Speech Recognition System Using Wav2vec Model (Punjabi Language)

By

Capt Kashif Yaseen

Capt Adeel Zafar

Maj Awais Ali

Supervised by:

Dr. Shibli Nisar

Submitted to the faculty of Department of Electrical Engineering,
Military College of Signals, National University of Sciences and Technology, Islamabad,
in partial fulfillment for the requirements of B.E Degree in Electrical Engineering.

June 2023

In the name of ALLAH, the Most benevolent, the Most Courteous

CERTIFICATE OF CORRECTNESS AND APPROVAL

This is to officially stated that the thesis work contained in this report

"Speech Recognition System Using Wav2vec Model

(Punjabi Language)"

is carried out by

Capt Kashif Yasin

Capt Adeel Zafar

Maj Awais Ali

under my supervision and that in my judgement, it is fully ample, in scope and excellence, for the degree of Bachelor of Electrical Engineering in Military College of Signals, National University of Sciences and Technology (NUST), Islamabad.

Approved by

Supervisor

Dr. Shibli Nisar

Department of EE, MCS

Date: ___________

DECLARATION OF ORIGINALITY

We hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification in either this institute or anywhere else.

ACKNOWLEDGEMENTS

Allah Subhan’Wa’Tala is the sole guidance in all domains.

Our parents, colleagues and most of all supervisor, Dr. Shibli Nisar without your guidance.

The group members, who through all adversities worked steadfastly.

Plagiarism Certificate (Turnitin Report)

This thesis has 13 % similarity index. Turnitin report endorsed by Supervisor is attached.

Capt Kashif Yasin
359302

Capt Adeel Zafar
359300

Maj Awais Ali
359293

Signature of Supervisor

ABSTRACT

Speech Recognition presents natural phenomena for the communication among man and machine. The purpose of Speech Recognition speech system is to convert the sequence of sound units in the form of text description. Technology for understanding spoken words by computers has improved a lot recently. But for languages like Punjabi, it's still hard for computers to understand speech well. The complexity of Punjabi phonology, compounded by variations in accent and pronunciation, poses substantial challenges for automatic speech recognition systems. As a result, the need for a robust Punjabi sound recognition system has become increasingly evident. Our project aims to solve this problem by using a special computer model called Wav2Vec. We train this model to understand Punjabi sounds better, so it can transcribe speech more accurately. So far, no work has been done in the field of Punjabi speech recognition system. Our approach involves pre-processing Punjabi audio data, training the Wav2Vec model, and fine-tuning it using transfer learning techniques. The final output is presented through a user-friendly Graphical User Interface (GUI), illustrating the outcomes of our Punjabi sound recognition system in a clear and accessible manner, facilitating easy interaction with transcribed speech for users of varying technical abilities. In this paper, the focus is on the development of the spontaneous speech model for the recognition of the Punjabi language. The GUI for Punjabi speech model also has been created and tested. The recognition accuracy is good for Punjabi sentences and much higher for Punjabi words. The python programming are used to build a speech model for Punjabi live speech.

Table of Contents

List of Figures

Chapter 1: Introduction

Engineering is the discipline of using scientific knowledge and mathematical tools to solve problems and create innovative solutions for the real world. It's the bridge between scientific theories and practical applications. The spectrum of engineering encompasses a remarkable range. On one end, we have engineers meticulously crafting intricate microchips, the invisible engines powering our modern technology. On the other hand, there are those who design and build massive structures like bridges and dams, shaping the very landscape we inhabit.

With the advancement in engineering knowledge, Machine learning has become a transformative force across various industries. It's a branch of artificial intelligence that empowers computers to learn and improve without explicit programming. Imagine feeding a computer vast amount of data, and it starts to identify patterns and make predictions on its own. That's the core idea behind machine learning. Machine learning is emerging as a powerful weapon in our arsenal for tackling modern problems. From revolutionizing healthcare to combating climate change, its applications are vast and hold immense potential to improve our lives. Machine learning is rapidly transforming how we approach modern challenges, and speech recognition is a prime example of its problem-solving prowess.

Speech and text are the most common modes of communication. Automatic speech recognition (ASR) is the technology that automatically converts speech to its text form. Nowadays, there are many speech recognition tools integrated into our everyday life such as Google assistant, Siri, Google Translate etc. With the increasing use of smart devices and huge data resources, many companies and government agencies are interested in the development of speech technology. Another important application of ASR is in language documentation. More than half of the world’s languages are currently at different levels of endangered state, which may not persist another century. First nations or tribal languages are amongst the highest severity of the endangered languages as most of them have few speakers and the new generations are diverting to major spoken languages due to economic, social or political reasons. Languages are carriers of cultural heritage, memory of important individuals or events, and they hold diverse information on linguistic evolution. Therefore, in order to preserve these languages, linguists have already started many documentation projects.

The Wav2Vec 2.0-XLSR-53 is a powerful model that was pre-trained to learn multilingual speech representation end-to-end in an unsupervised way. Dialect Identification (DID) and Accent Identification (AID) can be used to improve Automatic Speech Recognition (ASR) systems in languages with multiple distinctive dialects or accents. It is evaluated how the model performs when trained on low-resource datasets. Various experiments are conducted in the areas of AID in English and Spanish. In addition, evaluations were executed on short samples. To further explore the capabilities of wav2vec, an age and sex classifier is trained on German speech. The used corpora were extracted from Mozilla's Common Voice (Common Voice). Trained on more than 336,000 hours of publicly available speech recordings, XLS-R is based on wav2vec 2.0, approach to self-supervised learning of speech representations. That's nearly 10 times more hours of speech than the best, previous model we released last year, XLSR-53. Utilizing speech data from different sources, ranging from parliamentary proceedings to audio books, it is expanded to 128 different languages, covering nearly two and a half times more languages than its predecessor.

1.1 Overview

As the modern world continues to embrace technologies that facilitate seamless communication and streamline everyday tasks, the role of speech recognition systems has emerged as a cornerstone in this digital evolution. Speech recognition systems play a crucial role in bridging the gap between humans and machines by enabling the conversion of spoken language into text. This fundamental capability not only enhances accessibility and productivity but also unlocks new possibilities for innovation across diverse domains. From voice-activated virtual assistants to automated transcription services, speech recognition systems have revolutionized the way we interact with technology, offering unparalleled convenience and efficiency. However, the importance of speech recognition extends beyond mere convenience, particularly in the context of linguistic diversity.

In a world characterized by multilingualism, the absence of robust speech recognition systems for specific languages poses significant challenges. Without adequate support for speech recognition in a particular language, individuals may encounter barriers to accessing digital resources, communicating effectively, and participating fully in the digital realm. In light of these considerations, the development of speech recognition systems tailored to the linguistic needs of diverse communities is essential. By addressing the unique challenges associated with language diversity, such as dialectal variations, script complexities, and limited linguistic resources, these systems can empower individuals to engage more seamlessly with technology and overcome linguistic barriers. Our proposed system will facilitate the smooth transition from spoken Punjabi to written text, empowering Pakistani Punjabi speakers to effectively communicate, access information, and engage with digital content in their native language.

1.2 Problem Statement

Speech recognition systems have become an integral part of modern technology, facilitating various applications ranging from virtual assistants to voice-controlled devices. However, most existing speech recognition systems are primarily developed for widely spoken languages like English, leaving many other languages underserved, including Punjabi. Punjabi, with its rich linguistic heritage and widespread usage, demands attention in the domain of speech recognition for effective communication, accessibility, and technological advancement within Punjabi-speaking communities. Despite the growing demand for Punjabi speech recognition systems, the development in this area remains relatively limited compared to other languages. Existing systems often lack accuracy, robustness, and linguistic coverage specific to Punjabi, thereby hindering their practical utility.

This presents a significant challenge for Punjabi speakers who rely on speech recognition technology for various tasks, such as dictation, voice commands, and transcription. Furthermore, the complexity of Punjabi language poses unique challenges for speech recognition systems. Punjabi is a tonal language, characterized by a rich phonetic inventory and complex phonological processes. It encompasses a diverse set of dialects and accents, further complicating the task of accurately transcribing spoken Punjabi into text. Additionally, Punjabi script presents challenges in terms of orthographic variation and script-specific features, necessitating specialized approaches for effective recognition.

In recent years, significant advancements have been made in the field of speech recognition, particularly with the emergence of deep learning techniques. One such technique is the Wav2Vec model, which has shown promising results in speech recognition tasks for various languages. Wav2Vec utilizes self-supervised learning to pretrain on large corpora of unlabeled speech data, followed by fine-tuning on labeled data for specific tasks. This approach has demonstrated state-of-the-art performance in speech recognition tasks, especially for languages with limited resources. However, the application of Wav2Vec to Punjabi speech recognition remains relatively unexplored. Adapting Wav2Vec to the nuances of Punjabi language presents several challenges and opportunities. Firstly, the availability of large-scale, high-quality Punjabi speech corpora for pretraining is limited, which affects the model's ability to learn robust representations of Punjabi speech.

Additionally, fine-tuning Wav2Vec on labeled Punjabi speech data requires specialized methodologies to address the linguistic intricacies and acoustic variations inherent in Punjabi. Therefore, the primary objective of this thesis is to develop and evaluate a speech recognition system for the Punjabi language using the Wav2Vec model.

1.3 Proposed Solution

The major goal of our proposed solution is to develop a robust speech recognition system tailored to the linguistic characteristics of Pakistani Punjabi, overcoming the challenges of limited linguistic resources, dialectal variations, and script complexities. This system aims to enhance accessibility, communication, and digital inclusion for Pakistani Punjabi speakers by providing accurate and efficient transcription of spoken Punjabi into written text.

This solution comprises several key components aimed at data collection, model adaptation, fine-tuning strategies, handling orthographic challenges, and evaluation metrics. By implementing the proposed solution, it is anticipated that a robust and accurate speech recognition system tailored to the Punjabi language will be developed. This system will not only contribute to bridging the technological gap in Punjabi language processing but also serve as a foundation for further research and applications in speech technology for underrepresented languages. Moreover, the insights gained from this endeavor can be extrapolated to benefit other similar languages facing similar challenges, thereby promoting linguistic diversity and inclusivity in the field of speech recognition.

1.4 Working Principle

The project mainly works on the principles of signal processing and natural language processing amalgamated with machine learning algorithms. The simplified form of working of this model is shown in the picture below:-

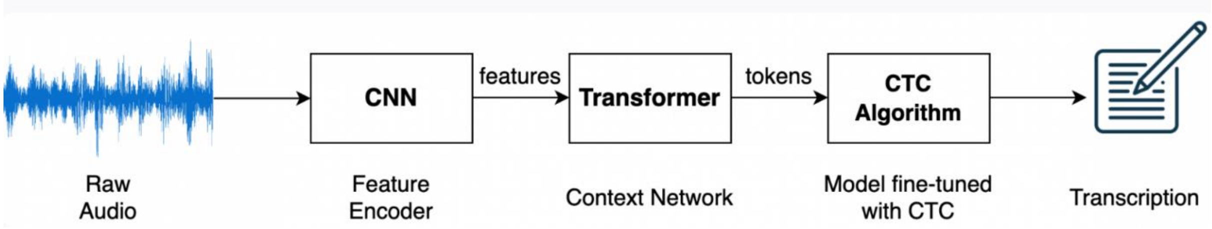


Figure 1: Working of Wav2Vec model [1]

The project is divided into different modulus and every module is inter-woven with the next module.

The list of modules is as under:

1.4.1 Data Collection:

The integral part of the project is the preparation of datasets. The dataset comprises of audio data of 5 hours, divided into segments each of 5 to 6 seconds.

1.4.1.1 Professional Recordings:

This project uses high-quality recordings of Punjabi speech in various dialects and speaking styles recorded in professional studios.

1.4.1.2 Internet-sourcing:

Speech data from a diverse range of Punjabi speakers is a helpful data. Our project uses these online platforms by keeping in mind the quality assurance.

1.4.2 Processing:

Preprocessing audio samples is a crucial step in preparing them for use in a speech recognition system. Following techniques are crucial here.

1.4.2.1 Resampling:

All audio files were resampled to a consistent sampling rate (e.g., 16 kHz) to match the requirements of the chosen wav2vec model.

1.4.2.2 Noise Reduction:

Background noise like traffic or conversations can be present in recordings. De-noising algorithms were applied to clean the audio data, addressing the specific types of noise present in the dataset.

1.4.2.3 Normalization:

Audio signals can vary in volume. Normalization techniques aim to bring all audio samples to a similar volume range. This helps prevent louder recordings from disproportionately influencing the model's learning process. Scaling techniques used to normalize audio features, preventing any one aspect from dominating the model's learning

1.4.3 Fine Tuning:

It involves adjusting the pre-trained Wav2Vec model's parameters to better understand Punjabi speech patterns. This process entails re-training the model using speech data to adapt its acoustic representations to the nuances of Punjabi phonetics. Fine-tuning involves initializing the Wav2Vec model with pre-trained weights and then updating those weights using Punjabi speech data through an iterative training process.

1.4.3 Evaluation:

During evaluation, we are assessing the performance and effectiveness of the model. We begin by dividing the available dataset into training, validation, and testing sets to ensure unbiased evaluation. Then, we use evaluation metrics, such as Word Error Rate (WER) and Character Error Rate (CER), to quantify the system's performance.

We compare the system's performance against baseline models to establish a benchmark. We validate the trained model using a separate validation dataset to ensure its generalization performance and identify potential over fitting. Next, we evaluate the performance of the validated model on unseen testing data to assess its accuracy and robustness in transcribing Punjabi speech. We conduct detailed error analysis to identify common errors made by the system and understand the challenges faced in transcribing Punjabi speech. Based on the evaluation results, further fine-tuning of the model is done to improve its performance.

1.4.5 GUI presentation:

The visual demonstration of the project is done through the aid of GUI (graphical user interface).

1.5 Objectives

1.5.1 General Objectives:

“To develop an advanced automatic speech recognition system leveraging Machine Learning (ML) techniques, tailored specifically for a low-resource language, Punjabi.”

1.5.2 Academic Objectives:

- Development of a Punjabi Speech Recognition System
- To implement Machine Learning techniques and simulate the results
- To increase productivity by working in a team
- To design a project that contributes to the welfare of society

1.6 Scope

The scope of this project entails the development of an automatic speech recognition system specifically designed for Punjabi, a low-resource language. The project focuses on addressing the linguistic challenges unique to Punjabi and aims to enhance accessibility and usability for Punjabi speakers in digital environments.

1.7 Deliverables

1.7.1 Military

a. Command and Control: Enable voice commands for controlling military systems and equipment.

b. Surveillance: Implement speech recognition for analyzing audio data in surveillance systems.

c. Communication: Facilitate secure voice communication in Punjabi among military personnel.

d. Operational Efficiency: Enhance the efficiency of military operations through voice-activated tools and systems

1.7.2 Civil:

a. Customer Service: Implement Punjabi speech recognition for better customer service in local businesses and government offices.

b. Accessibility: Enhance accessibility for Punjabi speakers in various applications, such as voice-activated devices and services.

c. Healthcare: Enable voice recognition for medical transcription and healthcare documentation in Punjabi.

d. Education: Facilitate speech-to-text features in educational tools for Punjabi-speaking students.

e. Help People with hearing problems.

1.8 Relevant Sustainable Development Goals

The project addresses the locally relevant socio-economic issue of digital exclusion among Punjabi-speaking communities in Pakistan. Due to the lack of robust language support and accessibility in digital technologies, Punjabi speakers often face barriers to accessing online resources, participating in digital communication, and benefiting from technological advancements. This initiative can empower individuals to fully engage with technology, access educational and economic opportunities, and participate in the digital economy, ultimately contributing to socio-economic development and empowerment within Punjabi-speaking communities. It covers two sustainable development goals.

1.8.1 SDG 3:

The project aims to enhance access to educational resources and opportunities for Punjabi-speaking communities in Pakistan. The development of a Punjabi speech recognition system contributes to achieving SDG 3 by promoting digital inclusion, enhancing access to educational opportunities, and fostering linguistic diversity and cultural preservation within Punjabi-speaking communities.

1.8.2 SDG 9:

The project also aligns with SGD 9 i.e. Industry, Innovation, and Infrastructure. The development of a Punjabi speech recognition system involves innovation in technology and infrastructure, particularly in the field of natural language processing and machine learning. The development of a speech recognition system tailored to Punjabi addresses the need for inclusive and accessible digital infrastructure, particularly for languages and communities that are traditionally underserved or overlooked in technological advancements.

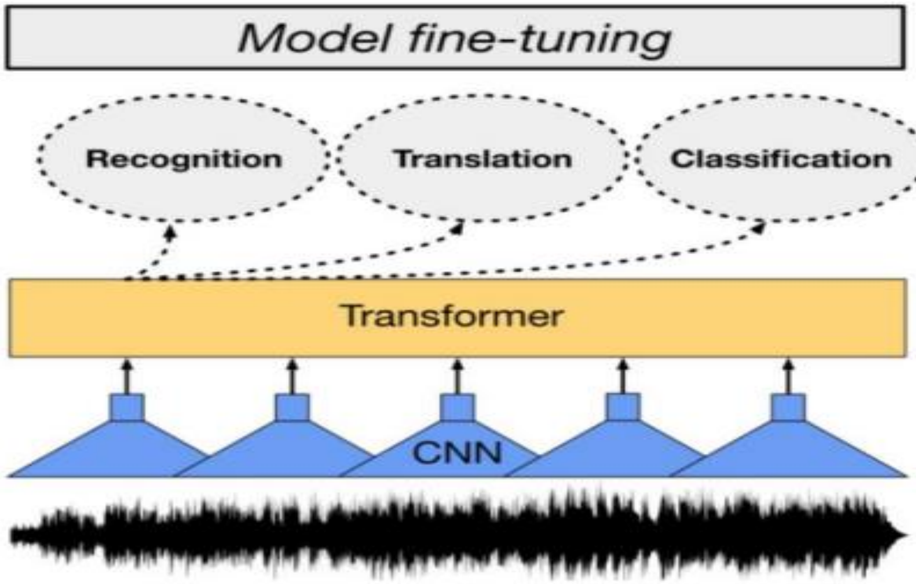


Figure 2: Fine-tuning of Wav2Vec model [2]

Chapter 2: Literature Review

In recent years, the intersection of natural language processing (NLP) and machine learning has led to significant advancements in speech recognition technology. However, while major languages have seen substantial progress in the development of speech recognition systems, low-resource languages such as Punjabi have often been overlooked.

Literature review is an important step for development of an idea to a new product. Our literature review is divided into the following points.

- Technological Background
- Existing models and their drawbacks

2.1 Technological Background

In order to understand the context and significance of developing a speech recognition system for Punjabi language using the Wav2Vec model, it is essential to delve into the technological background encompassing key concepts, methodologies, and advancements in the fields of speech recognition and deep learning. This section provides a comprehensive overview of relevant topics, laying the foundation for the subsequent discussion and analysis within the thesis.

2.1.1 Speech Recognition Technology: Speech recognition technology enables the conversion of spoken language into text, allowing computers to interpret and process human speech. The evolution of speech recognition systems can be traced back to the early systems based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). These systems relied on statistical modeling of acoustic features and language modeling techniques to transcribe speech accurately. However, they often suffered from limitations in robustness and scalability, particularly for languages with complex phonetic structures like Punjabi.

2.1.2 Deep Learning in Speech Recognition: Deep learning has revolutionized the field of speech recognition, enabling significant improvements in accuracy and performance. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been widely employed for acoustic modeling and sequence-to-sequence modeling in speech recognition tasks. These deep learning architectures learn hierarchical representations of speech data, capturing both local and global dependencies for more accurate transcription.

2.1.3 Self-Supervised Learning: Self-supervised learning has emerged as a powerful paradigm for pretraining deep neural networks on unlabeled data, followed by fine-tuning on labeled task-specific data. Techniques such as contrastive learning, autoencoding, and predictive modeling are commonly used for self-supervised learning. In the context of speech recognition, self-supervised learning facilitates the extraction of meaningful representations from raw audio signals, enabling the development of robust and language-agnostic models.

2.1.3 Wav2Vec Model: Wav2Vec is a state-of-the-art speech recognition model developed by Facebook AI Research (FAIR). It leverages self-supervised learning to pretrain on large-scale unlabeled speech corpora, learning contextualized representations of speech. Wav2Vec employs a novel contrastive predictive coding (CPC) objective, where the model predicts future audio frames conditioned on past frames. This pretraining stage enables the model to capture high-level semantic information from raw audio signals, making it suitable for downstream speech recognition tasks with limited labeled data.

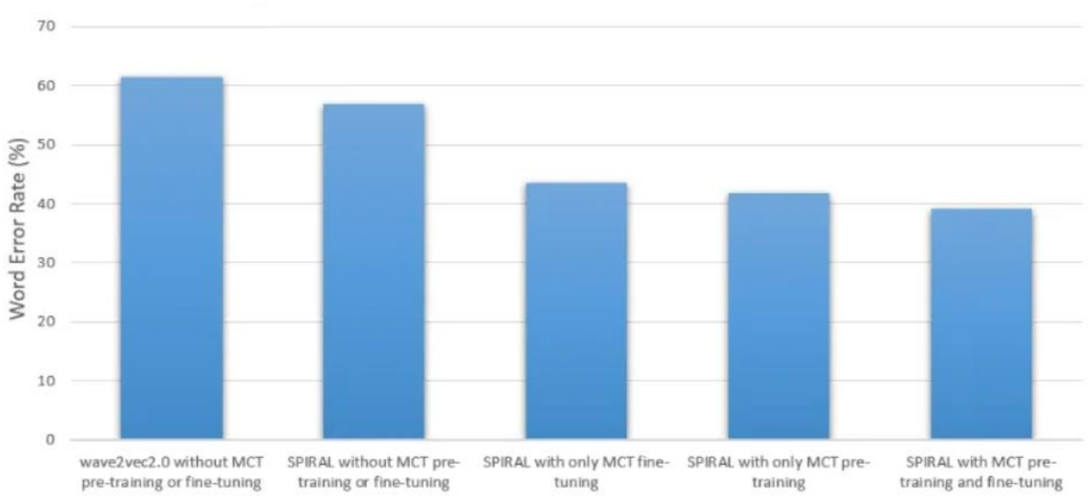


Figure 3: Average % WER of Wav2Vec Model [3]

2.1.5 Punjabi Language Characteristics: Punjabi is a language spoken predominantly in the Punjab region of South Asia, with significant populations in India, Pakistan, and diaspora communities worldwide. It is characterized by a rich phonetic inventory, including consonants, vowels, and tones, which contribute to its melodic and rhythmic qualities. Punjabi script, primarily written in Gurmukhi and Shah-Mukhi scripts, exhibits orthographic variations and script-specific features that pose challenges for automated transcription and text processing.

2.1.6 Challenges in Punjabi Speech Recognition: Developing a speech recognition system for Punjabi language entails addressing several challenges unique to the linguistic and acoustic properties of Punjabi. These challenges include dialectal variations, accent diversity, code-switching between Punjabi and other languages, and limited availability of annotated speech corpora for training and evaluation. Moreover, the absence of standardized orthographic conventions and resources further complicates the development of accurate and robust speech recognition models for Punjabi.

By exploring these technological aspects, the thesis aims to provide a comprehensive understanding of the underlying principles and methodologies relevant to the development of a speech recognition system for Punjabi language using the Wav2Vec model. Moreover, it sets the stage for addressing specific research questions and challenges outlined in the problem statement, guiding the subsequent experimentation, analysis, and interpretation within the thesis framework.

2.2 Existing models and their drawbacks

Speech recognition technology has witnessed significant advancements over the years, with various models and approaches developed to tackle the complexities of converting spoken language into

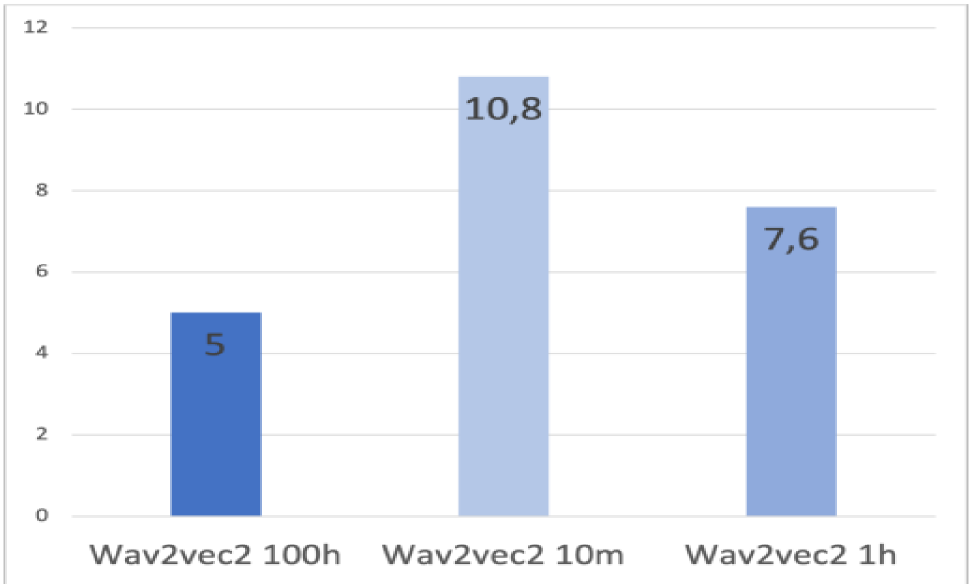


Figure 4: Average % WER of Wav2Vec Model trained on different datasets [4]

text. In this comprehensive exploration, we delve into the landscape of existing models, ranging from traditional statistical methods to state-of-the-art deep learning architectures. Additionally, we analyze the challenges associated with these models, highlighting their limitations and areas for improvement.

2.2.1. Hidden Markov Models (HMMs): Hidden Markov Models (HMMs) have long been the cornerstone of speech recognition systems. HMMs model the temporal dependencies in speech signals, treating speech as a sequence of phonemes or acoustic units. These models utilize statistical techniques to estimate the probabilities of transitioning between different states and emitting observed acoustic features. However, HMM-based systems often struggle with modeling long-range dependencies and capturing complex linguistic structures, leading to suboptimal performance, especially for languages with rich phonetic inventories like Punjabi.

2.2.2 Gaussian Mixture Models (GMMs): Gaussian Mixture Models (GMMs) are commonly used in conjunction with HMMs for acoustic modeling in speech recognition. GMMs represent the probability density of acoustic features using a mixture of Gaussian distributions. While GMMs have been effective in capturing the statistical properties of speech features, they suffer from limitations in modeling non-linear relationships and capturing contextual information. As a result, GMM-based systems may exhibit reduced accuracy and robustness, particularly in noisy environments or for speakers with diverse accents.

2.2.3. Deep Neural Networks (DNNs): Deep Neural Networks (DNNs) have revolutionized the field of speech recognition, offering superior performance and scalability compared to traditional HMM-GMM systems. DNNs employ multiple layers of neurons to learn hierarchical representations of speech features, enabling more effective modeling of complex relationships. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been widely adopted for acoustic modeling and sequence modeling in DNN-based speech recognition systems.

2.2.3. Challenges with DNN-based Models:

a. **Data Requirements:** DNN-based models often require large amounts of labeled data for training, which may be scarce or expensive to acquire, especially for low-resource languages like Punjabi.

b. **Overfitting:** DNNs are prone to overfitting, particularly when trained on limited data. Overfitting can lead to poor generalization performance on unseen data, impacting the robustness and reliability of the speech recognition system.

c. **Domain Adaptation:** DNN-based models may struggle with adapting to diverse linguistic environments, dialectal variations, and speaker characteristics. Domain adaptation techniques are required to enhance the model's adaptability and generalization capabilities across different domains and speaker demographics.

d. **Computational Complexity:** Training and inference with DNN-based models involve significant computational resources, including specialized hardware accelerators such as GPUs and TPUs. The computational complexity may pose challenges for deploying speech recognition systems in resource-constrained environments or on mobile devices.

2.2.5. End-to-End Models: End-to-End (E2E) models have gained traction in recent years due to their simplicity and effectiveness in directly mapping input speech to output text without the need for intermediate representations. E2E models typically consist of a single neural network that jointly learns feature extraction, acoustic modeling, and language modeling tasks. These models eliminate the need for handcrafted features and intermediate components, offering streamlined architectures and potentially higher accuracy.

2.2.6. Challenges with End-to-End Models:

a. **Data Efficiency:** End-to-End models often require large amounts of training data to achieve competitive performance, making them less practical for languages with limited annotated corpora like Punjabi.

b. **Lack of Interpretability:** The black-box nature of E2E models makes it challenging to interpret their internal representations and understand the decision-making process. This lack of interpretability may hinder the diagnosis of errors and the refinement of model behavior.

c. **Robustness to Noise:** E2E models may exhibit reduced robustness to noise and environmental variability compared to traditional modular systems. Robustness techniques such as data augmentation, multi-task learning, and adversarial training are necessary to improve the resilience of E2E models to noisy input conditions.

d. **Domain Adaptation:** Similar to DNN-based models, E2E models may require domain adaptation techniques to adapt to diverse linguistic environments, dialects, and speaking styles encountered in real-world applications.

2.2.7. Transformer-based Models: Transformer-based models have emerged as a powerful architecture for various natural language processing tasks, including speech recognition. Transformers leverage self-attention mechanisms to capture long-range dependencies and contextual information in input sequences. Models such as the Transformer, BERT (Bidirectional Encoder Representations from Transformers), and GPT (Generative Pre-trained Transformer) have demonstrated remarkable performance in capturing semantic and syntactic structures in text data.

Challenges with Transformer-based Models:

a. **Model Size and Complexity:** Transformer-based models tend to be large and computationally expensive, requiring substantial resources for training and inference. Model compression techniques, knowledge distillation, and efficient attention mechanisms are needed to reduce the computational burden and enable deployment on resource-constrained devices.

b. **Data Efficiency:** Like other deep learning models, transformer-based models benefit from large-scale annotated data for pretraining and fine-tuning. Strategies for leveraging unlabeled data and weak supervision are essential for improving data efficiency and adapting the models to low-resource languages like Punjabi.

c. **Handling Long Sequences:** Transformers may encounter challenges in processing long sequences, such as those encountered in continuous speech recognition tasks. Techniques like chunking, streaming, and hierarchical modeling are required to overcome these limitations and enable efficient processing of long-form audio data.

d. **Wav2Vec Model:** Wav2Vec is a recent advancement in speech recognition technology developed by Facebook AI Research (FAIR). Wav2Vec leverages self-supervised learning to pretrain on large amounts of unlabeled speech data, learning contextualized representations of speech. The model employs a contrastive predictive coding (CPC) objective to predict future audio frames conditioned on past frames, facilitating the extraction of high-level semantic information from raw audio signals.

Chapter 3: Interface Implementation and Functionality

In this section, we delve into the complicated domain of audio processing, a foundational aspect of any speech recognition system. We will navigate through two fundamental facets: the management of recorded audio files and the dynamics of real-time audio capture.

3.1 Audio Handling

Two types of audio samples are being processed in this project, already recorded audio samples and real-time sample. Working is as fol:-

3.1.1 Management of Recorded Audio The realm of recorded audio files encompasses a WAV formats. We embark on a journey to understand the working of this format and implications for our Punjabi speech recognition system. Exploring techniques for decoding, preprocessing, and loading these files is paramount in ensuring seamless integration into our system's pipeline.

3.1.2 Real-Time Audio Recording Real-time audio recording introduces a overabundance of challenges and opportunities. From harnessing the capabilities of the microphone hardware to mitigating latency and noise interference, every step is crucial in achieving accurate speech recognition. We divide the tones of accessing microphone streams, handling data in real-time, and implementing robust mechanisms for capturing pristine audio input.

3.2 GUI Integration

The graphical user interface (GUI) serves as the gateway to our speech recognition system, offering users a seamless and intuitive experience. This section delves into how the system's output is visualized on the GUI, presenting users with valuable insights and controls.

3.2.1 Audio File Selection Empowering users to select audio files from their local storage is a keystone of our GUI design. We explore the working of file browsing, selection dialogs, and the seamless integration of chosen files into our system's workflow. Providing users with a seamless experience from file selection to analysis is paramount in fostering engagement and usability.

3.2.2 Real-Time Audio Recording The ability to record audio in real-time adds a layer of dynamism to our system, enabling users to interact with it in a more immersive manner. We delve into the technical underpinnings of microphone access, recording controls, and data management during live audio capture. Addressing challenges such as latency and ensuring a reliable audio input stream is imperative for achieving optimal performance and user satisfaction.

3.2.3 Audio Playback Audio playback functionality enriches the user experience, allowing users to review recorded or selected audio files at their convenience. We explore techniques for decoding audio files, managing playback controls, and seamlessly integrating playback functionality into our GUI.

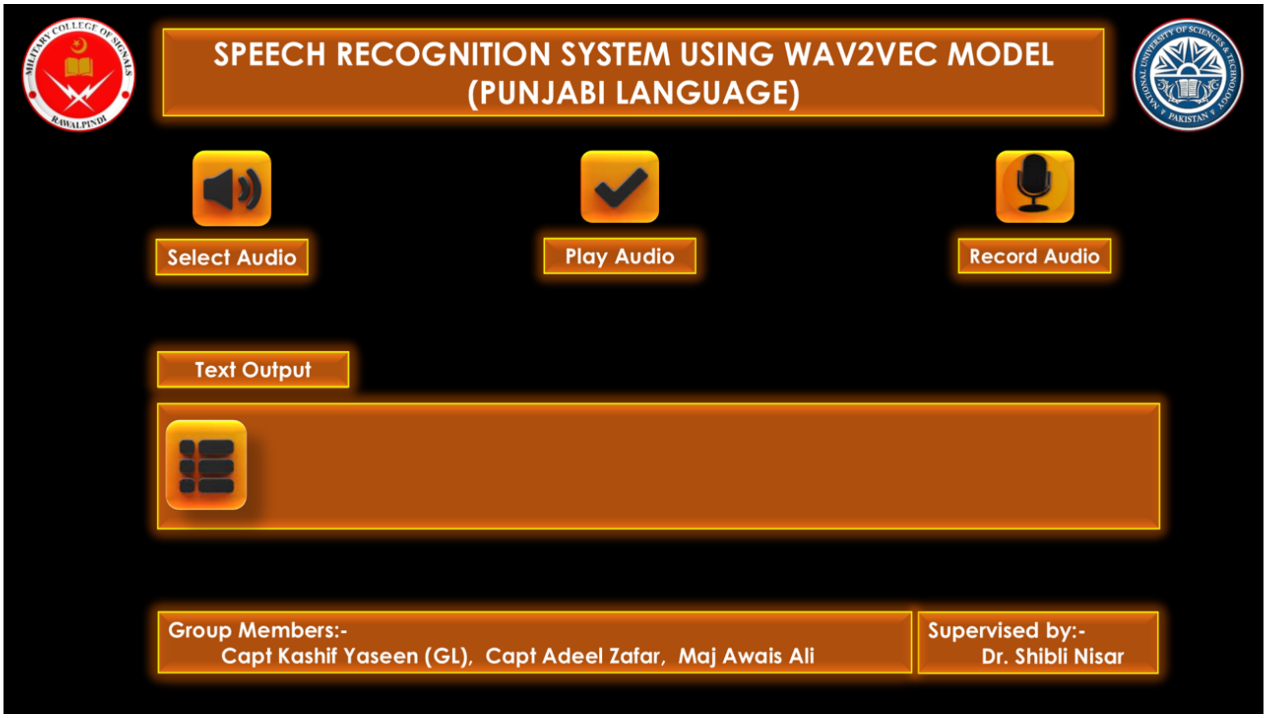
3.3 GUI Development Approaches The choice of GUI development approach significantly influences the user experience and system performance. This section explores various methodologies for creating the GUI of our Punjabi speech recognition system, each with its unique advantages and considerations.

3.3.1 Web-Based GUI Development Web-based GUI development exceeds platform limitations, offering cross-platform compatibility and easy deployment via web browsers. We have used a different approch to create user friendly GUI. This also enables the developer to make changes eaisly.

3.4 The developed GUI has three easy to choose options that includes a. Audio selection from the existing data set and then processing the audio file for further utilization by the model.

b. Recording the audio files in real time for processing and generating the output. Final GUI would look like fol:-

As it is evident from the above picture that GUI is user friendly and can eaisly be changed and modified according to the requirnment.



SPEECH RECOGNITION SYSTEM USING WAV2VEC MODEL
(PUNJABI LANGUAGE)

Group Members:-
Capt Kashif Yaseen (GL), Capt Adeel Zafar, Maj Awais Ali

Supervised by:-
Dr. Shibli Nisar

Figure 5: GUI Representation of Wav2Vec model [5]

Chapter 4: Methodology and Evalution of Model

4.1 HuggingFace

Numerous python libraries exist for the development of ASR, such as scikit-learn and NLTK. These are standalone packages that are highly useful for tuning hyperparameters and developing Natural Language Processing (NLP). For this research, complex development is not needed, rather testing of existing models is the focus. Testing and implementation of new and popular models is difficult in these environments, as they are not easily accessible yet. The platform HuggingFace is more accessible. The HF community provides a wide range of publicly accessible pretrained models and datasets that can be modified with ease and reuploaded to the HF hub for public use. Over 25 transformer tasks are available for use, e.g. ASR, image classification and text generation.

4.2 Pretraining of Model

For clarification, if a model is pretrained, it has already been adapted on data before. If a model is fine-tuned, it has been trained on a downstream task, such as ATC data transcription. So, a model can both be pretrained and fine-tuned. Models and datasets are constantly updated, which implies that SOTA tools are attainable. HF has a course for beginners and a decent documentation of the main libraries used:

4.2.1 Datasets and Transformers. The learning curve is high, though the models and data delivered outweigh the costs compared to other libraries.

4.3 Development of Various platforms.

Various platforms are used for the development of transcribing, evaluating, training and testing of ASR models. HF recommends using Google Colab for ease of programming and the absence of needing to own a high-end GPU for computation. For low computational power demanding tasks, such as small batch transcription, Google Colab is used, but the free to use platform has limitations that are diminishing our resources. GPUs can either be a T4, P100 or K80. These GPUs are inconsistently allocated by Google Colab and are unpredictable.

4.3.1 Resource Deprivation and loss of Progress: Fine-tuning models and transcribing full datasets regularly exceed the resources (GPU ram, disk space and system ram) made available by Google Colab, leading to resource deprivation and loss of progress. For this reason, local development in scripts is done in addition to cloud development. A GeForce GTX 1080 GPU is used for transcribing full datasets and training of models for more robust development.

4.4 Data Constraints

There is a distinct shortage of publicly available data for ATC that can be used for ASR, as most datasets do not include transcriptions (Badrinath & Balakrishnan, 2022). Websites such as ‘liveatc.net’ provide, as the domain name implies, live and very recent ATC communication audio. Based on airport codes and frequencies, different radio communication is selected and provided. This data does not include transcriptions and is therefore not suitable for training our models. Audio data is abundant and publicly available; however, audio data has to be manually transcribed, which is costly and time-consuming, leading to lower accessibility of transcribed ATC data (Zuluaga-Gomez et al., 2020).

4.4.1 Data Augmentation Techniques: These have been shown to give an increase in performance when applied to low resource data (T¨uske et al., 2014). Here, text data is normalized and unused tokens are removed, as they are deemed uninformative to transcription. 3.4 Models The models used are imported from the HF community. They are not trained or fine-tuned specifically on ATC data, as that domain has not been explored or published by the community yet. It is important that the models are recent, as ASR models are consistently updated and improved upon. Four models were considered for evaluating Wav2Vec2 models:

4.4.2 Base: Facebook/wav2vec2-base-960h Currently the most popular ASR model on HF that was updated 3 months ago. It has been pretrained and fine-tuned on 960 hours of Librispeech data. This model is called ‘base’ as it is mostly used for further training on a downstream task (finetuning). This model works well when clean, regular speech data is fed.

4.4.3 Robust: Facebook/wav2vec2-large-robust-ft-swbd-300h This model is a pretrained and fine-tuned version of the facebook/wav2vec2-large-robust model. The latest update to this model was 5 months ago. It has been fine-tuned on 300 hours of the Switchboard corpus. This is a telephone speech corpus that contains noisy data. This model has been chosen to see exactly how robust a robust model is when transcribing ATC data.

4.4.4 huBERT: Facebook/hubert-large-ls960-ft This model has been fine-tuned on 960 hours of Librispeech data and had its last update three months ago. The huBERT model follows the Wav2Vec2 architecture and extends upon it. huBERT’s training process is fundamentally different from Wav2Vec2, which the specifics of are out of scope. It has been shown that the model either improves or matches the performance of Wav2Vec2 (Hsu et al., 2021), so it is interesting to see if it performs better than the base and robust model on transcribing ATC data.

4.4.5 XLS-R-53: Jonatasgrosman/wav2vec2-large-xlsr-53-english The XLS-R model has been used for learning multilingual speech representation, dialect identification and accent identification (Pascal & Dominique, 2021). These challenges are apparent in ATC transcription as well, even though they are not the main objectives. The XLS-R model has been chosen, because the architecture uses contrastive learning and might perform better compared to the other models as a result. This specific model is updated regularly, with the last update pushed to HF being less than a month ago since beginning of research (August 2022).

4.5 Evaluation:

Evaluation of the model transcriptions was done by using the WER and CER. These are the standard for ASR models as the raw accuracy from speech to text is captured best (Malik et al., 2021). Metrics such as Glue or Bleu are not useful here, as they give information on the meaning of the output text compared to the reference text. These metrics are not needed in our ASR models and are thus omitted. The perplexity metric is used as well to determine how rare sentences are to appear in a gpt-2 text generation model. This model is trained on the text on 45 million Reddit web pages and contains 1.5 billion parameters. The gpt-2 model has been chosen, because it is the most popular text generation model found on HF with over 12 million downloads.

4.6 Fine-Tuning:

Fine-tuning (training) is done using the standard template given by user ‘patrickvonplaten’ on the HF blog. Hyperparameters are not tuned to the dataset and models, as resources are limited. Training is done locally, as Google Colab has limitations that are mentioned above. The number of training data steps in which the models were fine-tuned in are chosen from the three defined distributions: low, medium and high data ranges. Ranges being from 1-100, 101- 500 and 501-inf respectively. Due to time and resource restrictions, only five models are fine-tuned.

4.7 Language Model

The LM was made using an ARPA file containing only the training and validation transcriptions. Excluding the test cases from this file improves integrity and reduces overfitting. Overfitting is already a problem, as the dataset is quite small and only contains 743 unigrams. Because the unigram count is low, A maximum of 5-grams are stored. Higher rank n-grams were not used, as this could potentially lead to overfitting due to the nature of the dataset (largely containing short sentences). Higher/lower N-gram LMs are available, but are not studied here.

4.8 XLS-R Model

The reasoning behind choosing to finetune only the XLS-R model consists of three reasons. Firstly, the model is evaluated to be the most interesting to research on ATC data, as it has been shown that the model handles accents and multilingual data very well in comparison to the other models (Babu et al., 2021). Secondly, this model is exceedingly recent, as updates are posted regularly (August 2022 version used here). It could have been interesting to finetune all the models; however, our resources would not grant this, bringing forth the third reason.

4.8.1 Evaluation of Pretrained Models First, the pretrained models that are directly pulled from HF are evaluated. In figure it is shown that the WER and CER are relatively high. These models were not given an explicit LM. The base model performs the worst, which is expected, as it has not been pretrained on noisy speech data. It has the lowest WER and CER. The base model is the oldest model that is researched, which should be noted. The robust model performs better compared to the base model, as it is pretrained and finetuned on noisy telephone speech data. This results in a significant reduction in WER and a decent reduction in CER. The huBERT model possesses the smallest WER and CER. Improved or identical performance compared to standard Wav2Vec2 models is expected from the huBERT model (Hsu et al., 2021). A modified architecture might be the cause of this, as the model has been fine-tuned on the same data as the base and robust model. The XLS-R model was expected to have a lower WER than the base and robust model, as it ought to have an advantage in transcribing data that has multiple accents (Zuluaga-Gomez et al., 2022). However, this is not the case.

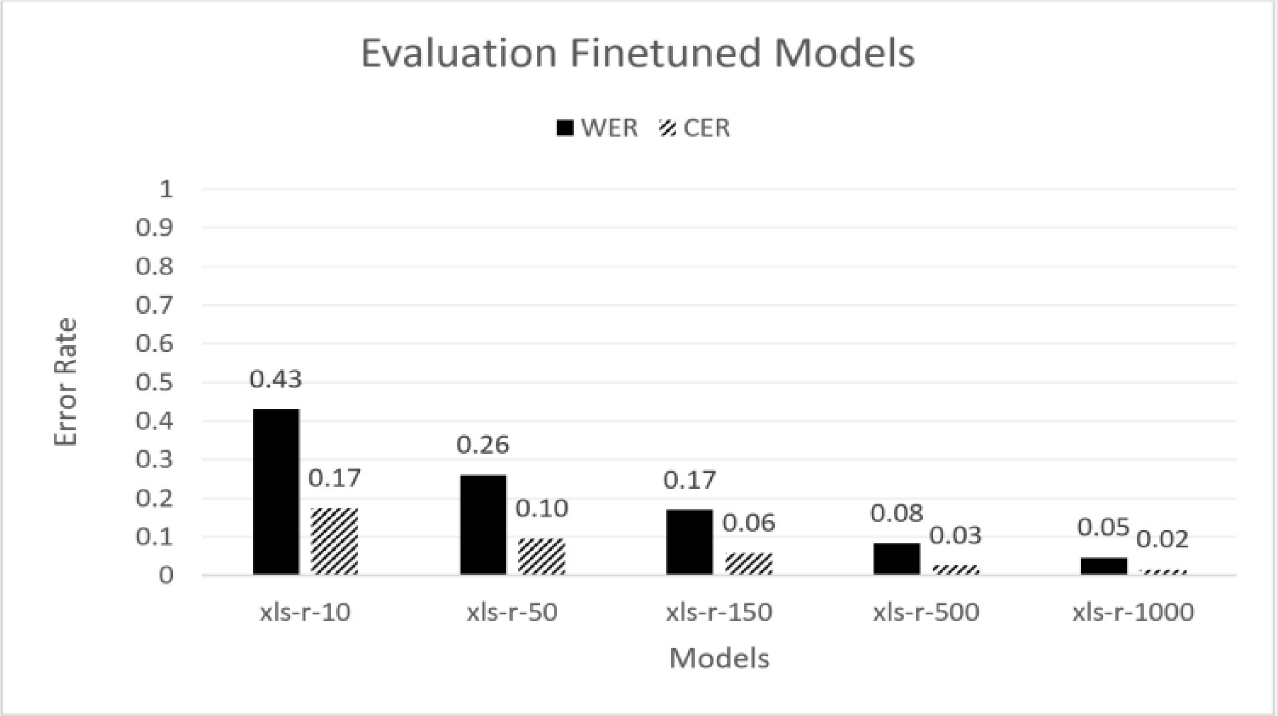


Figure 6: Evaluation of different Model trained on different datasets [6]

The cause for this might be that the robust model has a bigger advantage, as it is pretrained on noisy data, in contrast to the XLS-R model. The CER is relatively low for all models, which indicates that the models are able to recognize letters being spoken, however, they do not have an in-domain LM. Airport names and callsigns are uncommon in regular speech, relating to a high WER. Evaluation of the fine-tuned XLS-R models To continue, the XLS-R model is fine-tuned on increasing amounts of data (figure 2). We see a massive drop in WER and CER as the amount of data increases. The WER Reduction (WERR) for the XLS-R-10 model is already at ~39%. The highest WERR is seen found in the XLS-R-1000

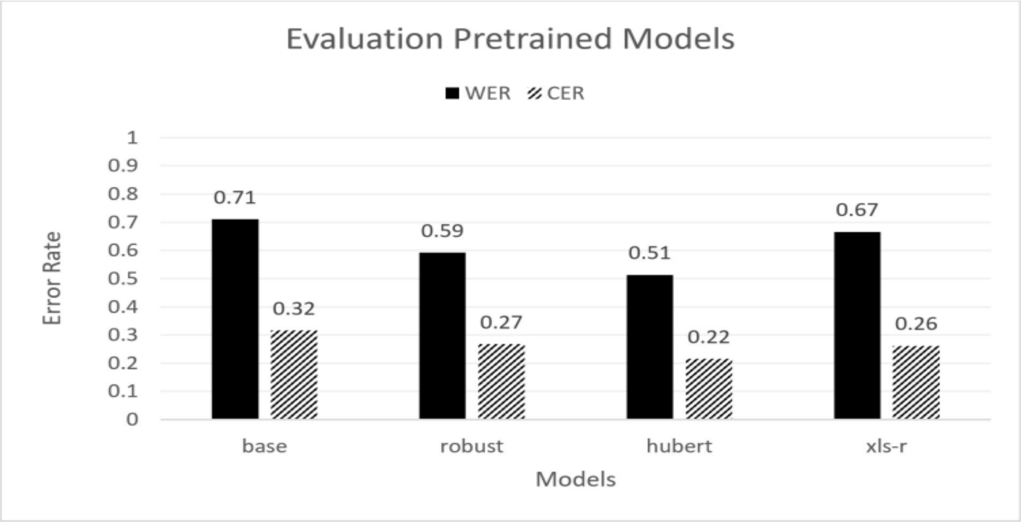


Figure 7: Evaluation of different Model trained on same datasets [7]

model at ~93%. CER Reduction (CERR) for the XLS-R-1000 model is significant as well (~92%)

4.9 Evaluation of the fine-tuned XLS-R models

In addition to a language model Next, an LM is added to the Wav2Vec2 processor. Absolute reduction of performance seems to be variable, but relative reduction of the WER and CER per model achieves ~33% decrease of WER and ~20% decrease of CER. The model with the lowest WER and CER is, as expected, the XLS-R-1000* model. However, the WERR of the lower data provided models are significant and impressive as well. Improvement of the XLS-R model The overall improvement of the XLS-R model is shown in figure 4. It shows that fine-tuning on ATC data can significantly improve the overall performance on WER and CER, especially in addition to

Chapter 5: Conclusion

In this thesis, we discussed a speech recognition system that can process audio smartly and more efficiently than the typically used language models. Our proposed system has an advantage over other traditional systems due to the low data requirement.

The completion of this model marks a significant milestone in our journey towards developing a Punjabi speech recognition system. Through meticulous exploration and detailed discussion, we've outlined a comprehensive framework for interfacing and working with audio data, laying the groundwork for a robust and user-friendly system. The inclusion of features such as audio file selection, real-time recording, and playback functionality serves to enrich the user interface, providing users with intuitive controls and seamless access to audio processing capabilities.

Furthermore, our examination of GUI development approaches has underscored the importance of selecting the right tools and methodologies to ensure cross-platform compatibility and responsiveness. As we move forward, it's essential to remain guided by the principles of usability, accessibility, and innovation. By incorporating user feedback, refining our algorithms, and embracing emerging technologies, we can continue to enhance the performance and functionality of our Punjabi speech recognition system. Ultimately, our goal is to empower users to interact with spoken Punjabi language effortlessly, opening up new possibilities for communication and collaboration in the digital age. With this model as our guide, we're poised to make significant strides towards realizing that vision.

Chapter 6: Future Work

Future milestones that need to be achieved to commercialize this project are the following.

6.1 Enhanced Accuracy and Performance: Continuously improving the accuracy and performance of the speech recognition system is crucial. Investing in research and development to refine algorithms, optimize models, and enhance language understanding capabilities will be essential for delivering a competitive and reliable product.

6.2 Scalability and Robustness: Scaling the system to handle a large volume of users and diverse usage scenarios is vital for commercial success. Ensuring robustness against various environmental factors, accents, and speech variations will be necessary to provide a consistent and seamless user experience.

6.3 Integration with Existing Platforms: Integrating the speech recognition system with existing platforms and applications will broaden its accessibility and utility. Investing in research and development of multimodal fusion techniques and user interface design will enable more natural and intuitive interactions.

6.3 Customization and Personalization: Offering customization and personalization features will cater to the unique needs and preferences of individual users and businesses. Implementing user profiles, adaptive learning algorithms, and customization options for language models and vocabulary will enhance user satisfaction and loyalty.

References and Work Cited

1. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Advances in neural information processing systems, vol. 33, pp. 12339– 12360, 2020.

2. A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in Proc. Interspeech, (Brno, Czechia), pp. 2326–2330, 2021.

3. F. Wu, K. Kim, J. Pan, K. J. Han, K. Q. Weinberger, and Y. Artzi, “Performance-efficiency trade-offs in unsupervised pre-training for speech recognition,” in Proc. ICASSP, (Singapore), pp. 7667–7671, IEEE, May 2022.

3. A. Vyas, W. Hsu, M. Auli, and A. Baevski, “Ondemand compute reduction with stochastic wav2vec 2.0,” in Proc. Interspeech, (Incheon, Korea), pp. 3038– 3052, Sept. 2022.

5 P. Vieting, C. Lüscher, W. Michel, R. Schlüter, and H. Ney, “On architectures and training for raw waveform feature extraction in ASR,” in Proc. ASRU, pp. 267–273, IEEE, 2021. 6.

S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in Proc. Interspeech, (Graz, Austria), pp. 3365–3369, Sept. 2019.

7. A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in International Conference on Learning Representations, Sept. 2019.

8. Z. Fan, M. Li, S. Zhou, and B. Xu, “Exploring wav2vec 2.0 on speaker verification and language identification,” in Proc. Interspeech, (Brno, Czechia), pp. 1509–1513, 2021

9. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.

10. A. Baevski, S. Shah, C. Fuegen, and M. Auli, “Wav2vec: State-of-theart speech recognition through self-supervision,” Sep 2019. [Online]. Available: https://ai.facebook.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

11. S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” arXiv preprint arXiv:1903.05862, 2019.

12. A. Baevski, A. Conneau, and M. Auli, “Wav2vec 2.0: Learning the structure of speech from raw audio,” Sep 2020. [Online]. Available:

13. H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 1, pp. 117–128, 2010.

13. E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with gumbel-softmax,” arXiv preprint arXiv:1611.01133, 2016.

15. T. Kendall, C. Vaughn, C. Farrington, K. Gunter, J. McLean, C. Tacata, and S. Arnson, “Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ing),” Frontiers in Artificial Intelligence, vol. 3, p. 638533, 2021.

16. R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” arXiv preprint arXiv:1912.06670, 2019. 17. “Nst swedish dictation,” Feb 2003. [Online]. Available: https: //www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-17/

18. E. Lagerlöf, “A swedish wav2vec versus google speech-to-text,” 2022.

19. M. Elfeky, M. Bastani, X. Velez, P. Moreno, and A. Waters, “Towards acoustic model unification across dialects,” in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 623–628.

20. Y. Huang, D. Yu, C. Liu, and Y. Gong, “Multi-accent deep neural network acoustic model with accent-specific top layer using the kldregularized model adaptation,” in Fifteenth Annual Conference of the International Speech Communication Association, 2013.

21. A. Jain, M. Upreti, and P. Jyothi, “Improved accented speech recognition using accent embeddings and multi-task learning.” in Interspeech, 2018, pp. 2353–2358.

22. B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018, pp. 3739–3753.

23. K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 3815–3819.

24. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. http://arxiv.org/abs/2111.09296

25. Badrinath, S., & Balakrishnan, H. (2022). Automatic Speech Recognition for Air Traffic Control Communications. In Transportation Research Record (Vol. 2676, Issue 1, pp. 798–810). SAGE Publications Ltd. https://doi.org/10.1177/03611981211036359

26. Baevski, A., & Mohamed, A. (2020). Effectiveness of Self-Supervised pre-training for ASR. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7694-7698, doi: 10.1109/ICASSP40776.2020.9054224.

27. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. http://arxiv.org/abs/2006.11477

28. Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 4774–4778. https://doi.org/10.1109/ICASSP.2018.8462105

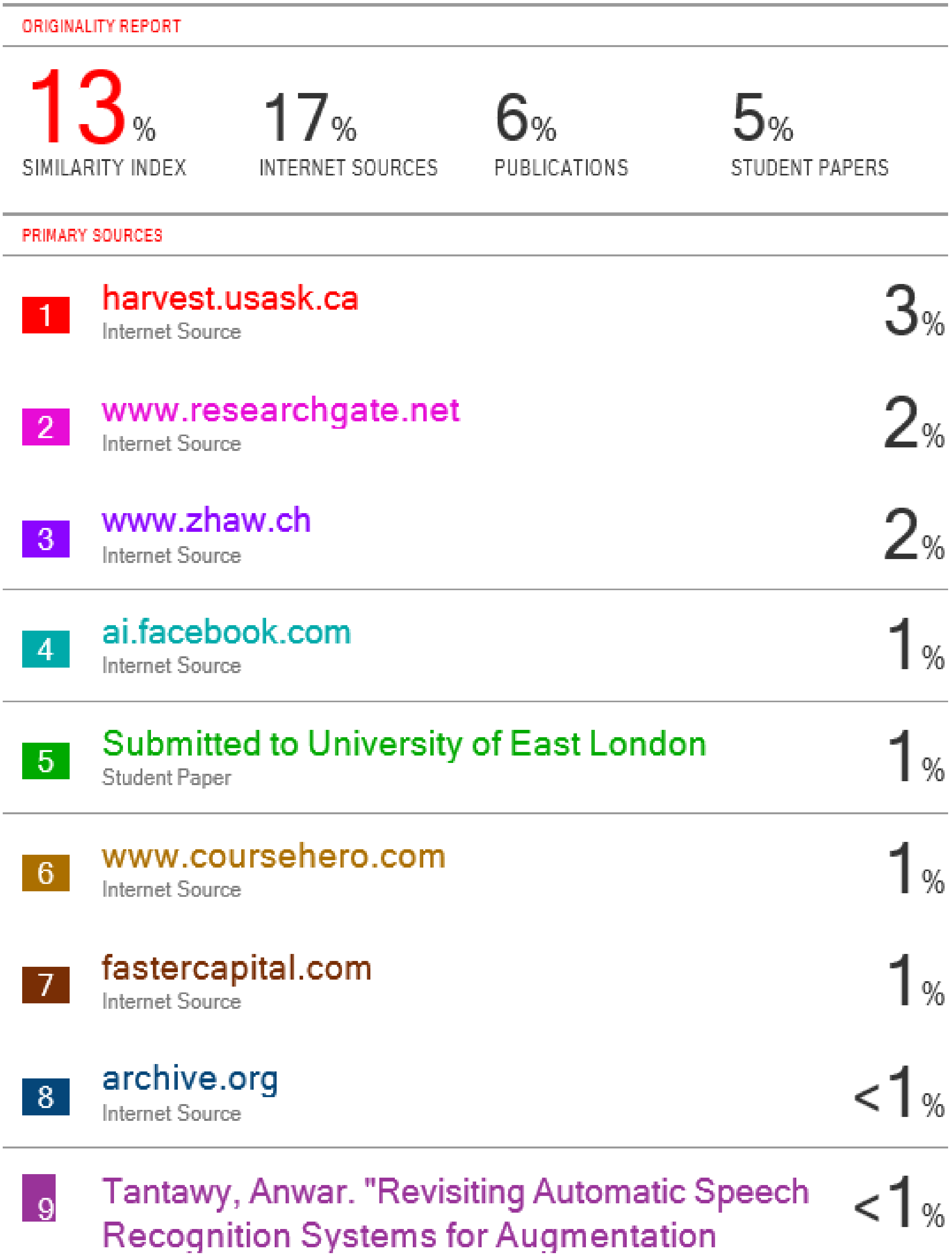
29. Helmke, H., Ohneiser, O., Muhlhausen, T., & Wies, M. (2016). Reducing controller workload with automatic speech recognition. AIAA/IEEE Digital Avionics Systems Conference - Proceedings, 2016-December. https://doi.org/10.1109/DASC.2016.7778024

30. Hofbauer, K., Petrik, S., & Hering, H. (n.d.). The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). http://www.lrec-conf.org/proceedings/lrec2008/pdf/545 paper.pdf

31. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio Speech and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

32. Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. http://arxiv.org/abs/2104.01027

33. Jia, G., Cheng, F., Yang, J., & Li, D. (2018). Intelligent checking model of Chinese radiotelephony read-backs in civil aviation air traffic control. Chinese Journal of Aeronautics, 31 (12), 2280–2289. https://doi.org/10.1016/j.cja.2018.10.001

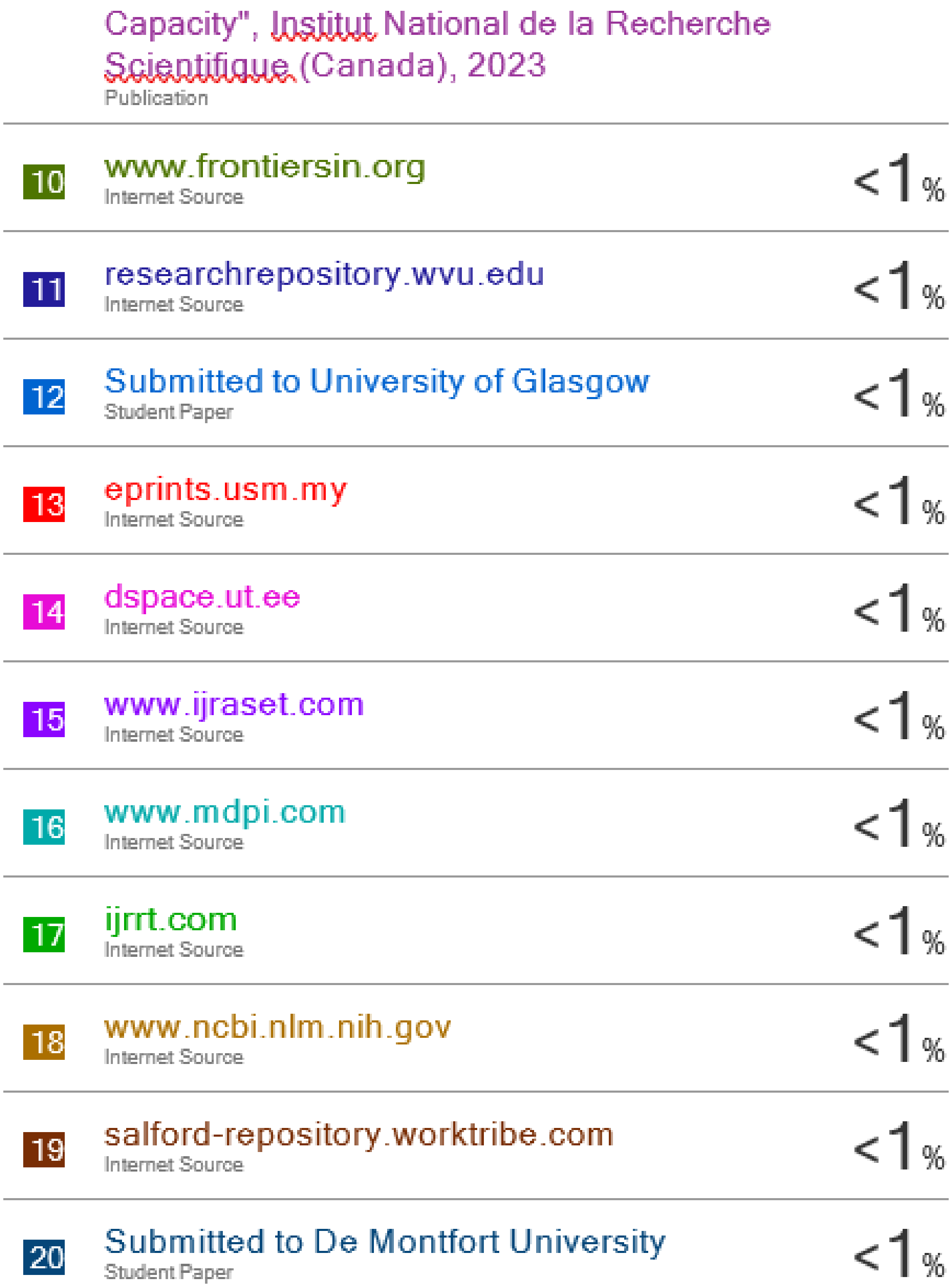


ORIGINALITY REPORT

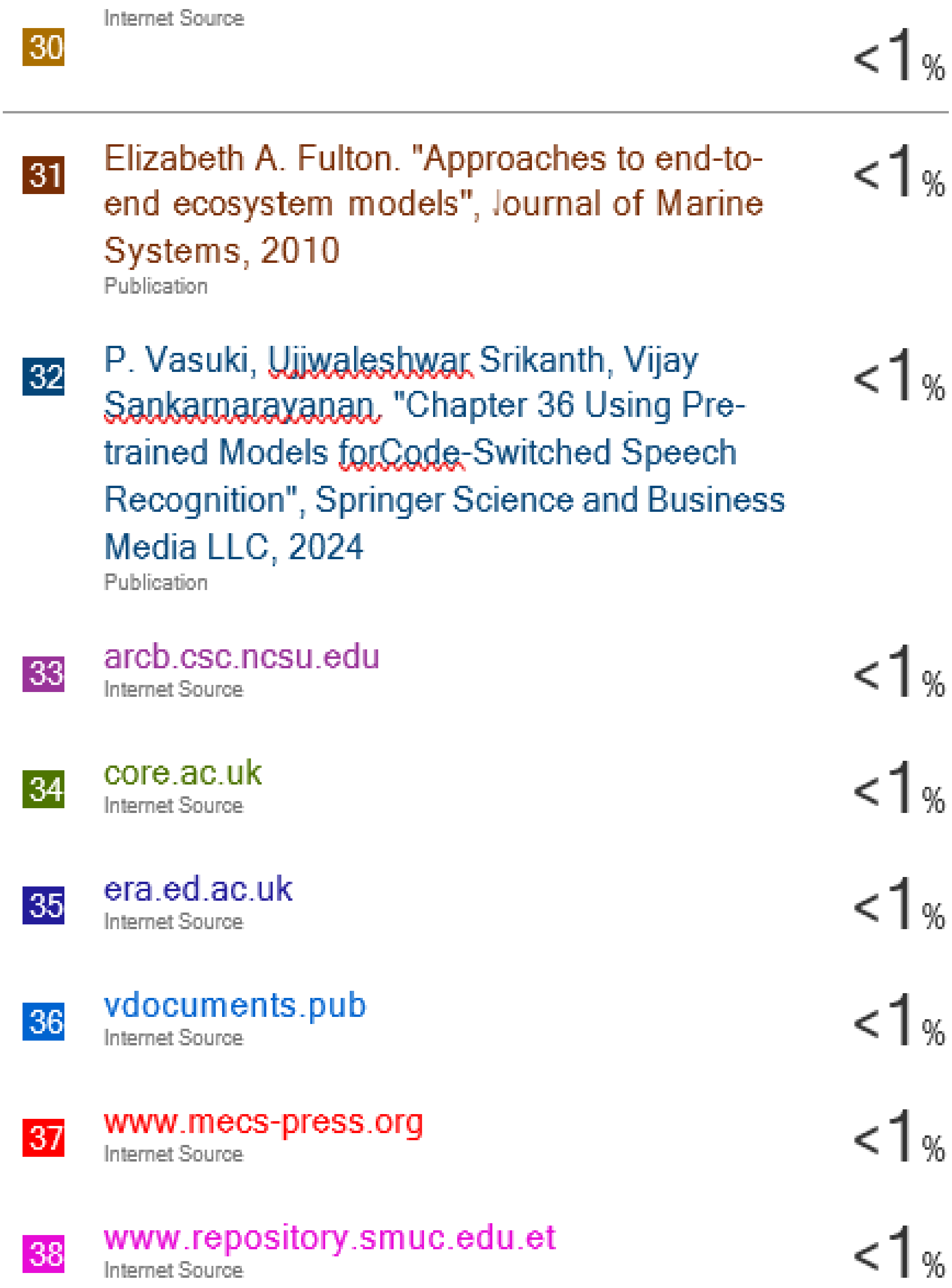
13%	17%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

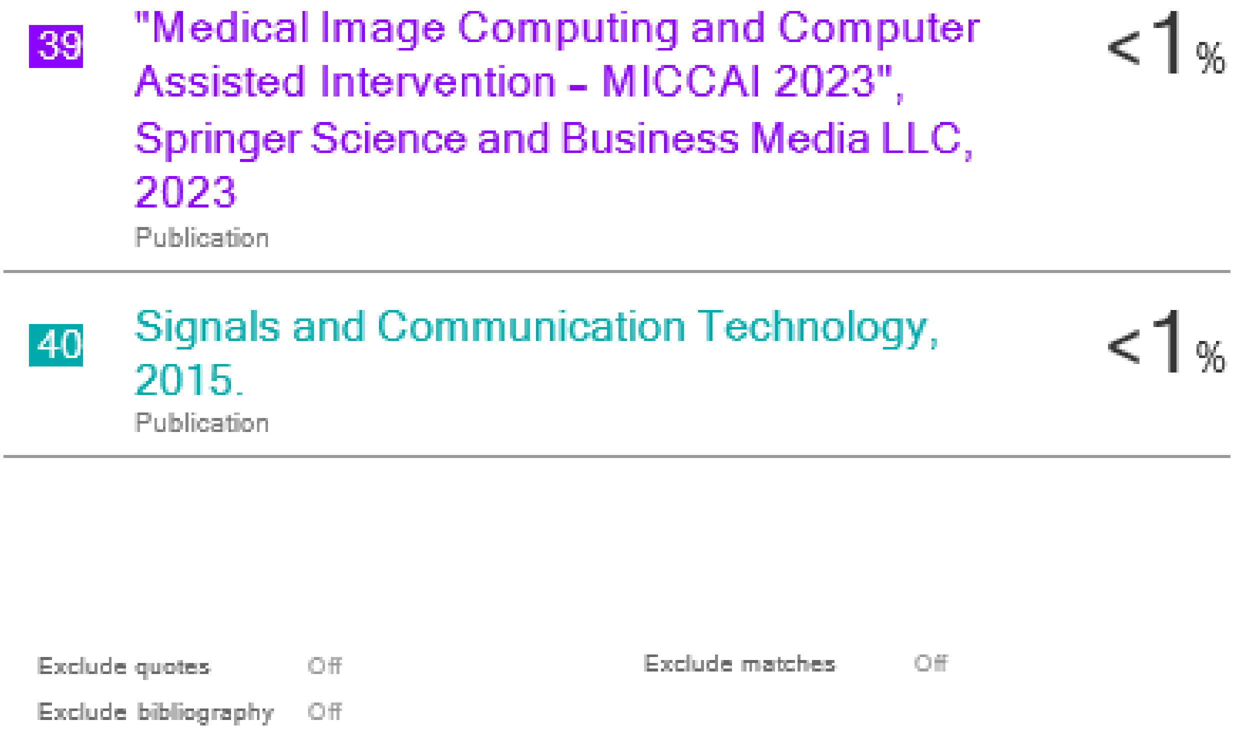
1	harvest.usask.ca Internet Source	3%
2	www.researchgate.net Internet Source	2%
3	www.zhaw.ch Internet Source	2%
4	ai.facebook.com Internet Source	1%
5	Submitted to University of East London Student Paper	1%
6	www.coursehero.com Internet Source	1%
7	fastercapital.com Internet Source	1%
8	archive.org Internet Source	<1%
9	Tantawy, Anwar. "Revisiting Automatic Speech Recognition Systems for Augmentation	<1%



	Capacity", Institut National de la Recherche Scientifique (Canada), 2023 Publication	
10	www.frontiersin.org Internet Source	<1%
11	researchrepository.wvu.edu Internet Source	<1%
12	Submitted to University of Glasgow Student Paper	<1%
13	eprints.usm.my Internet Source	<1%
14	dspace.ut.ee Internet Source	<1%
15	www.ijraset.com Internet Source	<1%
16	www.mdpi.com Internet Source	<1%
17	ijrrt.com Internet Source	<1%
18	www.ncbi.nlm.nih.gov Internet Source	<1%
19	salford-repository.worktribe.com Internet Source	<1%
20	Submitted to De Montfort University Student Paper	<1%



30	Internet Source	<1%
31	Elizabeth A. Fulton. "Approaches to end-to-end ecosystem models", Journal of Marine Systems, 2010 Publication	<1%
32	P. Vasuki, Ujjwaleshwar Srikanth, Vijay Sankarnarayanan. "Chapter 36 Using Pre-trained Models forCode-Switched Speech Recognition", Springer Science and Business Media LLC, 2024 Publication	<1%
33	arcb.csc.ncsu.edu Internet Source	<1%
34	core.ac.uk Internet Source	<1%
35	era.ed.ac.uk Internet Source	<1%
36	vdocuments.pub Internet Source	<1%
37	www.mecs-press.org Internet Source	<1%
38	www.repository.smuc.edu.et Internet Source	<1%



39	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2023", Springer Science and Business Media LLC, 2023 Publication	<1%
40	Signals and Communication Technology, 2015. Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		