Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 00 (2023) 1–14

Signal Processing

Augmented Wav2Vec 2.0: ASR Improvement Using Data Augmentation for Under-Represented Languages

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

[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

With the growing popularity in Deep Learning and Transformers, acoustic models for learning robust representations such as the well-known Wav2Vec2, achieve high performance with respect to state-of-the-art approaches for applications in the Speech Recognition (SR) field. Automatic speech recognition (ASR), is a main SR-based application that can be solved more efficiently thanks to fine-tuning capabilities of Wav2Vec2, by extracting voice characteristics and learning them in the form of representative vectors. Despite the improvement and efficiency of Wav2Vec2 compared to acceptable methods, the strength and robustness of Wav2Vec2, including the use of hundreds of times less of labeled data - for under-represented languages, the performance of Wav2Vec2 drops significantly. This paper presents an end-to-end framework that includes a full model training process, based on data augmentation technologies that help significantly improve an ASR system, that was fine-tuned on Wav2Vec2. In order to prove the effectiveness of our framework, we present a fine-grained experimental evaluation set that supports our hypothesis, by using three datasets from the well known Mozilla's Common-Voice datasets, in the following under-represented languages: Arabic, Russian and Portuguese. Moreover, in worldwide-spoken languages like Arabic, the abundance of accents and forms of pronunciation of words significantly complicates the success of the basic Wav2Vec2 in ASR, and our framework shows robustness to different diacritics. Finally, our approach yields an average relative improvement of 33.9% in Word Error Rate (WER), and 53.2% of relative improvement in Character Error Rate (CER), compared to the Wav2Vec2 baseline model.

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

1. Introduction

Automatic Speech Recognition (ASR) Avci & Akpolat (2006); Zoughi et al. (2020); Li et al. (2022) is a technique that processes human speech into readable text. ASR systems are also known as Speech-to-Text (S2T) or transcription systems Ronao & Cho (2016). Among nowadays applications, one can find virtual assistants such as Apple's Siri McCrocklin et al. (2022) or Amazon's Alexa Bräuer & Mazarakis (2022), which are utterly based on ASR systems. The field of speech-recognition in general, and particularly ASR, has grown exponentially over the last two decades, as

[*]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), revi85@gmail.com (Revital Marbel), Najeeb040@gmail.com (Najeeb Abdalla), nerya0001@gmail.com (Nerya Bigon), benmyara13@gmail.com (Benjamin Myara), roiyo@ariel.ac.il (Roi Yozevitch)

ASR systems became more and more popular in industries such as call centers Ha et al. (2020), education McCrocklin et al. (2022); Bräuer & Mazarakis (2022), finance Bandi & Kothari (2022), healthcare Sezgin & D'Arcy (2022), etc.

Given the popularity and demand of this technology, this paper proposes an ASR framework for under-represented languages, i.e., languages that have limited resources and data available for the development and implementation of accurate ASR systems. Such framework needs to be robust to a variety of dialects Dorn (2019), languages, and real-world sound quality - mainly noisy environments Kinoshita et al. (2020).

Recent advancements in the field of speech recognition have led to significant improvements in the accuracy of spoken language transcription Baevski et al. (2020a). However, some challenges are still relevant, that are making tough life to the technology; one of these challenges is background noise that greatly disturbs the system and can rarely be avoided. The noise of a helicopter landing, or a passing car are good examples. Another challenge, which is even worse, is different accents within the same language. For instance, in the Arabic language, the accent varies from one speaker into another, depending on the origin-country of that speaker.

Recently, it has been shown that pre-trained models, followed by fine-tuning on very little labeled speech data achieve competitive results with respect to state-of-the-art ASR systems Baevski et al. (2020a). For instance, using as little as merely few hours of labeled data, the Wav2Vec2 model Baevski et al. (2020a) yields a Word Error Rate (WER) Deléglise et al. (2009) of less than 5% on the clean test set of LibriSpeech dataset Garnerin et al. (2021). Despite of the fact that the Wav2Vec2 model has been proven to be effective for ASR systems, even whenever only a diminished amount of data is available - it might under-perform in case of under-represented languages, and the Arabic language is an excellent doctrine for that.

1.1. Arabic Dialects

Arabic, is a Semitic Huehnergard & Pat-El (2019) language spoken by more than 300 million people across the Middle-East, North-Africa and Asia. Arabic dialects are variations of the Arabic language that are spoken by different groups of people in different regions. While there is a standardized form of Modern Standard Arabic (MSA) Qwaider et al. (2019) that is used for formal written communication, spoken Arabic can vary greatly from one region to another Alhelbawy et al. (2020). These regional variations are known as dialects, and each dialect has its own distinct features and characteristics. These dialect-variations are that notable, so that some people consider them as different languages. These problem is not restricted to the variability of the Arabic language and can also be found in European dialects as well Khosravani et al. (2021); Alsayadi et al. (2022). Specifically, the main differences between Arabic dialects include pronunciation and vocabulary;

The pronunciation Caballero-Morales & Trujillo-Romero (2014) of certain consonants and vowels can vary greatly from one dialect to another. For instance, the *'d'* sound in Standard Arabic is pronounced as a *'z'* in some dialects, and the *'q'* sound is pronounced as a *'g'* in others. These differences can make it difficult for speakers of one dialect to understand speakers of another.

As for the vocabulary, while respected portion of words is shared across all dialects, many of them are specific to a particular one. In fact, the Arabic vocabulary difference can make it challenging for speakers of different dialects to communicate with each other. The differences in vocabulary between Arabic dialects can affect the performance of any S2T system, since it may not recognize words that are specific to a particular dialect. For instance, an ASR system that is trained on data from a dialect that uses the Arabic word "*Automobil*" for "car", may not accurately transcribe the word "*sayyara*" when it is spoken by a speaker of another dialect. Thus, robustness in S2T systems can only be achieved by taking the different dialects into account.

This challenges, i.e., the pronunciation and rich vocabulary of under-represented languages as the Arabic one, a simple training of a state-of-the-art ASR model such as the Wav2Vec2 is not sufficient; to improve the performance of an ASR system for Arabic dialects, it is important to train it on a diverse range of dialects and vocabulary. This can be done through the use of large, diverse datasets of transcribed speech, as well as advanced machine learning algorithms that can handle a wide range of dialects and vocabulary. By doing so, the ASR system can learn to recognize a variety of words and phrases, and can accurately transcribe speech from different Arabic dialects. In order to compute the ASR performance of our framework, additionally to the common WER Deléglise et al. (2009) metric, we compute the Character-Error-Rate (CER) Hou et al. (2020); Kumar et al. (2022), which is another common metric for computing the performance of an ASR systems, and is similar to WER, except for that the error computation is at the character level.

1.2. Our contribution

This paper offers a robust approach for fine-tuning the Wav2Vec2 speech recognition model using a novel data augmentation method Shahnawazuddin et al. (2020); Thomas et al. (2020). The algorithm was trained and tested on three different languages - Arabic, Portuguese and Russian. Those languages were not chosen randomly - all of them are under-represented languages (languages with low data availability). Thus, The fine-tuning process was done using limited audio data (Arabic and Portuguese - 17 hours, Russian - 30 hours). The model achieved an average accuracy improvement of 34% in the word level (WER) and an average accuracy improvement of more than 50% in the character level, i.e. CER. The results also show immunity to different languages with different grammatical rules, syntax etc. Furthermore, the algorithm's accuracy can be further improved once more labeled data is available. Overall, the proposed work makes a contribution to the field of speech processing for under-represented languages by providing a practical solution to the problem of low data availability and showing the effectiveness of the proposed approach on multiple downstream tasks, paving the way for more robust and reliable S2T transcription.

1.3. Paper Structure

The remainder of this paper is structured as follows: Section 2 surveys related work mainly regarding common ASR approaches in general, and the Wav2Vec2 architecture in particular, as well as self-supervision; Section 3 discusses dataset used in this paper and the pre-processing procedure; Section 4 presents the the different data-augmentation methods that were utilized in this work, and all of their possible combinations, as well as their influence over the ASR metrics of our framework; Section 5 presents the approach employed in this paper as part of the Wav2Vec2 architecture exploitation; Section 6 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 7 concludes the paper and offer some directions for future work. 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
CTC	Connectionist Temporal Classification
DER	Diarization Error Rate
S2T	Speech-2-Text
SCD	Speaker Change Detection
SD	Speaker Diarization
SR	Speech Recognition
SSL	Self Supervised Learning
WER	Word Error Rate

Table 1. List of Abbreviations.

2. Related Work

Representation-Learning Bengio et al. (2013), is a set of techniques for vector representation of data it is used for tasks like classification and clustering O'Shea et al. (2016). Thus, representation-learning can replace manual feature extraction and feature engineering. Transformers are a type of representation-learning that uses Self Supervised Learning (SSL) Nguyen et al. (2021); Jaiswal et al. (2020); Shurrab & Duwairi (2022) and self-attentionZhai et al. (2019) to learn the best representation of raw data for a given task. SSL is a machine learning technique that is used to train a model with unlabeled data, usually before re-training it again later with labeled data, a process usually referred to as fine-tuning. The idea of SSL has arisen as a machine learning framework to solve the problem of labeled data scarcity Babbar & Schölkopf (2019). It enables learning general data representations from unlabeled examples in a supervised learning task and fine-tunes the model on labeled data by adding a predictor to the model that takes in the representations learned by SSL.

The Wav2Vec2 Baevski et al. (2020a,b) is a Transformer-based Lin et al. (2022) model, mainly used in the context of speech recognition, and based on SSL. It uses raw audio waveform as input, and generates a vector-based language representation Sasajima et al. (1996), which uses a vectorial representation. Nowadays, the Wav2Vec2 is considered as the state-of-the-art transformer for its high-accuracy speech transcription. It is also able to handle a wide range of languages, accents, and speech styles.

Previous work on Wav2Vec2 focused on improving the performance of the model using various techniques, such as fine-tuning the model, adjusting the model architecture or hyper-parameters, and incorporating additional training objectives. One example is the Wav2Vec2-xlsr-53 Deschamps-Berger et al. (2022) model, which was trained on 53 languages and achieved state-of-the-art performance on a range of speech recognition tasks Shahgir et al. (2022). Another work Farias et al. (2022), has also explored the use of Wav2Vec2 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. Overall, the Wav2Vec2 model has been widely adopted in the speech recognition community and has demonstrated its effectiveness in a variety of applications, mainly due to the benefits of SSL.

Commonly, training and fine-tuning a speech recognition model with a limited amount of training data Thomas et al. (2020), does not yield a robust low-error-rate model. ASR models usually require large-scale and diverse datasets in order to learn efficiently. This robustness gap leads the main purpose of this paper, which is to cope with ASR systems that are based on low-resource and under-represented languages Shor et al. (2019). Additionally, training on a limited amount of data can also lead to over-fitting, where the model performs well on the training data but poorly on new, unseen data. One method suggested by Alsayadi et al. (2021) to overcome this limitation is the data augmentation method Shahnawazuddin et al. (2020).

Another approach is utilizing clustering methods of unlabeled data Bakheet (2021); Hsu et al. (2021) (which is easier to find). The problem in such methods is that low error-rate of the ASR are no guaranteed, and there is a question regarding the reliability of ASR systems whenever their WER Deléglise et al. (2009) is quite high;

Voice models are incredibly fragile because the variance between them is high, and the data is very noisy. ASR models are particularly sensitive because, in addition to phonemes recognition in the voice segments, there is another layer of language processing. Hence, many ASR models have not yet reached a low error performance, and this translates into a high WER. However, high WER does not always leads to a crisis; one example of a high WER result was presented in Anidjar et al. (2023), that have designed an end-to-end framework for the Speaker Change Detection (SCD) Anidjar et al. (2020); Meng et al. (2017); Hrúz & Zajíc (2017) and Speaker Diarization (SD) Lin et al. (2019); Shum et al. (2013); Silnova et al. (2020) challenges, which aims to answer the question *'who spoke when?'* in a given audio-recording (SD), according to the speaker-turns in it (SCD). The dataset on which their approach was suggested is ASR-based, i.e., contains WER and CER errors; it consists of three monolingual datasets in three different languages - English, French, and Hebrew. One main observation in Anidjar et al. (2023), is that the SCD and SD problems have been solved quite successfully, with 97.66% of F1-Score for the SCD, and 10.28 of Diarization Error Rate (DER) Deléglise et al. (2009), despite of the fact that the English ASR engine had a WER of 40.3%.

In another work, Këpuska & Bohouta (2017) have designed a tool that is used to test and conduct a comparison of several commercial ASR systems, such as the well-known Microsoft Speech API Këpuska & Bohouta (2017), or Google Speech API Anggraini et al. (2018), with an open-source ASR systems such as the Sphinx-4 Walker et al. (2004); the well-known Sphinx-4 Walker et al. (2004) has achieved 37% of WER Li et al. (2020); Nakatani (2019). It has been shown that despite such a relatively high WER value, the Sphinx-4 is still competitive for speech-recognition tasks Walker et al. (2004); Hafeez et al. (2014), when compared with low-WER ASR systems such as the Microsoft Speech API Këpuska & Bohouta (2017), or Google Speech API Anggraini et al. (2018) that achieved 18% and 9% of WER, correspondingly.

Radford et al. (2022) have studied the capabilities of ASR systems that are trained mainly in order to predict large amounts of transcriptions of audio-recordings on the internet. When their model is scaled to 680,000 hours of multilingual and multitask supervision, it achieves 9.9% of WER in English dataset and 29.2% WER in a multilingual dataset, based on a weak-supervision Kuang et al. (2022) on the Wav2Vec2 Baevski et al. (2020a,b) architecture.

Tran & Soleymani (2022) have presented a speech-representation anonymization framework, via selective noise perturbation whenever privacy and security are main concerns whenever audio-recordings are being processed in cloud services, in order to perform ASR, or Speech Emotion Recognition (SER) Guo et al. (2022). Even whenever the WER of the ASR systems is 39.6%, the framework suggested in Tran & Soleymani (2022) is capable of recognizing emotions in audio-recordings.

2.1. Whisper

Recently, Open.AI organization has launched a general-purpose speech recognition model - Whisper Radford et al. (2022). It is trained on a large data-set of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification Almeida et al. (2014). According to their GitHub page, in Arabic, the model reached a WER Sharma & Jayagopi (2021) of 16% on the Fleurs dataset (Conneau et al. (2022)). A major difference between Whisper and Wav2Vec2 is the model size. According to the original Open.AI paper Radford et al. (2022), Whisper is a much larger model than Wav2Vec2, with over 1.5 billion parameters in contrast to Wav2Vec2 that has ≈ 3 Million parameters (which is less than 1% of Whisper's size). However, three important points must be taken into account when comparing Whisper and Wav2Vec2:

(1). The Whisper WER is calculated using a custom-built text normalizer on the transcribed text *before* calculating the WER (as mentioned in the original Open.AI paper). The text normalizer is designed to improve the accuracy of the WER calculation by correcting common mistakes and normalizing the transcribed text to match the reference text.

(2). While this approach of working only with fillet data produce good results, it does not reflect real-world scenarios; in datasets of under-represented languages (such as Arabic), the training data itself is defected. For example, during the training process, one might come across many empty audio-recordings (that only contain silent), yet with a full sentence as a label.

(3). The most important difference, is that Whisper is a closed system and cannot be fine-tuned. It is what it is. Conversely, Wav2Vec2 can be further improved once more data is available.

3. Datasets

The Mozilla Common-Voice dataset Ardila et al. (2019); Berkson et al. (2019); Chachadi & Nirmala (2022) is a free collection of recorded speech data. It is publicly available, and fosters innovation and an acceptable as an enabler of commercial competition in machine-learning competitions that are based on speech technology. Moreover, Common-Voice is a multi-language dataset, which is considered as on of the largest publicly available voice datasets of its kind.

Common-Voice is used both to train and evaluate speech recognition algorithms. The data-set contains recordings of over 400,000 people speaking in multiple languages. It includes various voices, including different ages, genders, accents, and speaking styles. It also includes a diverse range of transcriptions, including informal conversations, news articles, and public service announcements. It is important to know that not all of the languages are equally represented in the dataset. For instance, the English language has over 2,000 hours of audio data, whereas the Arabic language has only 89 hours of audio data. As this work is devoted to ASR systems that are based on low-resource languages, the ones that are tackled in this work are the following 3 underrepresented languages - Arabic, Russian and Portuguese;

- **Arabic.** The Arabic version of the Common-Voice dataset consists of ≈ 89 hours of community-validated audio, 147 hours of audio, and 1,309 unique voices in mp3 format. It is important to note that **the train split only contains 17 hours of labeled data.**

- **Portuguese.** The Portuguese part of the Common-Voice dataset includes 126 hours of community-validated audio, 151 hours of audio and 2621 unique voices in mp3 format. **The train split only contains 17 hours of labeled data.**

- **Russian.** The Russian part of the Common-Voice dataset includes 180 hours of community-validated audio, 215 hours of audio, and 2731 unique voices and is in mp3 format. **The train split only contains 30 hours of labeled data.**

Unfortunately, the Common-Voice dataset contains more than a few samples with an empty audio-recording (only contains silent parts) that still has a label, or that the label does not match the audio - which will result in miss-classifications, and affect the WER and CER.

3.1. Data Pre-Processing

To prepare the data for training, the Wav2Vec2 framework imposes the following requirements on the audio data:

- The sample rate must be changed from 44kHz to 16kHz since that is the sample rate Wav2Vec2 can work with.
- Special characters were removed.
- Punctuation marks were removed.

4. Augmentation Methods

Data augmentation Shahnawazuddin et al. (2020) is a common technique that is used to artificially increase the size of a dataset by creating modified versions of existing data. In the realm of audio, this can be done by applying transformations such as pitch shifting, time stretching, or adding noise to the audio clip. Data augmentation is often used in machine learning to improve the performance and generalization of models by providing them with additional training examples. The new augmented data is viable since the augmentations do not interfere with the transcription of the original audio-recording Ragni et al. (2014). The new data can be used to further train the model.

Focusing on a minimal number of high-impacting augmentations is vital to successfully improving the model. On one hand, insufficient augmented data can cause over-fitting due to the insufficient variety of the data. Conversely, augmenting the data can lead to under-fitting since critical data can be lost in the process, making the words unrecognizable. Therefore, augmenting the data by increasing the variance and robustness of the ASR-system, is crucial to make any significant improvement. In this work, three augmentation methods are considered:

- Band-Stop Thai et al. (2019) (Section 4.1).
- Gaussian-Noise Scharenborg et al. (2017) (Section 4.2).
- Pitch-Shift Scharenborg et al. (2017); Thai et al. (2019) (Section 4.3).

Finally, we discuss in Section 4.4 about the combination of augmentation methods with respect to model construction.

4.1. Band Stop

A **Band-Stop** Roonizi & Jutten (2021) augmentation is a type of data augmentation that involves removing a specific frequency range from an audio signal. This is typically done by applying a band-stop filter to the audio, which attenuates frequencies within a certain range and allows frequencies outside that range to pass through. The frequencies that humans can hear are between $0 - 4000hz$. Thus, we used these frequencies as a minimum and maximum threshold. Secondly, we set the min/max cutoff range represented by the bandwidth fraction (the absolute bandwidth divided by the center frequency represented between $0 - 200\%$), which indicates the relative portion of the frequency spectrum to cut-off. Lastly, we set the steepness of the cutoff presented in dB. The objective of utilizing this augmentation technique is to artificially simulate various (Arabic) accents.

4.2. Gaussian-Noise

Artificially adding **Gaussian-Noise** augmentation El Helou & Süsstrunk (2020) to the training data can improve the robustness and generalization of an ASR model. The goal of this method is to make the model more resistant to variations in the input data, such as different accents or speaking styles, and more capable of handling real-world scenarios where the data may be noisy or contain errors. As a result, the model can handle sound signals that are slightly different from the ones it was trained on Hu et al. (2018). In order to augment the audio-recording, we created an array that is the same shape as the audio-recording and populate it with random samples from a uniform distribution over 0.001, 0.03 Hz). Next, the amplitude (Hz) of the original audio-recording is multiplied by the generated array. This multiplication-result is being done by a matrix addition to the amplitude of the original file and get the augmented file.

4.3. Pitch Shift

Pitch Shift augmentation Gfeller et al. (2020) is achieved by shifting the tempo of the entire audio-recording uniformly by a certain semitone; by doing so, one can synthesize different sounding voices. Many speech audio data sets are composed of a small variety of speakers, where each audio-recording has a large portion. The Arabic Common-Voice dataset has roughly 15 different speakers per hour. Thus, adding variance to the data set can lead to a significant improvement in generalizing the model in order to prevent over-fitting. For each audio-recording, the semitones are chosen randomly for which to shift between the range $[-6, 6]$ semitones.

4.4. Augmentations Combination

These three augmentations are inherently different from one to another. Pitch-Shift uniformly changes the semitone of the entire audio-recording by a fixed amount, Band-Stop removes a certain range of frequencies from the audio-recording and Gaussian-Noise changes the amplitude of the entire audio-recording by a small amount in order to create the effect of white noise. Each of these augmentations plays a crucial role in improving the model's accuracy.

5. Framework

5.1. Model Architecture

This paper discusses the effectiveness of fine-tuning the Wav2Vec2-xlsr-53 model Deschamps-Berger et al. (2022) on augmented data. The Wav2Vec2-xlsr-53 model, pre-trained on 53 languages by the team at Facebook AI Research in September 2020, is a state-of-the-art approach for converting raw audio wave-forms into high-quality text representations. It is based on the concept of SSL, in which the model learns to predict missing segments of the input waveform.

The main architecture of the model has three main building blocks.

- **Pre-Processing.** The model receives raw audio matrix as input and outputs latent speech representations for each time-step among T time-steps.
- **Speech-Encoding.** the speech representations are fed into a Transformer that creates T representations, extracting information from the sequence.
- **CTC-Clustering.** the feature encoder output is discretized in order to represent the targets (outputs) as a self-supervised-based objective function.

The Wav2Vec2 is composed of a multi-layer based convolutional feature encoder. The feature encoder contains a temporal convolution followed by a normalization layer and a GELU Hendrycks & Gimpel (2016) activation function. The encoder's total stride computes the amount of the T time steps, which serves as the Transformer's input. Then, The Transformer produces contextualized speech representations. The feature encoder output is fed into a context network that follows the Transformer architecture as in Devlin et al. (2018). Finally, the Wav2Vec2 is trained using the CTC Higuchi et al. (2022) loss since the S2T problem is a sequence alignment problem. The main change is that instead of fixed positional embeddings Devlin et al. (2018), which encode absolute positional information, the Wav2Vec2 exploits a convolutional layer that behaves as if it was a relative positional embedding.

CTC loss-function. Particularly, in ASR systems, the alignment is difficult - that is, the alignment of each character to its adequate location in an audio-recording. For this purpose, the CTC loss computes the loss between a continuous and unsegmented acoustic time-series signal data sample and a target sequence-based label that is represented by characters. This computation is done by summing over the probability distribution of possible alignments between the speech signal and the textual sequence label by producing a loss value that is differentiable, with respect to each input node. The alignment between the speech signal and the textual sequence label is assumed to be "many-to-one," posing a limitation on the length of the textual sequence label, such that it is required to be as same as the input length. By using the CTC loss, we hypothesize that fine-tuning the model on augmented data would improve its performance compared to fine-tuning the same model on clean data.

The model architecture and flow are presented in Figure 1.

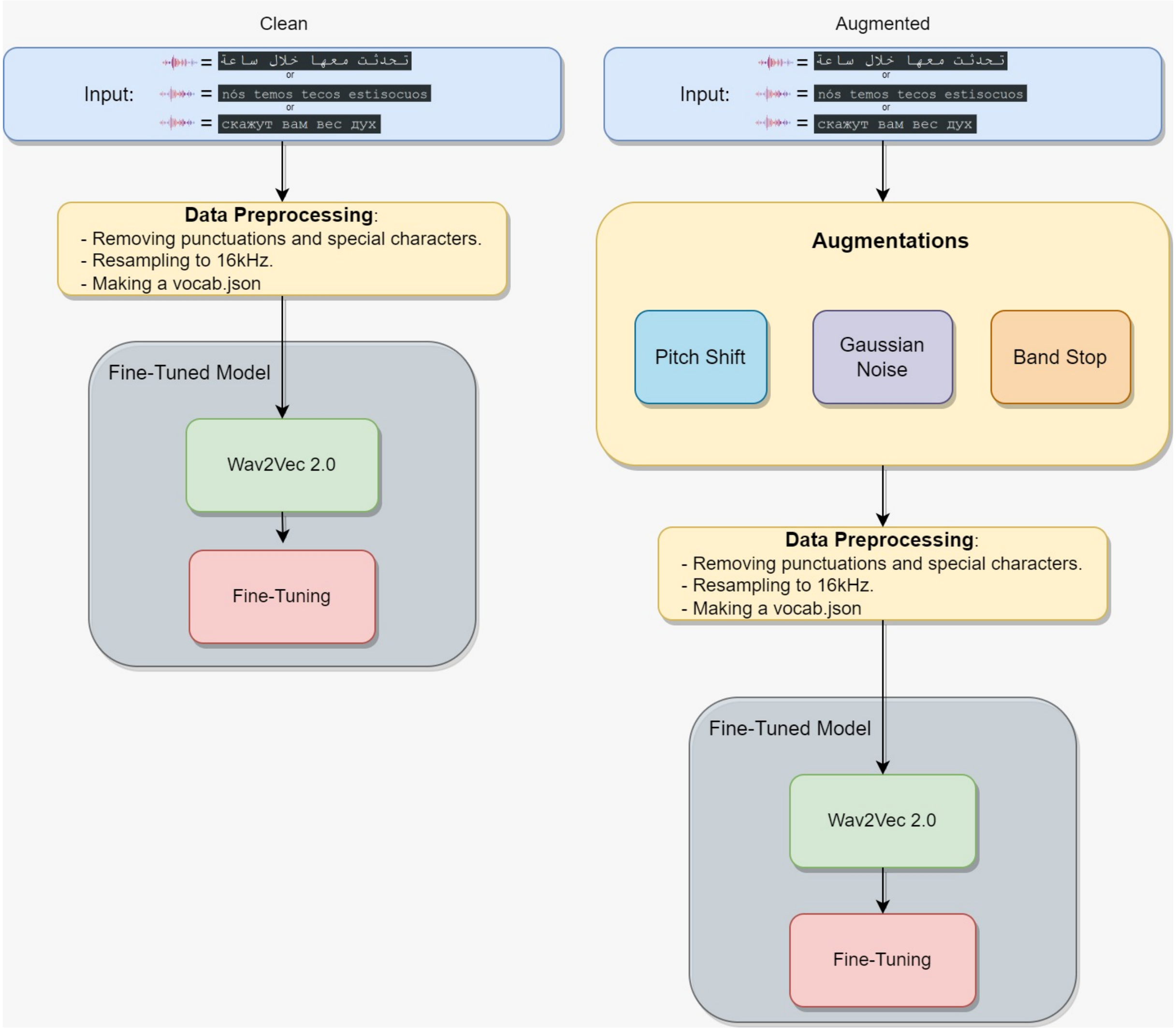

Figure 1. The framework presented in this paper.

We evaluated the performance of the fine-tuned models using the most common metrics in ASR systems which are WER and CER (Section 5.2), including their ability to accurately transcribe speech, capture the and content of the input audio, and generalize to new, unseen data. The results showed that fine-tuning the Wav2Vec2-xlsr-53 model on augmented data improved its performance compared to fine-tuning the same model on clean data.

5.2. ASR Precision Metrics

WER and CER are both most-common metrics of ASR systems performance. They are used to evaluate the accuracy of the system's transcription by comparing the system's output to a reference transcription of the same input. The main difference between WER and CER is the unit of measurement that is used. WER is based on the number of incorrect *words* in the system's transcription, while CER is based on the number of incorrect *characters*. The CER formula is given by Eq.(1):

$$CER = \frac{(I + S + D)}{N} \times 100 \tag{1}$$

where:

- I is the number of insertions (characters that are in the system's transcription but not in the reference transcription).
- S is the number of substitutions (characters in the system's transcription that are different from the corresponding characters in the reference transcription).
- D is the number of deletions (characters that are in the reference transcription but not in the system's transcription).
- N is the total number of characters in the reference transcription.

As for WER, its formula is very similar to the CER, and is given by Eq.(2):

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

where:

- I is the number of incorrect words in the system's output.
- S is the number of words that were correctly recognized but were out of order in the system's output.
- D is the number of words that were deleted from the reference transcription.
- N is the total number of words in the reference transcription.

5.3. *Arabic Diacritics*

Arabic diacritics are 11 symbols that are added to letters in the Arabic alphabet, to indicate vowel sounds and other phonetic features in specific words. For example, consider the following two identical sentences as presented in Figure 2. These diacritics heavily affect the reported WER - even a single error in one of the diacritics would fail the entire word. Thus, the diacritics-based WER computation does not capture the real model's performance.

وَرَوَى حُمَيْدُ عَنْ أَنَاسٍ أَنَ النَبِيَ صَلَىْ اللَه عَلَيْهِ وَسَلَمَ قَالَ

وروى حميد عن أناس أن النبي صلى الله عليه وسلم قال

Figure 2. An example of two identical sentences that are using different diacritics.

5.3.1. *WER vs CER*

In general, WER is more commonly used than CER because it is easier to understand and interpret. However, CER can be a useful metric in certain situations, such as when the transcription includes proper nouns or other words that are difficult to spell or when the transcription includes a large number of homophones (words that sound the same but are spelled differently). Moreover, if the system systematically fails to recognize a space between two words, CER and WER metrics will produce marginally different results. For example, consider the following two Portuguese sentences:

Reference : **é necessário fornecer quando formulado uma avaliação**

Prediction : **e necessário ponecer quando forme lado u mavalação**

The WER between the reference and prediction sentences is 85.7%, while the CER between them is only 17.3%

6. Experimental Evaluation and Results

This section presents the WER and CER results for the proposed model across the 3 tested languages. Across all languages, a considerable improvement was achieved. In all of the experiments, the following training parameters were used: 500 warm-up steps, $3e - 4$ learning rate, batch size of 16, evaluated every 100 steps.

6.1. Arabic

Throughout the experiment, the Wav2Vec2-xlsr-53 model was fine-tuned on a combination of clean and augmented data. As explained in section 4, the chosen augmentations were pitch shift, Gaussian noise, and band stop filter. To test our hypothesis that data-augmentation is of great utility for ASR systems, we first trained the Wav2Vec2-xlsr-53 model on 17 hours of clean audio data (Common-Voice 11.0) for the Arabic language only, since it is the most complex and dialect-rich. In order to determine what is the best augmentation-based combination (Section 4.4), the model was fine-tuned (train-split only) in the following manners:

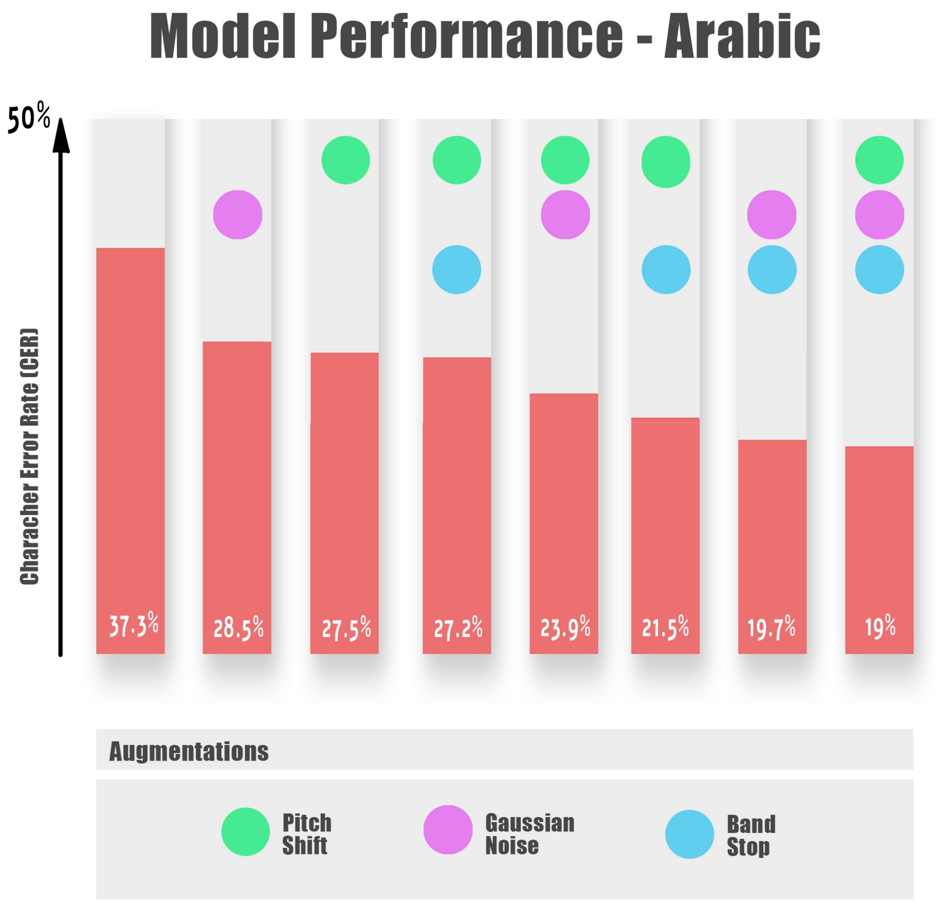

Figure 3. CER bar-chart for different mix of augmentation methods. One can see clearly that the more augmentations being used, the better the result. As can be seen, the best combination is the usage of the three augmentations: Pitch-Shift, Gaussian-Noise, and Band-Stop, which achieves 19% CER for a variety of diacritics in Arabic, which triples the size of the letters vocabulary, up to 80 letters.

(1). 100% clean data + 20% augmented data, so that a model was trained with 20% more augmented data, three times for the three different augmentations (Sections 4.1, 4.2 and 4.3). For each model, the aggregation of 20% augmented data was chosen randomly from the train split. Eventually, this step trains and evaluates three different models.

(2). 100% clean data + 20% augmented data from every pair of augmentations. That is, (i) one model with aggregation of 20% band-stop and 20% Gaussian-noise augmentations; (ii) one model with aggregation of 20% band-stop and 20% pitch-shift augmentations; and (iii) one model with aggregation of 20% Gaussian-noise and 20% pitch-shift augmentations. For each model, the aggregation of 40% augmented data (20% of one possible augmentation method and additional 20% of a second one) was chosen randomly from the train split. Eventually, this step trains and evaluates three different models.

(3). 100% clean data + 20% from each augmentation. That is, an aggregation of 20% band-stop augmentation, 20% Gaussian-noise augmentation, and 20% pitch-shift augmentation. For this model, the aggregation of 60% augmented data (20% of each augmentation) was chosen randomly from the train split. Eventually, this step trains and evaluates only model.

The CER bar-chart of all the 7 models is illustrated in Figure 3. As can be seen clearly from Figure 3, using augmented data significantly improve the model's performance, with the best result (CER=19.0%) achieved with all three augmentations. The more augmentations one utilize, the better the model. The CER of the non-augmented data is 37.5%. In addition, one can note that a single augmentation is inferior to a combination of (any) two augmentations and using all three is superior to all others.

6.2. Russian and Portuguese

Based on the results for the Arabic language, two more languages - Russian and Portuguese, were tested. While Arabic is a Semitic language, Portuguese and Russian are Latin and Slavic respectively. Thus, testing on these languages will prove model immunity to a specific language. Moreover, these language have a relatively small training data as presented in Section 3. The Portuguese training data consists of ≈ 17 hours of audio-recordings and the Russian training data consisted of ≈ 30 hours of audio-recordings (only train-splits from Common-Voice dataset). The WER and CER results are summarized in Tables 2 and 3:

Word Error Rate (WER)			
Language	Clean data [%]	Augmented [%]	Improvement [%]
Arabic	46.5	**27.6**	40.65
Russian	54.6	**35.8**	34.43
Portuguese	43.3	**31.8**	26.56

Table 2. Results table of WER for all languages.

Table 2 depicts the final WER for all three tested languages with and without the augmentation proposed in Section 4. While the average WER for the clean non augmented data is relatively high - ≈ 48%, the average WER for the augmented data is ≈ 32%. That is, the mean average accuracy improvement is ≈ 34%.

In addition, Table 3 depicts the CER results for the three languages:

Character Error Rate (CER)			
Language	Clean data [%]	Augmented [%]	Improvement [%]
Arabic	22.3	**9.0**	59.64
Russian	22.3	**10.2**	54.26
Portuguese	21.2	**11.5**	45.75

Table 3. Results table of CER for all languages.

Table 3 depicts the final CER for all three tested languages with and without the three augmentations proposed in section 4. The average CER for the clean, non-augmented data is ≈ 22%, while the average CER for the augmented data is ≈ 10%. The average accuracy improvement is ≈ 53%. The reason why a CER metric produces much better results was already explained in Section 5.3.1.

7. Conclusions and Future Work

The goal of this paper was to improve an ASR model for under-represented languages with low data availability. We have demonstrated the effectiveness of fine-tuning the Wav2Vec2-xlsr-53 model on a fusion of augmented data for improving its performance in transcribing speech, and capturing the meaning of the input audio. Some other speech-recognition tasks that could benefit from fine-tuning of the Wav2Vec2-xlsr-53 model on augmented data can include

Sentiment analysis, Speaker Identification, and Language Identification. A promising direction for future work may be investigating the use of different types of augmentations, such as adding background noise or altering the speed or pitch of the speech. This could help us understand the effects of different augmentations on the performance of the Wav2Vec2-xlsr-53 model. Another direction is studying the impact of different amounts of augmented data on the performance of the Wav2Vec2-xlsr-53 model. This could help in determine the optimal amount of augmented data needed to improve the model's performance. Least but not last, is the scenario on which the each dataset is multilingual; that is, a situation in which two (or more) languages are spoken in one single audio-recording, as can be seen in courts, inquiries, etc.

8. 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

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)*.

Alhelbawy, A., Lattimer, M., Kruschwitz, U., Fox, C., & Poesio, M. (2020). An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, *153*, 113365.

Almeida, S. G. M., Guimarães, F. G., & Ramírez, J. A. (2014). Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors. *Expert Systems with Applications*, *41*, 7259–7271.

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.

Alsayadi, H. A., Al-Hagree, S., Alqasemi, F. A., & Abdelhamid, A. A. (2022). Dialectal arabic speech recognition using cnn-lstm based on end-to-end deep learning. In *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)* (pp. 1–8). IEEE.

Anggraini, N., Kurniawan, A., Wardhani, L. K., & Hakiem, N. (2018). Speech recognition application for the speech impaired using the android-based google cloud speech api. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *16*, 2733–2739.

Anidjar, O. H., Estève, Y., Hajaj, C., Dvir, A., & Lapidot, I. (2023). 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*, .

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*, .

Avci, E., & Akpolat, Z. H. (2006). Speech recognition using a wavelet packet adaptive network based fuzzy inference system. *Expert Systems with Applications*, *31*, 495–503.

Babbar, R., & Schölkopf, B. (2019). Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, *108*, 1329–1351.

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`.

Bakheet, M. (2021). Improving speech recognition for arabic language using low amounts of labeled data.

Bandi, S., & Kothari, A. (2022). Artificial intelligence: An asset for the financial sector. *Impact of Artificial Intelligence on Organizational Transformation*, (pp. 259–287).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*, 1798–1828.

Berkson, K., Lotven, S., Thang, P. H., Thawngza, T., Sung, Z., Wamsley, J. C., Tyers, F., Van Bik, K., Kübler, S., Williamson, D. et al. (2019). Building a common voice corpus for laiholh (hakha chin). In *Proceedings of the Workshop on Computational Methods for Endangered Languages*. volume 2.

Bräuer, P., & Mazarakis, A. (2022). How to design audio-gamification for language learning with amazon alexa?—a long-term field experiment. *International Journal of Human–Computer Interaction*, (pp. 1–18).

Caballero-Morales, S.-O., & Trujillo-Romero, F. (2014). Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. *Expert Systems with Applications*, *41*, 841–852.

Chachadi, K., & Nirmala, S. (2022). Voice-based gender recognition using neural network. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020)* (pp. 741–749). Springer.

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).

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2022). Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, .

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*.

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*, .

Dorn, R. (2019). Dialect-specific models for automatic speech recognition of african american vernacular english. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 16–20).

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

Farias, F., Lobato, W., Cruz, W., & Amadeus, M. (2022). Bilingual asr model with language identification for brazilian portuguese and south-american spanish, .

Garnerin, M., Rossato, S., & Besacier, L. (2021). Investigating the impact of gender representation in asr training data: a case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing* (pp. 86–92). Association for Computational Linguistics.

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.

Guo, L., Wang, L., Dang, J., Chng, E. S., & Nakagawa, S. (2022). Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. *Speech Communication*, *136*, 118–127.

Ha, J.-W., Nam, K., Kang, J., Lee, S.-W., Yang, S., Jung, H., Kim, E., Kim, H., Kim, S., Kim, H. A. et al. (2020). Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *arXiv preprint arXiv:2004.09367*, .

Hafeez, A. H., Mohiuddin, K., & Ahmed, S. (2014). Speaker-dependent live quranic verses recitation recognition system using sphinx-4 framework. In *17th IEEE International Multi Topic Conference 2014* (pp. 333–337). IEEE.

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.

Hrúz, M., & Zajíc, Z. (2017). Convolutional neural network for speaker change detection in telephone speaker diarization system. In *ICASSP* (pp. 4945–4949).

Hsu, W.-N., Tsai, Y.-H. H., Bolte, B., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6533–6537). IEEE.

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.

Huehnergard, J., & Pat-El, N. (2019). Introduction to the semitic languages and their history. In *The Semitic Languages* (pp. 1–21). Routledge.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, *9*, 2.

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl*, *7*, 20–24.

Khosravani, A., Garner, P. N., & Lazaridis, A. (2021). Modeling dialectal variation for swiss german automatic speech recognition. In *Interspeech* (pp. 2896–2900).

Kinoshita, K., Ochiai, T., Delcroix, M., & Nakatani, T. (2020). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7009–7013). IEEE.

Kuang, Z., Arachie, C. G., Liang, B., Narayana, P., DeSalvo, G., Quinn, M. S., Huang, B., Downs, G., & Yang, Y. (2022). Firebolt: Weak supervision under weaker assumptions. In *International Conference on Artificial Intelligence and Statistics* (pp. 8214–8259). PMLR.

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).

Li, B., Chang, S.-y., Sainath, T. N., Pang, R., He, Y., Strohman, T., & Wu, Y. (2020). Towards fast and accurate streaming end-to-end asr. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6069–6073). IEEE.

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

Lin, Q., Yin, R., Li, M., Bredin, H., & Barras, C. (2019). Lstm based similarity measurement with spectral clustering for speaker diarization. *arXiv preprint arXiv:1907.10393*, .

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

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.

McCrocklin, S., Fettig, C., & Markus, S. (2022). Salukispeech: Integrating a new asr tool into students' english pronunciation practice. *Pronunciation in Second Language Learning and Teaching Proceedings*, *12*.

Meng, Z., Mou, L., & Jin, Z. (2017). Hierarchical rnn with static sentence-level attention for text-based speaker change detection. In *Conference on Information and Knowledge Management* (pp. 2203–2206).

Nakatani, T. (2019). Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech*.

Nguyen, T. T., Pham, M. T., Nguyen, T. T., Huynh, T. T., Nguyen, Q. V. H., Quan, T. T. et al. (2021). Structural representation learning for network alignment with self-supervised anchor links. *Expert Systems with Applications*, *165*, 113857.

O'Shea, T. J., Corgan, J., & Clancy, T. C. (2016). Unsupervised representation learning of structured radio communication signals. In *2016*

First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE) (pp. 1–5). doi:10.1109/SPLIM.2016.7528397.

Qwaider, C., Chatzikyriakidis, S., & Dobnik, S. (2019). Can modern standard arabic approaches be used for arabic dialects? sentiment analysis as a case study. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics* (pp. 40–50).

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*, .

Ragni, A., Knill, K. M., Rath, S. P., & Gales, M. J. (2014). Data augmentation for low resource languages. In *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association* (pp. 810–814). International Speech Communication Association (ISCA).

Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, *59*, 235–244.

Roonizi, A. K., & Jutten, C. (2021). Band-stop smoothing filter design. *IEEE Transactions on Signal Processing*, *69*, 1797–1810.

Sasajima, M., Kitamura, Y., Ikeda, M., & Mizoguchi, R. (1996). A representation language for behavior and function: Fbrl. *Expert systems with applications*, *10*, 471–479.

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).

Sezgin, E., & D'Arcy, S. (2022). Editorial: Voice technology and conversational agents in health care delivery. front. *Public Health*, *10*, 887492.

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. (2020). Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognition Letters*, *131*, 213–218.

Sharma, A., & Jayagopi, D. B. (2021). Towards efficient unconstrained handwriting recognition using dilated temporal convolution network. *Expert Systems with Applications*, *164*, 114004.

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*, .

Shum, S. H., Dehak, N., Dehak, R., & Glass, J. R. (2013). Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*, 2015–2028.

Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, *8*, e1045.

Silnova, A., Brümmer, N., Rohdin, J., Stafylakis, T., & Burget, L. (2020). Probabilistic embeddings for speaker diarization. In *Odyssey*.

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).

Tran, M., & Soleymani, M. (2022). Towards privacy-preserving speech representation for client-side data sharing. *arXiv preprint arXiv:2203.14171*, .

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1476–1485).

Zoughi, T., Homayounpour, M. M., & Deypir, M. (2020). Adaptive windows multiple deep residual networks for speech recognition. *Expert Systems with Applications*, *139*, 112840.