of Spanish words:

The first group comprises the keywords for the platform's workflow:

1. “agregar”, “buscar”, “actuak1 es milizar”,” vender”,”desactivar”,”generar”,”producto”, ”informe”.

The second group consists of product names that we will focus on for the application's recognition, which are:

1. Sodas: ”Inca Kola,” ”Coca Cola,” and ”Fanta.”
2. Cookies: ”San Jorge,” ”Margarita,” and ”Rellenita.”
3. Water: ”San Luis,” ”San Mateo,” and ”Cielo.”

The third group of data includes forty combinations of phrases used in the process of adding a product. These phrases follow the structure of ”add”: ”Key word phrase” + ”Quantity” + ”Product name” + ”Specifications” + ”Price.”

1. “agregar catorce rellenitas de cien gramos costo dos soles”
2. “vender veintitres inca kola de cuatrocientos mililitros costo dos soles cincuenta”
3. “compre setenta y nueve santos mateos de quinientos litros salio dieciocho soles con noventa”

From this last group of audio data, random but equivalent samples are obtained from the proposed fifty participants.

3.4 Data Preparation

To improve the model's accuracy in the proposed context, it is suggested to retrain the model with the newly collected data.

This way, the received audio inputs will have improved accuracy in recognizing the words within the transcribed phrase.

This precise transcription is essential for the subsequent stages of the platform's workflow.

Furthermore, since the model is already partially trained to recognize a wide range of Spanish words, the plan for increasing the data collection was to encompass variations in background noise, such as noise like a conversation in the background.

This involves a process of Data Augmentation for our data collection.

It allows us to obtain more data from a smaller number of users and optimize the model for specific situations, types of noises, or user speech speed.

For this purpose, a sample of 50 users was used in data collection, this group was chosen indiscriminately and is made up of Spanish speakers.

Each user-contributed with 37 audio tracks (8 from group 1, 9 from group 2, and 20 from group 3).

After compilation, we had a total of 1850 audios for retraining, and by adding to these new audios a background sound.

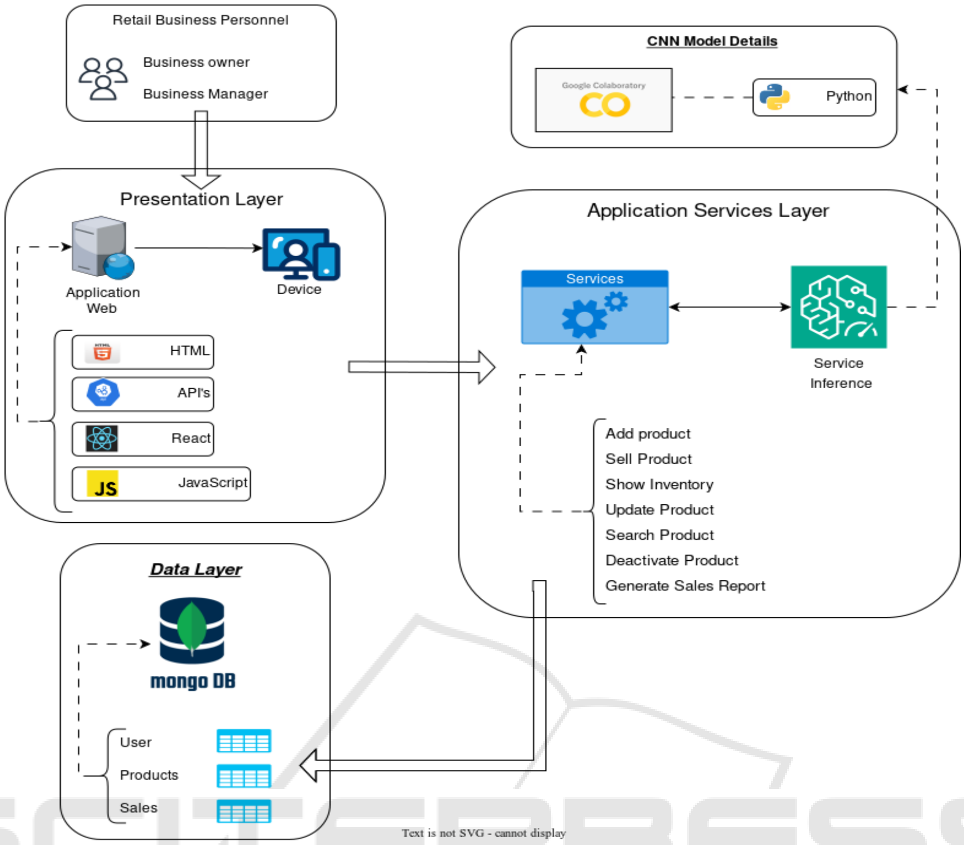

Figure 2: Logic Architecture Diagram of the solution.

We have more than 3700. However, we filtered some audio that were low quality and would not support training. And in the same way, we split a data set for testing.

3.5 Fine-Tuning Process Preparation

For this work we are using the Lightning Flash Library, with this we can use a function that help us to retrain the new model.

The structure is depicted in Figure 3.

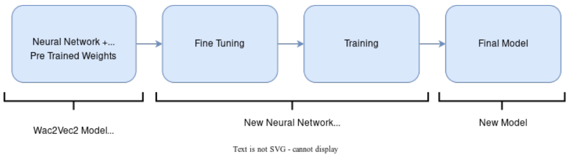

Figure 3: The structure of the pre-model learning framework of wav2vec2 applying the library with the freeze strategy.

The 'Backbone' hosts the Wav2Vec 2.0 model, which was trained on its respective datasets.

This model encompasses a neural network designed for Spanish speech recognition. The neural network has already acquired general features from the original dataset and will serve as the foundation for the new model.

On the other hand, the 'Head' constitutes another neural network, typically smaller in size, trained on the proposed dataset to learn mapping the general features extracted by the 'Backbone'.

The model is depicted in Figure 4.

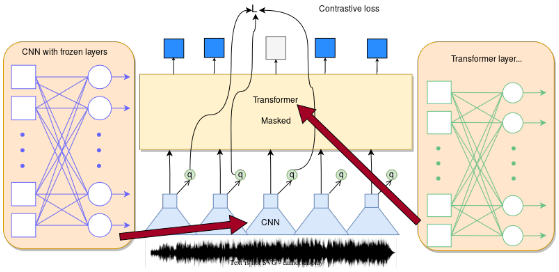

Figure 4: Fine-tuning applied to the pre-trained Wav2Vec2 model.

The "Freeze" strategy in Lightning Flash is primarily employed during the fine-tuning process of models. For the Wav2Vec 2.0 model, this strategy is used to "freeze" the pre-trained model's weights during training.

This entails that the weights of the neural network in the ”Backbone,” which are the convolutional neural networks, are not updated during the backpropagation process.

This is done to maintain the learned characteristics of the original dataset.

The aim is to enable the model to adapt to new data without completely forgetting the general features.

This is particularly useful when the new dataset is relatively small compared to the original one.

3.6 Final Output

Finally, after fine-tuning in Google Colab, we hosted the model file in Google Drive, with this model we created an API with Flask.

The frontend, developed in React, will receive user audio and pass it to the API of the model to convert the audio into JSON with the applied modifications and interpretations.

Subsequently, in the backend, developed in Node.js, it will carry out the necessary actions to perform the CRUD operations on the platform in Node.js.

4 EXPERIMENTS

In this section, we discuss about the experiments done for the project, the setup used and results obtained during the fine-tuning and test.

4.1 Experimental Protocol

In this part, we discuss about the setup where all the experiments were performed.

The re-training of the model used the Google Colab Pro. This premium version offers a Tesla V100-SXM2 with 16160MiB, using the CUDA version 12.0 and Python version 3.10.12. The setup of Colab of the fine-tuning is hosted in Google Drive. [1].

Also, we used 5% of the general audio data to serve as test data. This entire process is designed to improve the model's accuracy.

4.2 Training the Model

Potential scenarios were devised where the user inputs several extended sentences into the platform.

Moreover, the system was trained with potential words to act as specific commands for a particular platform feature.

All this data was compiled in ”ogg” format. Although the model accepts different audio formats, the documentation(Baevski et al., 2020) suggests that the audio to train the model has two characteristics, ”.wav” format and that it contains a sampling frequency of 16khz, so the audios had to be modified, even knowing that making this modification could generate a small amount of information loss.

However, since the data collected is not that extensive, that loss would not be as critical.

Taking the above into account, the data set was created.

They are made up of two columns: ”file”, which represents the location, stored in our Google Drive for easier access from the colab, and ”text”, which denotes the phrase transcribed from the audio.

We started by creating an instance of a pre-trained model for Spanish language recognition: MODEL ID = ”jonatasgrosman/wav2vec2-xls-r-1b-spanish”.

Utilizing the Lightning Flash framework provided a highly useful tool for fine-tuning, as elucidated in the previous section.

The training employed features such as the use of GPU if available, with a choice to conduct training over 10 epochs and to adopt the ”freeze” strategy, as detailed in the preceding section.

4.3 Results of the Training

Following the training phase, we conducted an evaluation using various official versions of the Wav2vec2 model. Additionally, models pre-trained by other users from the Hugging Face community were selected, which built upon Facebook's models and improved them using their respective datasets.

In this instance, four additional models were chosen:

Official Facebook Models:

- facebook/wav2vec2-large-xlsr-53-spanish (Model 2)
- facebook/wav2vec2-base-10k-voxpopuli-ft-es (Model 3)

Community Retrained Models:

- jonatasgrosman/wav2vec2-large-xlsr-53-spanish (Model 4)
- jonatasgrosman/wav2vec2-xls-r-1b-spanish (Model 5)

For this validation, an instance of each previously visited model was called to test its accuracy.

[1] Google Drive

In this test, an audio sample of 50 possible commands that the models could receive was taken.

To equalize conditions, the audio samples were processed exactly as each Wac2Vec2.0 architecture requires (Baevski et al., 2020).

This means that it was ensured that the audio had a sampling frequency of 16 kHz and a format acceptable for the models.

For speech recognition models, exist specific measures that provide a more detailed perspective on the performance of these models.

In particular, in this evaluation, metrics such as Word Error Rate (WER), Message Error Rate (MER), and Word Information Loss (WIL) have been employed to comprehensively assess the quality of the obtained transcriptions (Errattahi et al., 2015).

- WER (Word Error Rate): It is the most popular metric for ASR evaluation, measuring the percentage of incorrect words (Substitutions (S), Insertions (I), Deletions (D)) relative to the total number of words (Errattahi et al., 2015).

  $$\text{WER} = \frac{S+D+I}{N_1} = \frac{S+D+I}{H+S+D} \qquad (1)$$

  Where:
  I = total number of insertions,
  D = total number of deletions,
  S = total number of substitutions,
  H = total number of visits,
  N_1 = total number of input words.

- MER (Message Error Rate): Similar to WER, but it evaluates errors at the level of complete messages or phrases rather than individual words.
- WIL (Word Information Loss): Assesses the amount of information lost or preserved in the transcription or translation process, measuring the loss of information in terms of omitted, added, or changed words compared to a reference transcription or translation. Additionally, it serves as an approximation measure of RIL. However, unlike RIL (Relative Information Lost), WIL is easy to apply because it relies solely on counts of HSDI and is expressed as (Errattahi et al., 2015).

  $$\text{WIL} = 1 - \frac{H^2}{(H+S+D)(H+S+I)} \qquad (2)$$

For the following Table 1, decimal numbers are shown, these numbers represent the percentage of error in each category.

This means that the number that is closest to zero has fewer errors in the speech-to-text representation.

Table 1: Models Evaluation.

Model	WER	MER	WIL
Our Model	**0.143**	**0.143**	0.012
Model 2	0.447	0.447	0.126
Model 3	0.534	0.516	0.043
Model 4	0.218	0.218	**0.008**
Model 5	0.256	0.256	0.029

4.4 Discussion

In this subsection, we discussed the results obtained in the previous section. While the obtained results are favorable compared to other models. It is important to note that this model is still in its early stages. This means that the data we used for this first training was done for a small number of products and with that data we cannot satisfy the entire market. However, it has demonstrated a good ability to understand phrases in the context presented in this research.

4.5 Fine-Tuning Process

In the development of this project, various methods for training a model were researched.

We chose to employ a slightly more flexible method, as mentioned earlier in Section 4. However, this method has some limitations, such as the limited customization in hyperparameter settings, both in terms of freezing or unfreezing layers and their weight alterations.

Nevertheless, it serves as a good entry point into the field of machine learning.

The interface provided by Lightning Flash succinctly encapsulates the basic knowledge of how to train a model.

4.6 Comparison with Other Models

As we can see in Table 1, models 2 and 3 have higher Word Rate Errors, which means the transcription of these models can fail and bring a text that is not properly translated.

However, these models made by Facebook are the base of models 4 and 5, these models from the Hugging Face community are pre-trained models that use the original Facebook wac2vec2 model and are retrained with more data.

Looking at Table 1, these two models have a lesser World Rate Error than models 2 and 3.

However, the data that were used for the fine-tuning process were audios with generic voice liens in Spanish. And when we want to translate audio that is in the context of a Peruvian convenience store, the models can't recognize some keywords, like the name of the products, and it can return a text that is translated better than the other original models but it will have some errors, and required an extra process to transform that failed transcription, into the interpretation of the correct phrase.

And for our model, it looks like it has better results than the others, however, this does not mean that the model has a better precision than the others. A fact is that our model has better results due to the fine-tuning process for that specific context. As we can see, we retrained the model with our data to have that accuracy in the context of a Peruvian convenience store.

All other models were trained with general data for the Spanish language. Another fact is that the model, depending on the audio quality and the specific pronunciation of the user, cannot return the phrase that is 100% interpreted correctly.

5 CONCLUSIONS

For this project, fine-tuning was performed on a pre-trained speech-to-text model to improve the model in a specific context, the phrases used in convenience stores in Peru.

The goal was to implement this improved model on an online platform, allowing users to interact with the platform using their voice, our motivation for making this project was to help the owners of these traditional businesses adapt to technological solutions and be the first step to adapt the business to this new tools.

Through data collection, we achieved positive results, as depicted in the experiments section, specifically in the 'Training the Model' subsection. Our model outperformed the base models and pre-trained models in Spanish. Through this training using the proposed data, we achieved a Word Error Rate of 14.3%, demonstrating its effectiveness compared to other models for this specific context.

For future works, we aim to expand the product list to a more comprehensive one commonly used in these stores. In this initial training, we conducted as the project's prototype, we utilized a reduced list of products.

The objective is to deliver a high-quality online platform and help the independent owners of these traditional businesses with tasks that are more effective if you work with technological tools.

Additionally, another goal is to try to get better accuracy and attempt to reduce the Word Error Rate (WER) for the products already trained, and get the same WER for newly introduced products. Furthermore, other kinds of models might improve our metrics (Leon-Urbano and Ugarte, 2020; Ysique-Neciosup et al., 2022; Rodríguez et al., 2021).

REFERENCES

Aguirre-Peralta, J., Rivas-Zavala, M., and Ugarte, W. (2023). Speech to text recognition for videogame controlling with convolutional neural networks. In *ICPRAM*, pages 948–955. SCITEPRESS.

Androutsopoulou, A., Karacapilidis, N. I., Loukis, E. N., and Charalabidis, Y. (2019). Transforming the communication between citizens and government through ai-guided chatbots. *Gov. Inf. Q.*, 36(2):358–367.

Apicella, A., Isgrò, F., Pollastro, A., and Prevete, R. (2023). Adaptive filters in graph convolutional neural networks. *Pattern Recognit.*, 144:109867.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Errattahi, R., Hannani, A. E., and Ouahmane, H. (2015). Automatic speech recognition errors detection and correction: A review. In *ICNLSP*, volume 128 of *Procedia Computer Science*, pages 32–37. Elsevier.

Ghobakhloo, M., Asadi, S., Iranmanesh, M., Foroughi, B., Mubarak, M., and Yadegaridehkordi, E. (2023). Intelligent automation implementation and corporate sustainability performance: The enabling role of corporate social responsibility strategy. *Technology in Society*, 74:102301.

Leon-Urbano, C. and Ugarte, W. (2020). End-to-end electroencephalogram (EEG) motor imagery classification with long short-term. In *SSCI*, pages 2814–2820. IEEE.

Mallikarjuna Rao, G., Tripurari, V. S., Ayila, E., Kummam, R., and Peetala, D. S. (2022). Smart-bot assistant for college information system. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pages 693–697.

Peng, J. and Bao, L. (2023). Construction of enterprise business management analysis framework based on big data technology. *Heliyon*, 9(6):e17144.

Rodríguez, M., Pastor, F., and Ugarte, W. (2021). Classification of fruit ripeness grades using a convolutional neural network and data augmentation. In *FRUCT*, pages 374–380. IEEE.

Waqar, D. M., Gunawan, T. S., Kartiwi, M., and Ahmad, R. (2021). Real-time voice-controlled game interaction using convolutional neural networks. In *2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pages 76–81.

Ysique-Neciosup, J., Chavez, N. M., and Ugarte, W. (2022). Deephistory: A convolutional neural network for automatic animation of museum paintings. *Comput. Animat. Virtual Worlds*, 33(5).