Available online at www.sciencedirect.com

ScienceDirect

Machine Learning with Applications 00 (2023) 1–17

Deep Learning

Crossing Language Identification: *Multilingual* ASR Framework Based on Semantic Dataset Creation & Wav2Vec 2.0

Or Haim Anidjar[a,b,c,d,*], Roi Yozevitch[a,**], Nerya Bigon[a], Najeeb Abdalla[a], Benjamin Myara[a], Revital Marbel[e,a,b,**]

[a]*School of Computer Science, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[b]*Ariel Cyber Innovation Center, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[c]*Kinematics and Computational Geometry Lab (K&CG), Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[d]*Data Science and Artificial Intelligence Research Center, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[e]*Faculty of Information Systems and Computer Science, College of Law & Business, David Ben-Gurion 26, 5257346, Ramat-Gan, Israel.*

Abstract

This study proposes an innovative methodology to enhance the performance of multilingual Automatic Speech Recognition (ASR) systems by capitalizing on the high semantic similarity between sentences across different languages and eliminating the requirement for Language Identification (LID). To achieve this, special bilingual datasets were created from the Mozzila Common Voices datasets in Spanish, Russian, and Portuguese. The process involves computing sentence embeddings using Language-agnostic BERT and selecting sentence pairs based on high and low cosine similarity. Subsequently, we train the Wav2vec 2.0 XLSR53 model on these datasets and assess its performance utilizing Character Error Rate (CER) and Word Error Rate (WER) metrics. The experimental results indicate that models trained on high-similarity samples consistently surpass their low-similarity counterparts, emphasizing the significance of high semantic similarity data selection for precise and dependable ASR performance. Furthermore, the elimination of LID contributes to a simplified system with reduced computational costs and the capacity for real-time text output. The findings of this research offer valuable insights for the development of more efficient and accurate multilingual ASR systems, particularly in real-time and on-device applications.

Keywords: Wav2Vec 2.0, Automatic Speech Recognition, Speech-2-Text, Transformers, Word Error Rate, Character Error Rate, Language Identification.

1. Introduction

Language barriers can result in negative outcomes in various fields, such as healthcare and law enforcement, due to hindered effective communication. For instance, in healthcare, communication gaps may lead to errors in diagnosis, treatment, and overall quality of care Al Shamsi et al. (2020); Steinberg et al. (2016). In law enforcement, officials might encounter suspects or witnesses who speak multiple languages during questioning, making it challenging to obtain accurate information Richardson et al. (2014); Wangaryattawanich et al. (2016).

Multilingual Automatic Speech Recognition (MLASR) models are systems designed to transcribe spoken language from multiple languages into written text. These models hold the potential to bridge these communication

[*] Corresponding author: Or Haim Anidjar, School of Computer Science, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.
[**] These authors have contributed equally to the paper.

Email addresses: orhaim@ariel.ac.il (Or Haim Anidjar), roiyo@ariel.ac.il (Roi Yozevitch), neryabigon@gmail.com (Nerya Bigon), Najeeb040@gmail.com (Najeeb Abdalla), benmyara13@gmail.com (Benjamin Myara), revi85@gmail.com (Revital Marbel) gaps, allowing individuals to speak their native languages while enabling healthcare providers and law enforcement officials to understand their needs and provide appropriate care or gather crucial information. Accurate MLASR models can significantly impact these settings, such as reducing the time taken for patient diagnosis in healthcare, thus saving valuable time for both doctors and patients. Relying on human translators is not always feasible and can be extremely expensive.

Automatic Speech Recognition (ASR) Juang & Rabiner (2005); Cutajar et al. (2013) systems has witnessed significant advancements in recent years due to the rise of machine learning and natural language processing techniques Li et al. (2022a); Mustafa et al. (2022); Kumar & Singh (2019); Chowdhury et al. (2021) This progress has resulted in highly accurate transcription of various spoken languages.

However, despite significant advancements in mono-lingual ASR models Juang & Rabiner (2005); Cutajar et al. (2013), creating an MLASR model presents several challenges. First, training requires large amounts of transcribed multilingual speech data, which is a scarce resource. Second, correctly performing language identification (LID) Berns et al. (2023); Chakravarthi et al. (2022) and Code Switching (CS) Jose et al. (2020) is a difficult task. Finally, building a multilingual model is a computationally complex task. This paper aims to overcome those difficulties while providing a model that can accurately transcribe audio-speech segments that include two or more languages.

Wav2Vec2 Baevski et al. (2020a,b), a Transformer-based model Lin et al. (2022), has emerged as a leading architecture for speech transcription due to its superior accuracy and its ability to accommodate diverse languages, dialects, accents, and speech styles. The model operates by processing *raw audio waveform* inputs to produce vector-based language representations.

Prior research on the Wav2Vec2 has primarily aimed to enhance its performance by means of diverse techniques, including fine-tuning the model, adjusting the model architecture or hyper-parameters, and integrating supplementary training objectives. Among such endeavors, one can find the Wav2Vec2-xlsr-53 Deschamps-Berger et al. (2022), which was trained on 53 languages and achieved state-of-the-art performance on a range of speech recognition tasks Shahgir et al. (2022). Additionally, the Wav2Vec2 has been used of for tasks such as speaker identification Malek et al. (2022), language identification Chakravarthi et al. (2022), and keyword spotting Ahmed et al. (2022), among others. The Wav2Vec2 architecture has garnered significant traction within the speech recognition field, and has exhibited notable efficacy across various use cases. The principal driver of its success is the utilization of Self-Supervised Learning. (SSL) Saeedi & Giusti (2023).

Typically, when training and fine-tuning a speech recognition model with a limited amount of data Thomas et al. (2020), the resultant model may lack robustness, resulting in a high error rate. Given the complexity of ASR models, effective learning requires access to extensive and heterogeneous datasets. Consequently, fine-tuning a model for MLASR is a complex endeavor, as incorporating a LID solution could potentially introduce performance degradation and thus should be avoided, especially in the use-case of low-resource languages Shor et al. (2019). Furthermore, training a model on a limited dataset may result in over-fitting, which manifests as satisfactory performance on the training data, but poor generalization to novel, unseen data. One possible solution to overcome this obstacle Alsayadi et al. (2021), is data augmentation Shahnawazuddin et al. (2020a,b).

The development of an MLASR solution is highly advantageous, notwithstanding the challenges that may arise during the implementation process; the development of MLASR models has opened up new possibilities in various fields Yadav & Sitaram (2022); Salesky et al. (2021); Choutri et al. (2022). One such application can be found in DEA investigations. When questioning suspects, the interrogator may not be proficient in the suspect's language, causing key details to be missed during the conversation Kramsch (2014); Richardson et al. (2014). Access to an accurate conversation transcription can be a valuable tool in these scenarios, saving the cost of hiring an experienced interpreter for every interaction. Another potential use case is in mediation between groups that speak different languages Dendrinos (2006). A single model that is trained on the recording of the mediation segment can produce a reliable transcription, benefiting both parties. The ability to review the entire conversation can increase trust and assist in resolving conflicts.

Creating an MLASR Tachbelie et al. (2022); Abate et al. (2021) model is a difficult task as previously mentioned. To address the lack of data, this paper proposes a novel approach to artificially create multilingual conversations with accurate labels derived from pre-existing mono-lingual datasets. To help mimic real-world conversations, agnostic sentence embedding was utilized Datta et al. (2020); Feng et al. (2020a). Agnostic sentence embedding is a method for representing sentences as fixed-length vectors in a way that captures their meaning. The term *'agnostic'* refers to the fact that this method does not rely on any prior knowledge or assumptions about the task the sentence will be used for, such as sentiment analysis or text classification.

By combining similar sentence embedding vectors, the MLASR model can learn to generalize across different languages and contexts, making it more effective at a transcribing speech in low-resource languages. In addition, to prevent errors in language detection and CS, the proposed approach proposes fine-tuning a single model with the combined vocabulary of two languages in the dataset, focusing on minimizing error rates without the need to solve for LID.

This study aimed to develop an MLASR model to improve the accuracy of speech recognition systems for any two languages, including low-resource languages, in various fields such as healthcare, law enforcement, and education. The methodology involved several steps, beginning with the creation of a dataset. The agnostic sentence embedding technique was used to process each sentence in two different mono-lingual datasets. Subsequently, a sentence from language 'A' was paired with a sentence from language 'B' that had the highest vector similarity. The resulting audio segments were combined, and their corresponding transcriptions were stitched together. Following the dataset creation, the combined dataset was utilized to fine-tune an MLASR model, aimed at enhancing its performance. The proposed approach seeks to demonstrate the feasibility of developing MLASR models for different languages and showcasing its effectiveness in enhancing the accuracy of speech recognition systems.

1.1. Our contribution

This paper offers a robust approach for fine-tuning the Wav2Vec2 speech recognition model for acoustic multilingual recognition. The contributions of this paper can be summarized as follows:

- **Overcome the need for Language Identification model.** The integration of a Language Identification (LID) model into a Multilingual Automatic Speech Recognition (MLASR) system can lead to computational performance degradation and accuracy loss, unless the LID model provides a 100% success rate (which is a surrealistic assumption). To overcome this issue, we propose an MLASR approach that employs intelligent semantic dataset creation, which allows for the use of a **single** ASR model for multilingual conversations without the need for a separate LID model. This approach eliminates the accuracy loss associated with the use of another statistical model for LID, and ensures that the system maintains optimal computational performance.

- **Different languages immunity.** The proposed MLASR solution is designed to handle various grammatical rules and syntax across different languages, which is a critical aspect of efficient natural language processing systems. This adaptability enables seamless cross-lingual communication and knowledge sharing.

- **Language Change Detection.** One of the main concerns in MLASR is the potential for output errors due to pronunciation differences across languages. Guttural consonants, such as *'r','b','z'*, often sound different in different languages, while vowel letters (i.e., *'a,e,i,o,u'*) may sound the same. For instance, the combination of letters *'gui'* (appears in *guitarra*) in Spanish does not produce a typical *'U'* vowel sound. This can lead to incorrect output from the Wav2Vec2 model, producing a mixture of letters from different languages. While a Language Change Detection model could address this issue, such a model is not recommended due to computational performance degradation. However, our semantic dataset creation technique, incorporated into the Wav2Vec model and its loss function, effectively prevents such output errors. This contribution is detailed in Section 4.

- **Sentimental-Semantic relationship.** This paper presents a central hypothesis, which posits that there exists a relationship between the semantics of phrases from different languages and the acoustics of their speech signals. Specifically, it is argued that the magnitude and tonation of the speech signal reflect the semantic relationships between phrases from different languages. An experimental evaluation (Section 5) is conducted to examine this hypothesis. The results of this evaluation demonstrate that concatenating pairs of semantically correlated phrases from different languages leads to lower error rates compared to non-semantically correlated pairs.

- **Coping with real-life dataset.** We propose a pragmatic solution for the low data availability problem in ASR. Specifically, our approach employs data augmentation techniques to enhance the robustness and reliability of the transcription process. By adopting this approach, it is possible to mitigate the effects of low data availability and pave the way for more effective and accurate ASR in a variety of contexts. Based on our research, it appears that a bilingual dataset of this nature has not yet been established. Thus, we have employed innovative semantic techniques to generate a bilingual dataset, which serves to replicate situations where two distinct languages are spoken.

- **Improving CER for languages with the same vocabulary.** This study's results showcase that the proposed approach can be an effective solution for achieving high CER for languages with limited datasets. Additionally, it presents a useful technique for training the Wav2Vec2 model on underrepresented languages that share similar alphabetic characters, acting as a data augmentation method to increase the size of the training dataset.

1.2. Paper Structure

The remainder of this paper is structured as follows: Section 2 surveys related work mainly regarding common MLASR approaches in general, and the Wav2Vec2 architecture in particular; Section 3 discusses the dataset used in this paper and the pre-processing procedure; Section 4 presents the approach employed in this paper as part of the Wav2Vec2 architecture exploitation; Section 5 presents a fine-grained experimental evaluation process, that consists of an evaluation of our augmentation-based ASR approach, and the results of the comparison between our approach and a Wav2Vec2 baseline version; Finally, Section 6 provides a fundamental discussion, concludes and summarizes this paper. For ease of reading, Table 1 provides a list of abbreviations that are commonly used in this paper.

Abbreviation	**Meaning**
ASR	Automatic Speech Recognition
CER	Character Error Rate
CNN	Convolutional Neural Network
CS	Code Switching
CTC	Connectionist Temporal Classification
GELU	Gaussian Error Linear Unit
LID	Language Identification
MLASR	Multilingual ASR
NLP	Natural Language Processing
SR	Speech Recognition
SSL	Self Supervised Learning
WER	Word Error Rate

Table 1. List of common abbreviations used in this paper.

2. Related Work

LID Berns et al. (2023); Chakravarthi et al. (2022) is an essential component of ASR systems that aims to identify the language spoken in a given audio signal. It requires recognizing subtle differences in the acoustic features of speech signals. Several approaches have been proposed to address this problem over the years. One of the earliest approaches to LID is the acoustic feature-based method. This method uses statistical analysis of speech signals to identify the spoken language. Acoustic features such as pitch, formants, and energy in speech signals are analyzed to identify speech-specific patterns indicative of a particular language. This method has been studied extensively in the literature, and several studies have shown that it can achieve high accuracy in identifying the language of speech signals Zissman (1996); Zissman & Berkling (2001); Muthusamy et al. (1994). However, these acoustic feature-based methods have some limitations. Firstly, they are limited in their ability to detect many languages. They can only detect a small number of languages accurately, and their performance may degrade when dealing with less commonly spoken languages. Secondly, these methods are less accurate than more advanced ones that utilize deep learning-based approaches. These methods leverage deep neural networks Dominguez (2023) that can learn complex representations of speech signals, resulting in higher accuracy in language identification tasks.

The field of ASR has benefited significantly from recent advances in deep learning-based methods. One such approach involves using neural networks based on Mel Frequency Cepstral Coefficients (MFCCs) and LID models

Reda & Aoued (2005). MFCCs are commonly used in ASR systems to represent the spectral envelope of speech signals. They capture the frequency distribution of the audio signal and provide a compact representation that can be used as input to a neural network. The LID model is trained on a large dataset of speech samples from different languages to learn language-specific patterns. Once trained, the model can identify the language of a given speech segment with high accuracy. Another promising approach to language identification involves using Convolutional Neural Networks (CNNs) Saeedi & Giusti (2023). CNNs have been extensively used in image recognition tasks, and their application has also been adapted to speech recognition. In this approach, CNNs extract local features from spectrogram representations of speech signals and use these features to identify the spoken language.

Spectrograms are visual representations of speech signals that show the intensity of different frequencies at different time intervals. These representations can be fed into a CNN, which can learn to recognize patterns in the spectrograms indicative of specific languages. Recent studies have shown that CNNs can accurately identify languages from speech signals. For example, Singh et al. (2021) achieved a 98 percent accuracy rate in identifying 22 languages using a CNN approach. This demonstrates the potential of deep learning-based methods for improving the accuracy of ASR systems and advancing the field of speech recognition. Overall, these approaches highlight the importance of leveraging advances in deep learning and machine learning to enhance the performance of ASR systems and make them more robust and effective for a wide range of applications.

CS is an aspect of LID research where speakers switch between two or more languages in a single utterance. CS is a common phenomenon in multilingual communities and challenges language identification models. Several studies have proposed methods for identifying CS in speech, including neural network-based approaches and unsupervised clustering methodsLuo et al. (2018); Samih et al. (2016) Despite the advances in language identification techniques. There are still challenges in accurately identifying languages in CS speech. One of the challenges is the need for training data for CS speech Mustafa et al. (2022); Datta et al. (2020), which can limit the effectiveness of models trained on monolingual data. In addition, CS can vary by speaker, context, and language combination, making it challenging to develop a unified model for CS detection Mustafa et al. (2022). Our approach can increase the amount of available CS data so other researchers can benefit.

In recent years, there have been significant developments in ASR technology, with the emergence of E2E ASR systems among the most important. These systems can convert speech directly into text without intermediate representation. E2E ASR models are becoming increasingly popular due to their simplicity and improved performanceLi et al. (2022b); Mustafa et al. (2022); Vanderreydt et al. (2022a) Several studies have been conducted to explore the effectiveness of E2E models for multilingual ASR, including the work of Li et al. (2022b), whose proposed Transformer-based E2E ASR system achieved top performance on a multilingual ASR task.

Pre-training techniques have great potential for improving ASR performance. Among these techniques, Wav2Vec 2.0 has received special attention in recent years. Wav2Vec 2.0 is a method that learns speech representations from large amounts of unlabeled data and fine-tunes them on downstream ASR tasks, based on the attention Javeed (2023); Ren et al. (2022) mechanism, transfer-learning and SSL Saeedi & Giusti (2023). This leads to significant improvements in ASR performance, as shown by several studies, including multilingual ASR. Multilingual ASR is the task of recognizing speech from multiple languages with a single model. This task is challenging due to the diversity and complexity of different languages and domains. For example, Choi & Park (2022) proposed a Wav2Vec 2.0-based multilingual ASR system that achieved peak performance on several benchmark datasets across different languages and domains. They also showed that their system could generalize well to unseen languages and domains using zero-shot and few-shot learning techniques. Feng et al. (2020a) and Chuang et al. (2020) also explored multilingual ASR systems based on Wav2Vec 2.0 and other pre-training methods. They compared different strategies for pre-training, fine-tuning, and decoding multilingual speech data and analyzed the factors affecting multilingual ASR systems' performance.

Word Habbat et al. (2021); Anidjar et al. (2021b) and sentence Feng et al. (2020b) embedding techniques are commonly used in ASR to capture the semantic and syntactic information of words and sentences. These embeddings effectively improve ASR performance by providing additional features for the ASR model. Various embedding techniques have been explored, including word and sentence-level agnostic embeddings. Word-level agnostic embeddings Feng et al. (2020b) do not depend on the specific language or domain of the words. In contrast, sentence-level agnostic embeddings do not depend on the specific language or domain of the sentences. These agnostic embeddings can help improve multilingual ASR by enabling cross-lingual and cross-domain transfer learning Ford et al. (2022) and reducing the data requirements for low-resource languages and domains. For example, Choi & Park (2022) used sentence-level agnostic embeddings based on multilingual BERT to enhance their Wav2Vec 2.0 Anidjar et al. (2023a) based multilingual ASR system. They showed that these embeddings could improve the performance of zero-shot and few-shot learning scenarios. Feng et al. (2020a) used word-level agnostic embeddings based on FastText Anidjar et al. (2020, 2021a) to improve their multilingual ASR system. They showed that these embeddings could improve the performance of low-resource languages and domains. Chuang et al. (2020) used word-level and sentence-level agnostic embeddings based on multilingual BERT Karthikeyan et al. (2020) and XLM-R to improve their multilingual ASR system. They showed that these embeddings could improve the performance of high-resource and low-resource languages and domains.

3. Datasets

The present study recognizes the significance of developing MLASR models that can cater to the needs of diverse and multilingual populations. However, creating such models demands an extensive and varied training dataset. The Mozilla Common Voice Ardila et al. (2019) project offers a solution to this challenge by providing an open-source, multi-language dataset of voices that can be utilized to train speech-enabled applications. The Common Voice dataset, currently the most extensive publicly available resource, offers a wide range of voices of various ages, genders, accents, and speaking styles. It also comprises diverse transcriptions such as informal conversations, news articles, and public service announcements. Currently, the dataset includes 17,690 validated hours in 108 languages, with more voices and languages frequently added to the project. In this study, the Common Voice datasets for Russian, Portuguese, and Spanish languages were employed to create training datasets for the development of a multilingual ASR model, as presented in Table 2.

Monolingual datasets statistics - Russian, Portuguese and Spanish			
Ent.	Language	Hours of community-validated audio (≈)	# of unique speakers
1	Russian	209	2,901
2	Portuguese	151	3,099
3	Spanish	482	25,096

Table 2. Statistics of hours of community-validated audio, and amount of unique speakers in the Mozilla Common Voice Ardila et al. (2019) datasets in Spanish, Russian, and Portuguese.

The Common Voice datasets are widely employed to train and assess ASR models, particularly for low-resource languages and domains. Apart from offering substantial amounts of publicly available voice data, the Common Voice datasets incorporate demographic metadata, including age, gender, and origin, which can aid in personalizing ASR models or investigating the impact of speaker variability on speech recognition. Notably, these datasets have facilitated the development of various ASR models, such as SpeechBrain Ravanelli et al. (2021), Project Euphonia Clark et al. (2020), and Whisper Radford et al. (2022). However, while the existence of diverse and extensive datasets like Common Voice is crucial for creating multilingual ASR models that cater to diverse populations, it is not sufficient. Since there are minimal pre-existing datasets that contain code-switching Jose et al. (2020), researchers must create synthetic datasets themselves, including the present study. In the absence of a large-scale, open-source collection of such datasets, each researcher must develop their unique dataset, limiting the comparability of results among researchers in the MLASR domain. The lack of a standard benchmark for evaluation can lead to distorted outcomes influenced more by the datasets than the framework employed.

4. Framework

4.1. Dataset Creation and Augmentation

In this study, we employed a custom synthetic multilingual dataset to train a Wav2vec 2.0 XLSR53 model for transcribing code-switching (CS) speech audio. Our dataset was created by augmenting and combining existing speech recording datasets in Spanish, Russian, and Portuguese, with the aim of leveraging the cross-lingual transfer learning capabilities of the Wav2Vec 2.0 XLSR53 model, which is a state-of-the-art self-supervised speech recognition model that can learn from unlabeled speech data in multiple languages. To prepare our dataset, we utilized a novel approach based on sentence embeddings and data augmentation techniques Shahnawazuddin et al. (2020a); Ahmed et al. (2023).

Since there is currently no standardized benchmark for evaluating MLASR model performance across different languages and domains, researchers often construct their datasets. To create our dataset, we began with the Mozilla Common Voice datasets in Spanish, Russian, and Portuguese, which offer high-quality and diverse speech data covering a wide range of topics and accents. However, these datasets do not contain CS speech, which is essential for training a robust MLASR model from scratch. Therefore, we augmented these datasets by generating synthetic speech data by combining them.

4.2. Sentence Embeddings and Cosine Similarity

The first step in our process was to calculate the sentence embedding for each dataset using Language-agnostic BERT (LaBSE) Feng et al. (2020b), a pre-trained model that can produce high-quality sentence embeddings for 109 languages Feng et al. (2020a). A sentence embedding is a vector representation of a sentence that captures its structure and meaning. This technique is commonly used to cluster sentences based on their themes or measure their semantic similarity. To convert a sentence to a vector representation using LaBSE, several steps are involved. First, the input sentence is tokenized into subword units using WordPiece tokenization Kawazoe et al. (2021). Then, the pre-trained LaBSE model, which employs a transformer-based Rodrawangpai & Daungjaiboon (2022) neural network architecture, is used to encode the sequence into a fixed-length vector representation. The hyperparameters of the model determine the size of the vector. In the original LaBSE model, the output vector size is 768, meaning that each sentence is encoded into a fixed-length vector of 768 dimensions. The final hidden state of the *CLS* token, a special token added to the beginning of the input sequence, is used to obtain this vector representation, which captures the sentence's syntactic and semantic properties. The resulting sentence embedding can be applied to various Natural Language Processing (NLP) Khodadadi et al. (2022); Casola et al. (2022) tasks, including information retrieval, sentiment analysis, and text classification.

Once the sentence embeddings were obtained, the cosine similarity between each pair of sentence embeddings was calculated using the cosine similarity formula. The formula for calculating the cosine similarity between two vectors, x and y, is defined as in Eq.(1):

$$Cosine - Similarity = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{1}$$

where '$\cdot$' represents the dot product of the two vectors, $||x||$ and $||y||$ represents the magnitude or length of the respective vectors. The cosine similarity metric measures the cosine of the angle between two vectors and ranges from -1 to 1, with 1 indicating that the vectors are identical and -1 indicating that they are completely dissimilar. A cosine similarity score of 0 indicates that the two vectors are orthogonal. To calculate the cosine similarity between the two different language datasets, the sentence embeddings for each dataset are compared pairwise, resulting in a matrix of cosine similarity scores. We took all the sentences with a cosine similarity score above 0.75, and all the pairs with a cosine similarity of under 0.25. These pairs of sentences were combined into a single dataset by concatenating the audio segments and their respective transcriptions. During the concatenation process, we removed all special characters from the transcription and converted the audio files to Unicode. We repeated this process for the following languages; Russian-Portuguese, Russian-Spanish, and Spanish-Portuguese. Resulting in two datasets for each, one for low cosine similarity, and one for high cosine similarity. For the Russian-Portuguese dataset, we ended up with $14,000$ sentences for the high cosine variant and $14,000$ for the low cosine variant. For the Russian-Spanish dataset, we ended up $14,000$ sentences for the high cosine variant and $14,000$ for the low cosine variant. For the Spanish-Portuguese dataset, we ended up with $14,000$ sentences for the high cosine variant and $14,000$ for the low cosine variant.

4.3. Training the Wav2vec 2.0 XLSR53 Model

Once we finished the pre-processing phase, we trained the Wav2vec 2.0 – XLSR53 model Deschamps-Berger et al. (2022); Vanderreydt et al. (2022b) separately on each of our custom bilingual datasets. The Wav2Vec2-xlsr-53 model, was pre-trained on 53 languages by the Facebook AI Research team. It is a state-of-the-art approach for converting raw audio waveforms into high-quality text representations. The model is based on semi-supervised learning, where the model learns to predict missing segments of the input waveform. We used the Connectionist Temporal Classification (CTC) Higuchi et al. (2022); Sailor et al. (2021) loss function which is a popular choice for speech recognition tasks. This loss function works by predicting the probability distribution over all possible output sequences given an input audio waveform. By mapping the output sequence to a blank symbol and collapsing repeated symbols, the model is able to handle variable-length inputs and outputs. The big advantage of our framework is the absence of LID. Being LID-free has several advantages. Firstly, it allows for a truly multilingual experience where users can interact in any language interchangeably without explicitly setting the language of the conversation. Secondly, it eliminates the need for the LID model, which can be cumbersome to maintain and not suitable for on-device applications as the model size increases linearly with the addition of languages. Thirdly, it reduces the computational cost as the forward propagation through the shared encoder is performed only once for all languages, whereas in the case of monolingual models, each individual model has to perform a separate forward propagation per language. Finally, it enables real-time text output.

4.3.1. Wav2Vec2 Architecture

The Wav2Vec2 system comprises a convolutional feature encoder that is organized into multiple layers. Specifically, this encoder consists of a temporal convolution, which is followed by a normalization layer and a Gaussian Error Linear Unit (GELU) Hendrycks & Gimpel (2016) activation function. The encoder's total stride is responsible for determining the number of time steps T that are used as input to the Transformer. The resulting contextualized speech representations are generated by the Transformer. Subsequently, the output of the feature encoder is fed into a context network that follows the same architecture as the transformer in Devlin et al. (2018). Finally, the training process for Wav2Vec2 involves the use of the CTC Higuchi et al. (2022) loss function, given that the task at hand is a sequence alignment problem. Notably, instead of relying on fixed positional embeddings that encode absolute positional information, Wav2Vec2 utilizes a convolutional layer that effectively functions as a relative positional embeddings Devlin et al. (2018). This represents a key modification to the conventional approach.

CTC loss-function. In the context of ASR systems, achieving accurate alignment between individual characters and their corresponding locations in an audio recording is a challenging task. To address this, the CTC loss function is employed. This loss function operates by computing the difference between a continuous and unsegmented acoustic time-series signal data sample and a target sequence-based label that consists of characters. The computation is carried out by summing over the probability distribution of possible alignments between the speech signal and the textual sequence label, which results in a loss value that is differentiable with respect to each input node. Notably, the alignment between the speech signal and the textual sequence label is assumed to be ”many-to-one,” which places a constraint on the length of the textual sequence label, requiring it to be the same length as the input. In this paper, we hypothesize that fine-tuning the model on augmented data using the CTC loss function will lead to improved performance compared to fine-tuning on clean data alone.

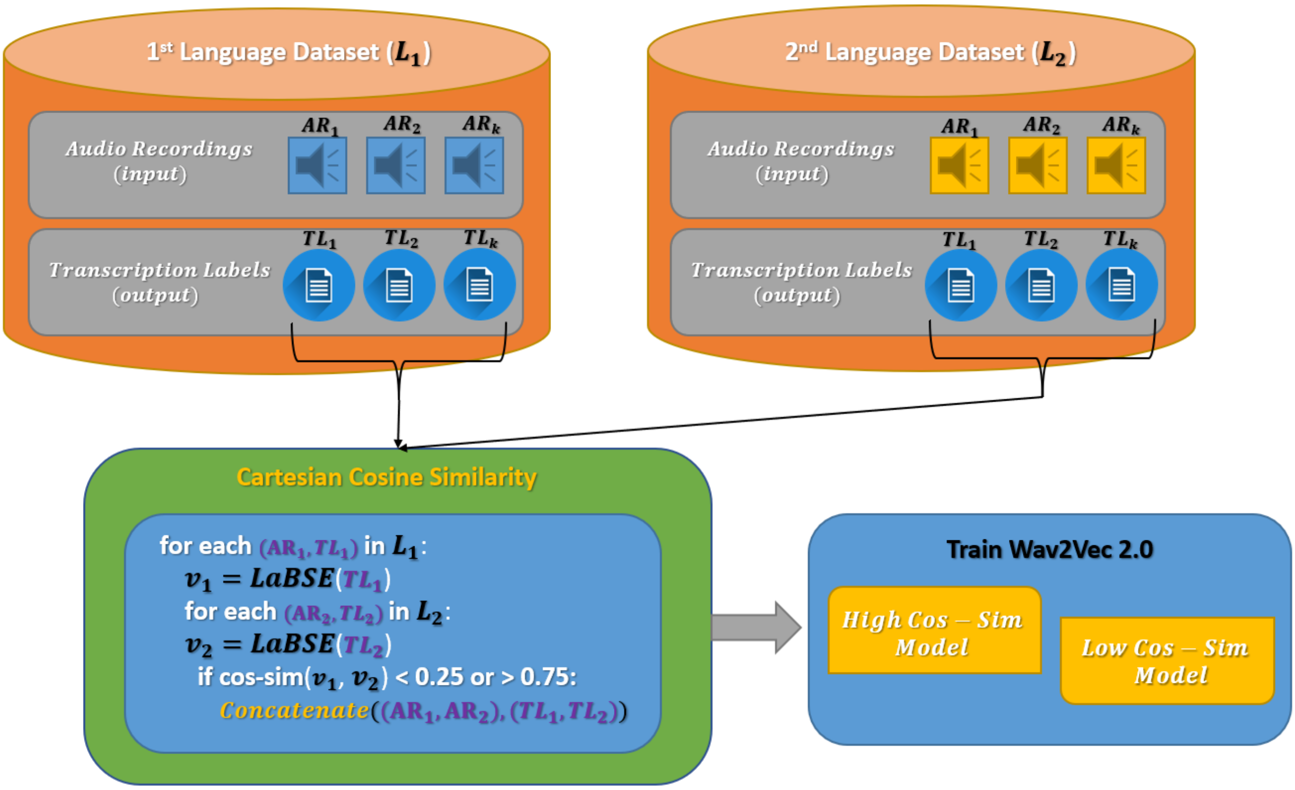
4.4. Evaluation and Comparison

For the final evaluation stage, we applied standard metrics such as Word Error Rate (WER) Deléglise et al. (2009); Anidjar et al. (2023b) and Character Error Rate (CER) Hou et al. (2020); Kumar et al. (2022) on the test set to assess the performance of our model. CER is a more reliable indicator of the model's capabilities than WER. CER measures the number of character-level errors in the transcription, regardless of the length and complexity of the words in the text. WER, on the other hand, counts the number of word-level errors, which can be skewed by long or rare words that are prone to be misrecognized or misaligned by the model. This difference becomes more pronounced when dealing with multiple languages that have different word structures and vocabularies. Moreover, we compared the performance of each model to its corresponding co-sign similarity model, which is a variant that matches sentences with similar meanings in different languages. This comparison allowed us to evaluate the effect of semantic alignment on the quality of the transcription. Finally, one can note in Figure 1 an illustration of the whole framework proposed in this paper, and described in this section.

5. Experimental Evaluation and Results

5.1. Evaluating ASR System Performance

Two widely used metrics for assessing the performance of ASR systems are WER and CER. These metrics determine the accuracy of the system's transcriptions by comparing them to a reference transcription of the input audio.



```
for each $(AR_1, TL_1)$ in $L_1$:
  $v_1 = LaBSE(TL_1)$
  for each $(AR_2, TL_2)$ in $L_2$:
  $v_2 = LaBSE(TL_2)$
    if cos-sim$(v_1, v_2)$ < 0.25 or > 0.75:
      $Concatenate((AR_1, AR_2), (TL_1, TL_2))$
```

Figure 1. An illustration of the framework presented in this paper. At first, a custom synthetic multilingual dataset is created, by combining existing speech recording datasets out of a subset of two languages (L_1, L_2) from the following three: Spanish, Russian, and Portuguese. Next, the sentence embeddings computed for each dataset using Language-agnostic BERT (LaBSE) Feng et al. (2020b,a), a pre-trained model that can produce high-quality sentence embeddings for 109 languages. Once the sentence embeddings were obtained, the cosine similarity between each pair of sentence embeddings was calculated using the cosine similarity formula (Eq.(1)). Then, all the sentences with a cosine similarity score above 0.75, or under 0.25 are considered for training the Wav2Vec2 model. These pairs of sentences were combined into a single dataset by concatenating the audio segments and their respective transcriptions. The resulting datasets were then used to train the Wav2vec 2.0 XLSR53 model. Eventually, two models are trained separately on each of the custom bilingual dataset; one model for the *High* cosine-similar pairs of sentences (i.e. with cosine score above 0.75), and another model for the *Low* cosine-similar pairs of sentences (i.e. cosine score under 0.25).

The key difference between WER and CER lies in their units of measurement: WER focuses on the number of erroneous *words* in the system's transcriptions, whereas CER considers the number of incorrect *characters*. The formula for CER is as follows (Eq.(2)):

$$CER = \frac{(I + S + D)}{N} \times 100 \tag{2}$$

where:

- I denotes the number of insertions (characters present in the system's transcription but absent in the reference transcription).
- S denotes the number of substitutions (characters in the system's transcription that differ from the corresponding characters in the reference transcription).
- D denotes the number of deletions (characters present in the reference transcription but absent in the system's transcription).
- N denotes the total number of characters in the reference transcription.

The formula for WER is quite similar to that of CER, as shown below (Eq.(3)):

$$WER = \frac{(I + S + D)}{N} \times 100 \tag{3}$$

where:

- I denotes the number of incorrect words in the system's output.
- S denotes the number of words that were correctly identified but appeared out of sequence in the system's output.
- D denotes the number of words removed from the reference transcription.
- N denotes the total number of words in the reference transcription.

5.1.1. Comparing WER and CER

Although WER is more widely used than CER due to its ease of comprehension and interpretation, CER can be advantageous in certain scenarios. For instance, CER is particularly helpful when transcriptions involve proper nouns or words with challenging spellings, or when homophones (words that have identical pronunciations but different spellings) are present in large numbers. Additionally, if the system consistently fails to recognize spaces between words, the difference between the results obtained from CER and WER metrics will be minimal. For example, consider the following two Spanish sentences:

> **Reference**: *'El misterio de* ***ochenta años*** *de* ***antigüedad****: ¿por qué* ***la corona del*** *Sol es más caliente* ***que su superficie****?'*
>
> **Prediction**: *'El misterio de* ***ochentaños*** *de* ***antiguo edad****: ¿por qué* ***lako rona de el*** *Sol es más caliente* ***queso super ficie****?'*

The WER between the reference and prediction sentences is 57.9%, while the CER between them is only 12.5%. Therefore, the entire results section in Section 5.2, reports both the WER and CER for all experiments conducted.

5.2. Results

This study presents a series of experiments designed to evaluate the performance of our proposed model across different language pairs and various levels of cosine similarity. Specifically, we conducted experiments using the Russian, Portuguese, and Spanish languages and focused on assessing the impact of high and low-similarity samples on the system's performance.

To carry out the experiments, we used a training set of 10,000 samples for each language pair. In addition, we created validation and test sets, each containing 2,000 samples, to ensure that our results were reliable. We conducted a total of six training sessions, two for each pair of languages, to evaluate the impact of high and low cosine similarity on the performance of our model. During each training session, we focused on training our model on samples with either high cosine similarity or low cosine similarity. To further evaluate the performance of our model, we performed four additional experiments where we trained the high model on low data and vice versa. Each model was trained for ten epochs.

Our results showed that the performance of our model varied significantly across different language pairs and levels of cosine similarity. Specifically, we found that the model performed better on language pairs with higher cosine similarity, which is one main hypothesis of this paper, and its performance degraded as the level of similarity decreased. Furthermore, our cross-check experiments indicated that our model's performance was affected by the level of cosine similarity, highlighting the importance of training our model on a diverse range of samples. Overall, the experiments demonstrate the efficiency of the proposed framework in handling different language pairs and levels of cosine similarity, providing a useful tool for various natural language processing applications.

5.2.1. Monolingual and Multilingual Comparison

In this section, we present the results of our experiments and compare them with the baseline results obtained from models only trained on *Monolingual* languages. The results of the *Monolingual* comparison are shown in Table 3. Next, the *Multilingual* training on language pairs, yields comparable results to the models trained on *Monolingual* languages, as demonstrated in Tables 4, 5, and 6. These findings suggest that our model trained on language pairs is as accurate and usable as the model trained on a corresponding *Monolingual* language.

The performance of the pair of languages is not negatively affected and is comparable to that of the models trained on individual languages, despite the assumption that the pair of languages should have lower accuracy. We attribute this to the fact that the pairs of languages have been trained on a larger amount of data and for a longer period of time. As a result, the need for a language recognition element is eliminated, which simplifies the model and reduces computational costs.

Our experimental results strongly support our thesis that building the dataset from samples with high similarity significantly improves the model's performance. In each language pair, the models trained on high-similarity samples consistently outperformed those trained on low-similarity samples.

Monolingual datasets training - Russian, Portuguese and Spanish			
Ent.	Language	WER [%]	CER [%]
1	Russian	35.82	10.29
2	Portuguese	27.07	9.23
3	Spanish	20.00	7.00

Table 3. The results of the models trained on *Monolingual* datasets for (i) Russian, (ii) Portuguese and (iii) Spanish are presented in terms of their respective CER and WER.

In Table 4 Entry 1, a comparison between the *multilingual* model, and two models trained on *monolingual* datasets is presented for the language pair of Russian and Portuguese. The *multilingual* model yields a WER of 26% and a CER of 8%, indicating a smaller difference in accuracy compared to the *monolingual* models, which have WER of 35.8% and 27%, and CER of 10.2% and 9% for Russian and Portuguese, respectively. These findings suggest that the *multilingual* model can effectively handle both languages concurrently, providing comparable CER and WER to models trained on *monolingual* datasets.

***Mono**lingual* and ***Multi**ingual* comparison - Russian & Portuguese			
Ent.	Language(s)	WER [%]	CER [%]
1	Russian & Portuguese	**26.22**	**8.21**
2	Russian	35.82	10.29
3	Portuguese	27.07	9.23

Table 4. The performance of each language (Russian & Portuguese) trained in a *monolingual* manner is compared to the performance of the language pairs trained together. One can observe that the WER and CER gap between the two approaches is negligible.

Next, Ent.(1) in Table 5 presents a comparison between the language pair of Spanish and Russian and the individual languages. The multilingual model achieves a CER of 10% and a WER of 28%, while the individual language models have WER of 35.8% and 20% and CER of 10.2% and 7% for Russian and Spanish, respectively. These results indicate that the multilingual model performs similarly to the individual language models, thus supporting the efficacy of our approach in handling multiple languages.

Finally, the performance of the multilingual model to that of models trained on monolingual datasets is presented for Portuguese & Spanish. Table 6 shows the results of the Spanish and Portuguese language pair in comparison to each language trained monolingually. The multilingual model achieves a CER of 7% and a WER of 19%, whereas the monolingual models have a WER of 27% and 20% and a CER of 9% and 7% for Portuguese and Spanish, respectively. These findings suggest that the multilingual model is just as effective as models trained on monolingual datasets, with a negligible gap in CER and WER between the monolingual datasets and the language pair.

Mono*lingual* and ***Multi****ingual* comparison - Russian & Spanish			
Ent.	Language(s)	WER [%]	CER [%]
1	Russian & Spanish	28.94	10.00
2	Russian	35.82	10.29
3	Spanish	**20.00**	**7.00**

Table 5. Comparison of each language (Russian & Spanish) trained in a *monolingual* manner vs. multilingual training: negligible gap in CER and WER between languages, whenever trained together.

Mono*lingual* and ***Multi****ingual* comparison - Spanish & Portuguese			
Ent.	Language(s)	WER [%]	CER [%]
1	Portuguese & Spanish	**19.91**	7.47
2	Portuguese	27.07	9.23
3	Spanish	20.00	**7.00**

Table 6. Performance comparison between monolingual and paired language training, demonstrating reduced CER and WER gap between monolingual and paired models.

5.2.2. Multilingual Comparison of High-Similarity & Low-Similarity Based Models

In this section, we present the results of training models on three *Multilingual* semantic datasets: (i) Russian & Portuguese; (ii) Russian & Spanish; and (iii) Spanish & Portuguese.

Table 7 shows the results of the models trained and tested on high-similarity data, while Table 8 shows the performance of the same models trained on high-similarity data and tested on low-similarity data. The results demonstrate that selecting data with high similarity is critical in achieving accurate and reliable change detection, which is our third contribution of Language Change Detection. For example, Ent.(1) in Table 7 shows that the Russian and Portuguese language pair achieved a CER of 8% and a WER of 26% when trained and tested on high-similarity data. In contrast, Ent.(1) in Table 8 shows that when the same language pair was trained on high-similarity data and tested on low-similarity data, the CER performance degraded by 112.5%, and the WER degraded by 130.8%. Similarly, Ent.(3) in Table 7 shows that the Spanish and Portuguese language pair achieved a WER of 19% and a CER of 7% when trained and tested on high-similarity data. In contrast, Table 8 shows that the model's performance declined when the same language pair was trained on high-similarity data and tested on low-similarity data, with a 68.6% increase in WER and a 66.7% increase in CER.

Our findings emphasize the crucial role of high-similarity data in ensuring precise and reliable language change detection, which further strengthens the relevance of our research contributions.

Comparison between *Multilingual* datasets **trained** on high-similarity and low-similarity models					
Ent.	Language Pair	High Similarity		Low Similarity	
		CER [%]	WER [%]	CER [%]	WER [%]
1	Russian & Portuguese	**8.21**	**26.22**	13.53	50.93
2	Russian & Spanish	**10.00**	**28.94**	12.90	43.44
3	Spanish & Portuguese	**7.47**	**19.91**	10.43	33.34

Table 7. Performance comparison of models trained and tested on data with different levels of similarity within the same language pairs. The results show the impact of similarity on the models' CER and WER metrics, highlighting the differences between high-similarity and low-similarity data. This analysis provides insights into the models' ability to handle variations in the input data and sheds light on the importance of training and testing on data with varying levels of similarity.

The results in this section found that the optimal performance of multilingual ASR models is achieved when models are trained and tested on data with similar levels of similarity. Specifically, we observed that models trained on high-similarity data yielded the best performance when tested on high-similarity data. On the other hand, models trained on low-similarity data performed better when tested on high-similarity data than models trained on high-similarity data and tested on low-similarity data. These findings support the importance of high-similarity data for multilingual ASR system success. In addition, it has been also shown that removing language identification and using high-similarity data significantly improved the performance of the ASR models; sentences with high semantic

Comparison between *Multilingual* datasets **inferenced** on opposite-similarity models					
Ent.	Language Pair	High on low [a]		Low on high [b]	
		CER [%]	WER [%]	CER [%]	WER [%]
1	Russian & Portuguese	17.88	60.95	**13.27**	**42.16**
2	Russian & Spanish	14.67	47.63	**10.18**	**33.68**
3	Spanish & Portuguese	11.87	36.67	**9.83**	**28.29**

Table 8. Results of inference language models on high-similarity data on low-similarity data, and vice versa, demonstrating the impact of similarity level on the model's performance. Specifically, the models were trained on high-similarity data and tested on low-similarity data and vice versa. The table showcases the differences in CER and WER metrics when models are exposed to mismatched similarity levels, indicating the importance of training and testing the models on similar data.
[a] High semantic similarity inference on low semantic similarity trained model, [b] Low semantic similarity inference on high semantic similarity trained model.

similarity tend to have similar intonation patterns, which the ASR models can leverage to better transcribe speech. This is supported by our CER and WER metric results, which consistently showed improvements when models were trained on high-similarity data.

For instance, consider the following sentences which have high similarity:

Reference: *’se encuentra en argelia libano portugal y espana o reino unido de portugal brasil e algarves’*

Prediction: *se encuentra en argelia libano portugal y espana o reino unido de portugal brasil e algarves’*

One can note that the prediction is perfectly accurate in terms of WER(0%) and CER(0%), in contrast, consider the following sentences:

Reference: *’conocer los* ***productos en los que se ha*** *usado* ***amianto*** *puede* ***ayudar a identificarlo*** *presidente* ***prudente’***

Prediction: *’conocer los* ***prologuicos la secidad*** *usado* ***a mianto*** *puede* ***a ludar allo n didicamino*** *presidente* ***pordente’***

In this example, the model's prediction was significantly poorer, with a WER of 81.25% and a CER of 31.68%. It is noteworthy that the model struggled more with predicting these sentences.

Our experiments reveal that ASR models tend to perform better on language pairs that share similar linguistic features such as vocabulary, grammar, and phonetics. The results suggest that the combined vocabulary size of the ASR model is smaller for similar languages, which makes it easier for the model to identify the correct phonemes since there are fewer options to choose from. This observation highlights the potential of leveraging linguistic similarity to improve the performance of multilingual ASR systems.

In addition to the improved accuracy, our approach has the advantage of not relying on a separate language identification model. Language identification models can be computationally expensive and difficult to maintain, especially when dealing with multiple languages. Our approach eliminates the need for language identification models, reducing the complexity and computational cost of the ASR system. This can be particularly useful for real-time applications and on-device implementations, where computational resources are often limited. By leveraging linguistic similarity and eliminating the need for language identification models, our approach demonstrates a promising way to improve the performance and reduce the complexity of multilingual ASR systems.

5.2.3. Data Augmentation Based Model

Data augmentation Shahnawazuddin et al. (2020a); Ahmed et al. (2023) is a widely used technique to expand datasets artificially by generating altered versions of existing data. Its application is particularly common in deep learning, where it enhances the performance and generalization of models by providing additional training examples.

To successfully enhance the model, it is important to focus on a small number of high-impact augmentations Temraz & Keane (2022). On one hand, inadequate augmentation may result in over-fitting due to the limited diversity of the data. On the other hand, over-augmentation can lead to under-fitting, as essential data may be lost, making the words unrecognizable. Thus, augmenting the data by increasing the variance and robustness of the ASR system is crucial for achieving any significant improvement. In this study, three augmentation techniques are being explored:

- **Band-Stop -** Thai et al. (2019) an audio signal can be modified by removing a specific frequency range using a Band-Stop augmentation, which is a form of data augmentation. To achieve this, a band-stop filter is usually applied, which attenuates frequencies within a certain range while allowing those outside it to pass through. The range of frequencies that humans can hear falls between $0 - 4000Hz$, so we used this range as a minimum and maximum threshold. Additionally, we determined the min/max cutoff range through the bandwidth fraction, which represents the relative portion of the frequency spectrum to cut-off, expressed between $0\% - 200\%$ and centered around the frequency spectrum's center frequency. Lastly, we set the steepness of the cutoff in dB. The primary objective of using this augmentation technique is to simulate various accents artificially.

- **Gaussian-Noise -** Scharenborg et al. (2017) involves adding artificial Gaussian noise El Helou & Süsstrunk (2020) to the training data to enhance the ASR model's robustness and generalization. The objective of this method is to make the model more resistant to variations in input data, such as different accents or speaking styles, and more adept at handling real-world scenarios where the data may be noisy or contain errors. Consequently, the model can handle sound signals that deviate slightly from those on which it was trained Hu et al. (2018). To augment audio recordings, an array of the same size as the audio recording is created and populated with random samples from a uniform distribution between 0.001 and 0.03 Hz. Next, the amplitude (Hz) of the original audio recording is multiplied by the generated array, and this product is added to the original file's amplitude through matrix addition to obtain the augmented file.

- **Pitch Shift -** Scharenborg et al. (2017); Thai et al. (2019) involves altering the pitch of an entire audio recording uniformly by a certain number of semitones Gfeller et al. (2020), resulting in the synthesis of different-sounding voices. Many speech audio datasets contain a limited variety of speakers, where each audio recording constitutes a substantial portion of the dataset. For instance, the Spanish Common-Voice dataset comprises around 2,901 distinct speakers. Consequently, incorporating variance into the dataset can significantly enhance the model's ability to generalize and prevent over-fitting. To perform the pitch shift augmentation, a random number of semitones are selected from the range of $[-6, 6]$ for each audio recording.

The proposed methodology takes advantage of augmented data samples, employing the three mentioned-above data-augmentation techniques. By integrating these noisy samples into the training set, our approach enables the ASR model to become more robust and resilient to various types of acoustic variations and distortions commonly found in real-world speech. Consequently, this can result in enhanced performance and more precise speech transcriptions across diverse environments and conditions. Although the results in Table 8 exhibit minimal change, the derived model demonstrates increased robustness.

Multilingual training comparison - augmentation-based datasets and non-augmented datasets					
Ent.	Language Pair	Data Augmentation		Non Data Augmentation	
		CER [%]	WER [%]	CER [%]	WER [%]
1	Russian & Portuguese	**8.01**	**25.28**	8.21	26.22
2	Russian & Spanish	**8.79**	**27.88**	10.00	28.94
3	Spanish & Portuguese	**7.17**	**19.63**	7.47	19.91

Table 9. Results of models trained on data with 20% of augmented data added, compared to models without augmentation. This table presents the performance of the models in terms of CER and WER metrics, showcasing the impact of incorporating data augmentation techniques on model robustness and resilience across the various language pairs. The improvement appears to be minimal but in actuality in real-world use cases the model train with the added augmented data is more robust.

Turning our attention to Table 9, we can analyze the impact of data augmentation on our models. For the Russian and Portuguese language pair, in entry 1 the model that was trained with augmentation achieves a CER of 8% and a WER of 25%, compared to the model trained without augmentation, which yields a CER of 8% and a WER of 26%. Similarly, for the Russian and Spanish language pair, the model with augmentation achieves a CER of 8% and a WER of 27%, as opposed to a CER of 10% and a WER of 28% for the model without augmentation. Lastly, for the Spanish and Portuguese language pair, both models with and without augmentation achieve a CER of 7% and a WER of 19%. Although the improvements in the metrics appear minimal, the augmented models exhibit increased robustness in real-world use cases.

6. Conclusions and Future Work

This paper presented a novel approach for improving multilingual ASR systems through intelligent semantic dataset creation. Firstly, we overcome the need for a Language Identification model, simplifying the system and reducing computational costs. Secondly, we enabled the system's immunity to different languages, allowing for cross-lingual communication and knowledge sharing. Thirdly, we eliminated output errors caused by pronunciation differences across languages without the need for a Language Change Detection model. Fourthly, we demonstrated that there exists a sentimental-semantic relationship between phrases from different languages and the acoustics of their speech signals, which can be harnessed to improve ASR performance. Finally, we provided a pragmatic solution to the problem of low data availability, which enhances the robustness and reliability of ASR systems.

Our experimental results have shown that models trained on high-similarity samples consistently outperform those trained on low-similarity samples, highlighting the importance of high semantic similarity for accurate ASR performance. Through the elimination of LID, the proposed approach reduces computational costs and enables real-time text output, which is beneficial for on-device implementations and real-time applications.

In conclusion, our work has contributed significantly to the field of MLASR by addressing key challenges that hinder its development. We believe that the proposed approach has significant potential for future research and practical applications in multilingual speech recognition. Nevertheless, there is still room for future work, by extending our approach to additional languages, exploring the impact of varying degrees of semantic similarity, and investigating other pre-trained models and architectures. Additionally, we aim to assess the effectiveness of different data augmentation techniques in making ASR systems more robust to various acoustic variations and distortions.

7. Acknowledgments

This work was supported by the Ariel Cyber Innovation Center in conjunction with the Israel National Cyber directorate in the Prime Minister's Office, and Ariel Data Science and Artificial Intelligence Research Center.

References

Abate, S. T., Tachbelie, M. Y., & Schultz, T. (2021). End-to-end multilingual automatic speech recognition for less-resourced languages: the case of four ethiopian languages. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7013–7017). IEEE.

Ahmed, H., Traore, I., Mamun, M., & Saad, S. (2023). Text augmentation using a graph-based approach and clonal selection algorithm. *Machine Learning with Applications*, *11*, 100452.

Ahmed, S., Shumailov, I., Papernot, N., & Fawaz, K. (2022). Towards more robust keyword spotting for voice assistants. In *31st USENIX Security Symposium (USENIX Security 22)*.

Al Shamsi, H., Almutairi, A. G., Al Mashrafi, S., & Al Kalbani, T. (2020). Implications of language barriers for healthcare: a systematic review. *Oman medical journal*, *35*, e122.

Alsayadi, H., Abdelhamid, A., Hegazy, I., & Taha, Z. (2021). Data augmentation for arabic speech recognition based on end-to-end deep learning. *International Journal of Intelligent Computing and Information Sciences*, *21*, 50–64.

Anidjar, O. H., Barak, A., Ben-Moshe, B., Hagai, E., & Tuvyahu, S. (2023a). A stethoscope for drones: Transformers based methods for uavs acoustic anomaly detection. *IEEE Access*, .

Anidjar, O. H., Estève, Y., Hajaj, C., Dvir, A., & Lapidot, I. (2023b). Speech and multilingual natural language framework for speaker change detection and diarization. *Expert Systems with Applications*, *213*, 119238.

Anidjar, O. H., Hajaj, C., Dvir, A., & Gilad, I. (2020). A thousand words are worth more than one recording: Nlp based speaker change point detection. *arXiv preprint arXiv:2006.01206*, .

Anidjar, O. H., Lapidot, I., Hajaj, C., & Dvir, A. (2021a). A thousand words are worth more than one recording: Word-embedding based speaker change detection. In *Interspeech* (pp. 3121–3125).

Anidjar, O. H., Lapidot, I., Hajaj, C., Dvir, A., & Gilad, I. (2021b). Hybrid speech and text analysis methods for speaker change detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2324–2338.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, .

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020a). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 12449–12460). Curran Associates, Inc. volume 33. URL: `https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf`.

Berns, T., Vaessen, N., & van Leeuwen, D. A. (2023). Speaker and language change detection using wav2vec2 and whisper. *arXiv preprint arXiv:2302.09381*, .

Casola, S., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: An empirical comparison. *Machine Learning with Applications*, *9*, 100334.

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2022). Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, (pp. 1–42).

Choi, K., & Park, H.-M. (2022). Distilling a pretrained language model to a multilingual asr model. *Interspeech*, . doi:10.48550/arxiv.2206.12638.

Choutri, K., Lagha, M., Meshoul, S., Batouche, M., Kacel, Y., & Mebarkia, N. (2022). A multi-lingual speech recognition-based framework to human-drone interaction. *Electronics*, *11*, 1829.

Chowdhury, S. A., Hussein, A., Abdelali, A., Ali, A., Ali, A., Ali, A. H., Ali, A., & Ali, A. (2021). Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. *arXiv: Computation and Language*, . doi:10.21437/interspeech.2021-1809.

Chuang, S.-P., Liu, A. H., Liu, A. H., Sung, T.-W., Sung, T.-W., & yi Lee, H. (2020). Improving automatic speech recognition and speech translation via word embedding prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, . doi:10.1109/taslp.2020.3037543.

Clark, L., Cowan, B. R., Roper, A., Lindsay, S., & Sheers, O. (2020). Speech diversity and speech interfaces: Considering an inclusive future through stammering. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (pp. 1–3).

Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Processing*, *7*, 25–46.

Datta, A., Ramabhadran, B., Emond, J., Kannan, A., & Roark, B. (2020). Language-agnostic multilingual modeling. *arXiv: Audio and Speech Processing*, . doi:10.1109/icassp40776.2020.9053443.

Deléglise, P., Esteve, Y., Meignier, S., & Merlin, T. (2009). Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate? In *Tenth Annual Conference of the International Speech Communication Association*.

Dendrinos, B. (2006). Mediation in communication, language teaching and testing. *Journal of Applied Linguistics*, *22*, 9–35.

Deschamps-Berger, T., Lamel, L., & Devillers, L. (2022). Investigating transformer encoders and fusion strategies for speech emotion recognition in emergency call center conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction* (pp. 144–153).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Dominguez, A. R. (2023). Portfolio optimization based on neural networks sensitivities from assets dynamics respect common drivers. *Machine Learning with Applications*, *11*, 100447.

El Helou, M., & Süsstrunk, S. (2020). Blind universal bayesian image denoising with gaussian noise level learning. *IEEE Transactions on Image Processing*, *29*, 4885–4897.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020a). Language-agnostic bert sentence embedding. *Annual Meeting of the Association for Computational Linguistics*, . doi:10.18653/v1/2022.acl-long.62.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020b). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, .

Ford, E., Maneparambil, K., Kumar, A., Sant, G., & Neithalath, N. (2022). Transfer (machine) learning approaches coupled with target data augmentation to predict the mechanical properties of concrete. *Machine Learning with Applications*, *8*, 100271.

Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., & Velimirović, M. (2020). Spice: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1118–1128.

Habbat, N., Anoun, H., & Hassouni, L. (2021). A novel hybrid network for arabic sentiment analysis using fine-tuned arabert model. *International Journal on Electrical Engineering and Informatics*, *13*, 801–812.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, .

Higuchi, Y., Karube, K., Ogawa, T., & Kobayashi, T. (2022). Hierarchical conditional end-to-end asr with ctc and multi-granular subword units. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7797–7801). IEEE.

Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J., & Shinozaki, T. (2020). Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. *Babel*, *37*, 10k.

Hu, H., Tan, T., & Qian, Y. (2018). Generative adversarial networks based data augmentation for noise robust speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5044–5048). IEEE.

Javeed, A. (2023). A hybrid attention mechanism for multi-target entity relation extraction using graph neural networks. *Machine Learning with Applications*, *11*, 100444.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th international conference on advanced computing and communication systems (ICACCS)* (pp. 136–141). IEEE.

Juang, B.-H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Kawazoe, Y., Shibata, D., Shinohara, E., Aramaki, E., & Ohe, K. (2021). A clinical specific bert developed using a huge japanese clinical text corpus. *Plos one*, *16*, e0259763.

Khodadadi, A., Ghandiparsi, S., & Chuah, C.-N. (2022). A natural language processing and deep learning based model for automated vehicle diagnostics using free-text customer service reports. *Machine Learning with Applications*, *10*, 100424.

Kramsch, C. (2014). Language and culture. *AILA review*, *27*, 30–55.

Kumar, T., Mahrishi, M., & Meena, G. (2022). A comprehensive review of recent automatic speech summarization and keyword identification techniques. *Artificial Intelligence in Industrial Applications*, (pp. 111–126).

Kumar, Y., & Singh, N. (2019). A comprehensive view of automatic speech recognition system-a systematic literature review. In *2019 international conference on automation, computational and technology management (ICACTM)* (pp. 168–173). IEEE.

Li, B., Pang, R., Zhang, Y., Sainath, T. N., Strohman, T., Haghani, P., Zhu, Y., Farris, B., Gaur, N., & Prasad, M. (2022a). Massively multilingual asr: A lifelong learning solution. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, . doi:10.1109/icassp43922.2022.9746594.

Li, J. et al. (2022b). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, .

Luo, N., Jiang, D., Zhao, S., Gong, C., Zou, W., & Li, X. (2018). Towards end-to-end code-switching speech recognition. *arXiv preprint arXiv:1810.13091*, .

Malek, J., Jansky, J., Koldovsky, Z., Kounovsky, T., Cmejla, J., & Zdansky, J. (2022). Target speech extraction: Independent vector extraction guided by supervised speaker identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2295–2309.

Mustafa, M. B., Yusoof, M. A., Khalaf, H. K., Rahman Mahmoud Abushariah, A. A., Kiah, M. L. M., Ting, H. N., & Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, *12*, 9541.

Muthusamy, Y. K., Barnard, E., & Cole, R. A. (1994). Reviewing automatic language identification. *IEEE Signal Processing Magazine*, *11*, 33–41.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, .

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J. et al. (2021). Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, .

Reda, A., & Aoued, B. (2005). Artificial neural network & mel-frequency cepstrum coefficients-based speaker recognition. In *3rd International Conference: Science of Electronic, Technologies of Information and Telecommunication (SETIT 2005)* (pp. 27–31).

Ren, Z., Qian, K., Dong, F., Dai, Z., Nejdl, W., Yamamoto, Y., & Schuller, B. W. (2022). Deep attention-based neural networks for explainable heart sound classification. *Machine Learning with Applications*, *9*, 100322.

Richardson, B. H., Taylor, P. J., Snook, B., Conchie, S. M., & Bennell, C. (2014). Language style matching and police interrogation outcomes. *Law and human behavior*, *38*, 357.

Rodrawangpai, B., & Daungjaiboon, W. (2022). Improving text classification with transformers and layer normalization. *Machine Learning with Applications*, *10*, 100403.

Saeedi, J., & Giusti, A. (2023). Semi-supervised visual anomaly detection based on convolutional autoencoder and transfer learning. *Machine Learning with Applications*, *11*, 100451.

Sailor, H. B., T, K. P., Agrawal, V., Jain, A., & Pandey, A. (2021). Sri-b end-to-end system for multilingual and code-switching asr challenges for low resource indian languages. *Interspeech*, . doi:10.21437/interspeech.2021-1578.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., & Post, M. (2021). The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*, .

Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L., & Solorio, T. (2016). Multilingual code-switching identification via lstm recurrent neural networks. In *Proceedings of the second workshop on computational approaches to code switching* (pp. 50–59).

Scharenborg, O., Ciannella, F., Palaskar, S., Black, A., Metze, F., Ondel, L., & Hasegawa-Johnson, M. (2017). Building an asr system for a low-research language through the adaptation of a high-resource language asr system: preliminary results. In *Proc. Internat. Conference on Natural Language, Signal and Speech Processing (ICNLSSP)* (pp. 26–30).

Shahgir, H., Sayeed, K. S., & Zaman, T. A. (2022). Applying wav2vec2 for speech recognition on bengali common voices dataset. *arXiv preprint arXiv:2209.06581*, .

Shahnawazuddin, S., Adiga, N., Kathania, H. K., & Sai, B. T. (2020a). Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognition Letters*, *131*, 213–218.

Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., & Ahmad, W. (2020b). Voice conversion based data augmentation to improve children's speech recognition in limited data scenario. In *Interspeech* (pp. 4382–4386).

Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., Vieira, F., McNally, M., Charbonneau, T., Nollstadt, M. et al. (2019). Personalizing asr for dysarthric and accented speech with limited data. *arXiv preprint arXiv:1907.13511*, .

Singh, G., Sharma, S., Kumar, V., Kaur, M., Baz, M., & Masud, M. (2021). Spoken language identification using deep learning. *Computational Intelligence and Neuroscience*, *2021*.

Steinberg, E. M., Valenzuela-Araujo, D., Zickafoose, J. S., Kieffer, E., & DeCamp, L. R. (2016). The "battle" of managing language barriers in health care. *Clinical pediatrics*, *55*, 1318–1327.

Tachbelie, M. Y., Abate, S. T., & Schultz, T. (2022). Multilingual speech recognition for globalphone languages. *Speech Communication*, *140*, 71–86.

Temraz, M., & Keane, M. T. (2022). Solving the class imbalance problem using a counterfactual method for data augmentation. *Machine Learning with Applications*, *9*, 100375.

Thai, B., Jimerson, R., Arcoraci, D., Prud'hommeaux, E., & Ptucha, R. (2019). Synthetic data augmentation for improving low-resource asr. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)* (pp. 1–9). IEEE.

Thomas, S., Audhkhasi, K., & Kingsbury, B. (2020). Transliteration based data augmentation for training multilingual asr acoustic models in low resource settings. In *INTERSPEECH* (pp. 4736–4740).

Vanderreydt, G., REMY, F., & Demuynck, K. (2022a). Transfer learning from multi-lingual speech translation benefits low-resource speech recognition. *Interspeech*, . doi:10.21437/interspeech.2022-10744.

Vanderreydt, G., Remy, F., & Demuynck, K. (2022b). Transfer learning from multi-lingual speech translation benefits low-resource speech recognition. In *Interspeech2022* (pp. 3053–3057).

Wangaryattawanich, P., Chavali, L. S., Shah, K. B., Gogia, B., Valenzuela, R. F., DeMonte, F., Kumar, A. J., & Hayman, L. A. (2016). Contrast-enhanced reformatted mr images for preoperative assessment of the bridging veins of the skull base. *Radiographics*, *36*, 244–257.

Yadav, H., & Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2202.12576*, .

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing*, *4*, 31.

Zissman, M. A., & Berkling, K. M. (2001). Automatic language identification. *speech communication*, *35*, 115–124.