WAV2VEC 2.0 FOR IRISH ASR:
A MULTILINGUAL APPROACH TO UNDER-RESOURCED LANGUAGES

Sarah Faste
Rijksuniversiteit Groningen
Campus Fryslân

A thesis submitted
in partial fulfillment of
the requirements for the degree of
Master of Science in Voice Technology
August, 2022

Supervising advisors:
Shekhar Nayak and Matt Coler

Abstract

Under-resourced languages have very little or no data recorded, making it an exceptional challenge to create automatic speech recognition systems for them. Using multilingual methods, new models have been developed to use data from other languages to create under-resourced language systems with very small datasets. The Wav2Vec 2.0 XLSR-53 is a large multilingual model that uses 53 languages to pre-train in a self-supervised manner. In this research, I conclude that it is possible to fine-tune the XLSR-53 model with less than 5 hours of data and achieve a WER of less than 50%. Using the Irish dataset from Mozilla's Common Voice with only 4 hours of validated data, the multilingual Wav2Vec 2.0 XLSR-53 is able to achieve a WER of 46.88%.

Abstract **1**

1 Introduction **3**

2 Literature Review **7**

2.1 Applications of Automatic Speech Recognition 8

2.2 Challenges of Automatic Speech Recognition 10

2.3 Understanding Automatic Speech Recognition 12

2.4 Multilingual Automatic Speech Recognition 15

2.4.1 Deep-Neural Networks (DNN) 15

2.4.2 Multilingual BERT 16

2.4.3 Wav2Vec 2.0 17

3 Research Question and Hypothesis **20**

4 Methodology **21**

4.1 Data 21

4.2 Model 21

5 Results and Discussion **23**

5.1 Challenges 24

5.2 Limitations and Recommendations 24

6 Conclusion **26**

References **28**

1 Introduction

In the last few decades, advances in technology have achieved exponential growth. With the improvement of computers and machine learning techniques we are able to bring technology into new disciplines everyday. Computers operate with their own language, so it is fortunate that we can use them to study human language. The use of speech technology by combining the fields of linguistics and computer science has been an incredible human achievement. It allows humans to solve all kinds of problems that may otherwise be difficult or impossible to achieve without the help of computers. There are many applications, including "Speech Coding, Text-to-Speech Synthesis, Speech Recognition, Speaker Recognition and Verification, Speech Enhancement, Speech Segmentation and Labeling (Transcription), Language Identification, Prosody, Attitude and Emotion recognition, Audio-Visual Signal Processing and Spoken Dialog Systems" (S & Chandra, 2016). From these, a prominent facet of speech technology is automatic speech recognition (ASR). This form of technology takes human speech as input, forms an understanding in the computer, then converts the speech into text (Benkerzaz, 2019). Today this technology is used for various forms of translation, transcription, forensics, security, medicine, and a wide range of everyday tasks. This provides new opportunities for technology to aid with issues such as accessibility for those with disabilities, improved medical care, and language preservation for those in danger of extinction.

With nearly 7,000 languages in the world, only a small fraction have sufficient resources available to them. With more than 90% of the world's languages in danger of extinction by the end of the century, it is crucial that we find solutions for language preservation (Prud'hommeaux & Jimerson, 2018). Language is an important part of culture, identity, and daily life, so it is pertinent that we use the tools available to us to support the continued use of every language. When a language is lacking these resources, researchers call it an under-resourced (URL) or a low-resource language (LRL). For continuity purposes, we will refer to these languages as under-resourced for the duration of this research. The concept of an under-resourced language can suggest different meanings. In this case, we will be referring to under-resourced languages by the definition of having very little to no recorded data or labeled data available (Prud'hommeaux & Jimerson, 2018; Cieri et al., 2016). This suggests that audio, video, and written data for an under-resourced language is very limited. It is necessary to make this distinction because although a language may be spoken by many people, it may not have much linguistic data recorded. Regardless of the number of speakers or the level on the scale of endangerment, they can still be considered an under-resourced language.

Many of these languages face extinction due to majority languages pushing them out of human use. Colonization has shifted language use throughout history and colonized groups are often forced to speak the language of the invaders in a phenomenon called linguistic colonization. Through the years, the lasting effects of this issue combined with an increasingly open world has caused these languages to die out. Unfortunately, linguistic colonization has caused both written and oral history of languages to disappear to the extent that revitalization may even be impossible. Even if the spoken language is preserved, the written form may be lost due to records being destroyed. Speakers of these languages either completely lose their language, or the language use is confined to small communities often spoken at home. While many members of these communities continue to speak majority languages, there are many members who wish to reclaim or save their dying language as a part of their cultural history and identity. As Besacier et al. writes, "individual and community memories, ideas, major events, practices, and lessons learned are all preserved and transmitted through language." Thus, preserving the language also preserves far more than just the language itself.

Automatic speech recognition can help preserve endangered languages by offering opportunities to learn through language learning apps and translation services. It also allows speakers to use it everyday with documentation dictation services, hands free texting, and security measures in businesses. Since it offers so many options in daily life, it can help improve continued use for future generations. If technology is so widely used everyday, then it would be a great benefit to use it with their native (or second) language. It also allows other people to engage with the language with translation services and subtitles so speakers of other languages can learn it themselves, or interact with media in the original language.

With increasing data on the state of the world's languages and the rise of technology to offer new options, we now have more opportunities than ever before to create these systems. However, the problem here lies in many different factors: time, money, resources, and expert assistance. While these new technologies such as ASR exist, the solutions they provide are difficult to implement if there are no experts, time, money, or resources allocated to language communities in order to actually create them. Advocating for language preservation is an important task for communities who wish to implement this technology, and it can be a long and difficult process. The goal with language preservation is to provide communities with the proper language resources to successfully provide tools for preservation, along with education to sustain the systems for the future. In addition to endangered languages, the technology can also be implemented for languages with large numbers of speakers. While there are many speakers of the language, there may not be recorded data available. With ASR technology, these systems could provide services to large numbers of the world's population that are often overlooked.

The Irish language is an often overlooked European language that is spoken in The Republic of Ireland and Northern Ireland. Due to the island's colonization by the English, the Irish language has slowly been replaced with the English language. Recently, a large effort has been made by the Irish government and people to reclaim their language, and automatic speech recognition (ASR) systems could be one way to assist in that effort.

Describing any language in its entire history, linguistic properties, and cultural significance would likely take up several textbooks. Here we will present the basics of the Irish language in order to better understand it in both linguistic and historical context, and its role as an under-resourced language in automatic speech recognition. An Coimisinéir Teanga (The Language Commissioner) is a government role appointed by the President of Ireland to oversee language rights in the country according to the Official Languages Act. According to the official An Coimisinéir Teanga website, Irish belongs to the Celtic language family along with Scottish, Gaelic, and Manx (*About the Language*). It is spoken today in The Republic of Ireland and Northern Ireland in addition to any countries where individual Irish speaking expats have settled. Under the umbrella of

Irish, there are three main dialects of the language: Ulster, Connacht, and Munster which are named after the regions they are spoken in. As of 2016, The Central Statistics Office (An Phríomh-Oifig Staidrimh) has reported more than 1.76 million Irish speakers in the census, which is approximately 40% of the population (*Irish Language and the Gaeltacht - CSO - Central Statistics Office*, 2016). Officially it has been declared an endangered language by UNESCO (*Gaelic vs. Irish*).

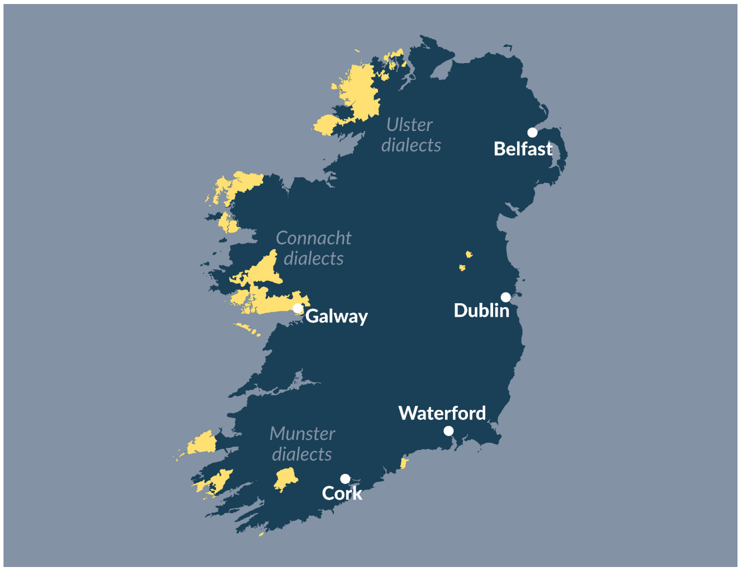

Fig. 1: Map of Irish Dialects
Image from *The Irish Language in Ireland*

Historically, Irish was the predominant language in Ireland until the 16th and 17th centuries with the invasion of England. Irish was steadily replaced with English as the predominant language, resulting in only 1% of the population being monolingual Irish speakers by the end of the 19th century (*About the Language*). While less than half of the population today identify as Irish speakers, this number is quite significant. The Irish people have put a significant effort into preserving their language through education and government efforts. Based on an article published by the European Commission, while the European Union recognizes it as an official language, it was not mandated to provide translators or interpreters for the language until earlier this year in January 2022. The efforts to secure it as an official EU language, gain translation rights, educate in schools, and offer learning opportunities on apps like Duolingo may help with its preservation (*Gaelic vs. Irish*). These accomplishments are a prime example of the determination and work required by under-resourced language communities in order to gain basic linguistic rights next to high-resource languages. In this case, the Irish speaking community has had some major successes and will likely continue to make strides for the Irish language in the coming years.

The rise of technology and automatic speech recognition could aid the cause for Irish language preservation. Different applications such as public transportation announcements, subtitles, and language documentation help create new environments to use Irish on a daily basis, and to also encourage new speakers as well. This helps to keep the language alive since it is convenient and can be used everyday, and also allows new speakers to use it with technology in a variety of situations. The variety of ASR applications offers the opportunity for both Irish and other under-resourced languages to survive (and hopefully thrive) in the future.

In contrast to under-resourced languages, a high-resource language is a language with a large amount of accessible recorded data. In this paper, we will look at the use of high-resource and multilingual language models to improve word error rate (WER) in automatic speech recognition models for under-resourced languages. Automatic speech recognition does not operate on just one type of model. There are different types, techniques, models, and forms of evaluation that make up an ASR system (Arora & Singh, 2012). While many modern models achieve a high accuracy, under-resourced languages cannot always benefit from these systems because of the added challenges. If a model requires a large amount of data, under-resourced languages are automatically at a disadvantage. If additional data is needed, researchers have developed techniques to attempt to solve this problem. Using high-resource languages or several other languages in tandem with the URL is a common technique, where using a high-resource is referred to as cross-linguistic ASR and using multiple languages is referred to as multilingual ASR (Sailor et al., 2018).

Multilingual methods are used in many different model types, including deep neural networks (DNN), the multilingual BERT model (M-BERT), and various versions of the Wav2Vec 2.0 model. Diwan et al. (2021) and Miao & Metze (2013) have performed research on hybrid DNN-HMM models for under-resourced languages. Huang et al. (2013) has conducted a study on a shared hidden layer multilingual DNN (SHL-MDNN), and overall these studies have shown promising results. In most experiments, the multilingual methods outperform the traditional monolingual or bilingual models by comparison. Furthermore, in studies by Conneau et al. (2021) and Xu et al. (2021), different Wav2Vec 2.0 models were tested against each other, and the multilingual models also outperformed their monolingual, bilingual, or smaller multilingual counterparts.

This high performing multilingual model is the Wav2Vec 2.0 XLSR-53 by Facebook AI. This open source model uses self-supervised methods to train on a large amount of unlabeled data, then fine-tune it on a smaller amount of labeled data (Conneau et al., 2021). This method makes it ideal for under-resourced languages as it can benefit from the large multilingual datasets in training, then get a better result from a smaller amount of data. In this research, we will use the XLSR-53 model fine-tuned with Irish data to determine if we can achieve a WER below 50% based on previous models.

2 Literature Review

Humans often use speech as a primary form of communication. It is an important part of understanding human language and how we communicate with one another. So as the development of the computer over the decades has improved, so has the study of computational linguistics and how we can communicate with machines through speech. Automatic speech recognition (ASR) is the method of taking spoken language and converting it into text through a computer, thereby creating a system that allows machines to understand human language. Another way of describing ASR is speech-to-text (STT) (S. Arora & Singh, 2012, Saini & Kaur, 2013). We can trace the existence of speech recognition to as early as 1920, however, the 1980's provided a significant leap in progress due to the new use of Hidden Markov Models (HMMs) (Arora & Singh, 2012, Gaikwad et al., 2010). When discussing the topic of speech recognition, it is important to acknowledge that language and communication is not only confined to spoken language. Humans have many other forms of communication including facial expressions, gestures, and deliberate hand motions that allow us to communicate through sign languages and non-verbal signals. In this particular study we will focus on spoken language as it pertains to speech recognition. However, it is important to address these other forms of communication because they are both meaningful to human communication, and they pose possible problems to the full functionality and accuracy of ASR systems (Arora & Singh, 2012, Petkar, 2016). Overall, ASR systems have a plethora of applications in a variety of different fields and can offer unexpected solutions to many different problems.

Many different fields in the world can use this technology to solve problems and create better solutions. It can be used in medicine, travel, transportation, communications, gaming, emergency services, language studies, security, historical preservation, and more. It opens up whole new possibilities for accessibility, safety, language preservation, and healthcare that can create a safer and more equitable world. However, as we will discuss in the challenges section, these systems do not always function as intended. They experience bugs, inaccuracies, and difficulties. The purpose of discussing both the applications and challenges of ASR is to understand how they function in real world scenarios and how we can improve them. It is also important to understand these applications and challenges since under-resourced languages are so highly impacted. By discussing the many different ways we can use ASR, we understand what URL communities are excluded from when their language is not included. In addition, by discussing challenges we can compare high-resource challenges with URL challenges and devise solutions to these problems.

In addition to the challenges and applications of automatic speech recognition, understanding how ASR and multilingual ASR functions is important for this research. Previous studies on multilingual ASR for under-resourced languages provide insight into how we can create new experiments for the future, and what those results mean. The context and results behind the research helps determine what kind of data is used in an experiment and which model is best for the desired outcome of the study. It is also beneficial to see how past research and this experiment are significant for under-resourced language systems in particular.

2.1 Applications of Automatic Speech Recognition

Applications of automatic speech recognition are varied for many different types of tasks. Today, translation is one popular application of ASR. Using applications or tools with ASR technology, we can travel around the world without the need for linguistic fluency. While ASR cannot replace a human learning a language, it can offer the ability to translate one language to another on the spot so that people can communicate with one another despite the languages in their repertoire (Yu & Deng, 2015). When especially reliable, this technology has potential to aid with international business opportunities or doctor-patient communication in the event that a language interpreter is not available. On the other side of this, it can also help people learn new languages. Some modern language learning applications or programs offer ASR technology for speaking practice and pronunciation study. Through the expertise of educators and ASR engineers, we can attempt to apply this as a learning tool for language learners around the world (Carrier, 2017). It can also help in under-resourced language education. Adding URLs to the language learning applications with this technology could be integral in getting new speakers comfortable with the language from anywhere in the world.

Transcription is another predominant use of automatic speech recognition. We can use speech to text to make tasks more convenient, but also for accessibility and safety. Using their voice, text dictation allows the user to send text messages, write documents, or receive written versions of voice messages (Yu & Deng, 2015). Some people may have trouble typing and may find it easier to dictate their thoughts orally. Writing or typing can be a very difficult task for some people. For example, if they experience a learning disorder such as dyslexia or if they do not have full physical use of their hands due to arthritis, the physical act of typing may be strenuous or painful. ASR transcription can provide an option for people to write essays, emails, or text messages without ever using their keyboard. It's an important step in accessibility for people who have previously had a difficult time performing writing tasks. It can also offer a safer option when replying to text messages while driving. Instead of looking down and replying to a text message, the user can use their voice to reply so they can keep their eyes on the road. By including URLs in this technology, we can offer safer and more convenient options to everyone who uses the technology regardless of language.

Professional dictation of documents is a growing field of ASR that is used in areas like the medical field and air traffic control. Medical dictation allows medical professionals to use ASR technology to chart patient information and document the situation differently than they normally would with traditional writing methods. Air traffic control can use this technology to monitor live operations or log information of controller-pilot speech (Dodiya & Jain, 2016, Raut & Deoghare, 2016). The main asset ASR offers these fields is an increased safety and efficiency. Lives can often be at stake in these situations, and it is extremely important to increase the chances of success. It is possible that dictation can remove extra time, stress, or physical exertion from the job to help these professionals perform at their best and keep an accurate record of their work. It is absolutely crucial that these systems are accurate due to the sensitive nature of the subject matter, so this is a highly important application of transcription. If this can be extended to URL systems, we can offer this service to professionals who work in minority languages and increase their chances of safety and efficiency in the workplace.

Subtitles are one of the more familiar uses of transcription in our everyday lives and have long been a function on television (Dodiya & Jain, 2016). In recent decades we have also seen them function on live streaming, social media platforms, and content streaming services. This can be a convenience for people who prefer subtitles with their content, but also provide an extremely important opportunity for accessibility. Subtitles offer an option for people who are deaf, hard of hearing, or non-native speakers of the language. With the option to read what is happening on the screen, people now have a far wider range of content to consume regardless of their capability of understanding the audio provided. This can be used with games, movies, or entertainment videos, but also for important news announcements and educational content. It opens the door for both recreational and necessary consumption of information across popular platforms for a whole population of people in the world.

Transcription also offers incredible potential for under-resourced languages in the area of language preservation. While not all URLs are endangered languages, many fall into that category. Today, linguists can use ASR technology to help preserve languages. By creating datasets with recordings of the language, we can increase records of the speech. In addition, the transcribed speech and the output of the ASR system can create large amounts of written data that would otherwise have been done manually. This is not always possible due to lack of speakers, time, funding, and experts to record it. If an ASR system can be implemented for a language, we can exponentially increase the recorded data for that language and possibly decrease the amount of time it takes to record that data. This method provides archives for the speakers of that language to use as they prefer. It can be used to educate future generations to prevent linguistic extinction, or to have a full record of the language in the event of an unfortunate extinction event.

Automatic speech recognition can also be used for many basic tasks such as setting timers, scheduling an appointment, phone calls, searching the internet, or performing tasks in video games (Yu & Deng, 2015). These benefits can be convenient, like setting a timer while you cook when your hands are occupied, or playing a video game and interacting in a new and fun way. However, these tasks can also increase accessibility for many users. Voice based tasks can be extremely beneficial for those with limited mobility or use of their limbs. Being able to contact family members, turn off the lights, and search the internet without moving or using any limbs offers more independence for these users, and could even be used to contact emergency services in the event of an emergency. This offers people the opportunity to become more independent. There are many people who may need extra assistance in their home due to their age or physical limitations. It is possible that completing small tasks with ASR can increase independence for these individuals and improve their quality of life and self confidence. Although some people may prefer not to use the system, we can offer people the choice to decide for themselves. By creating these systems with URLs, we can offer these services to a much weirder range of people and help to provide accessibility opportunities no matter what language the person speaks.

Speaker recognition is a particular subset of speech recognition that uses ASR technology and applies it to an individual speaker using biometrics. Human voices are unique to one another based on vocal tract length, sex, cadence, emotion, health, accent and much more. Based on this, we can create ASR systems for individuals when we study and apply these vocal parameters. It can be used for applications such as authentication, identification, surveillance, and forensics. One such application of authentication or identification is banking, where voice can be used to identify the customer as a unique security measure. The customer can use their voice as a replacement of (or alongside) passwords or other forms of identification to ensure that their belongings are secure. Adding under-resourced languages to the system allows people to use their native language for security purposes. When used for surveillance, it is necessary to comply with ethical practices. When used according to law, this application can help to collect data to improve customer service experience or gather further data for an ASR dataset which can be highly beneficial to building up URL datasets. Finally, speaker recognition can offer new revolutionary technology to forensic investigative work. If a crime is committed and law enforcement has a recording of the voice of the suspect, speaker recognition can help identify the individual. Adding more languages to this application increases the chances of catching the suspect, since they may not be speaking the majority language of the system. There are also several ethical factors to consider with this application according to where it is used, and it needs to be completely accurate if it is used in a court of law. However, it does offer new opportunities to solve current crimes or cold cases with voice analysis (Singh et al., 2012).

2.2 Challenges of Automatic Speech Recognition

Creating a system that understands human languages can present many challenges. By understanding these possible difficulties, we can solve problems as they arise or simply avoid them to the best of our ability. Since ASR systems deal with human speech, automatic speech recognition has also inherited the problems found in regular human speech such as stutters, pauses, false starts and repetitions (S. Arora & Singh, 2012). This can be difficult to translate to a written form when the speech is converted into text. When creating a system, we need to determine how these sounds will be processed. Will they be included in the written transcript? Edited out for clarity? Will the machine even recognize them as utterances? These questions are important to consider when building an ASR system, or else the system will encounter errors when faced with unknown input. Furthermore, by understanding the challenges faced by ASR systems, we can give context to the models created by past researchers presented in this paper and also understand that URLs face these challenges in a unique way and how that affects an ASR system.

Each person has a unique way of speaking, which is why we are able to have the branch of speech recognition that is speaker identification or verification (speech recognition for an individual speaker) (S & Chandra, 2016). However, this can lead to several roadblocks when building a traditional ASR system. Individuals will vary by speaking style, dialect, pronunciation, and sex. Someone may have a raspy voice, or have a strong accent in their second language. They may speak with a specific dialect, or they may be speaking with high emotion. Their voice may even change depending on their wellness. With such variability between people, variability with phone pronunciation in individuals, and extrinsic factors like wellness, we need to be prepared to create systems that can handle all of this information (Petkar, 2016, S. Arora & Singh, 2012). They need to recognize speech with multiple pronunciations, dialects, styles, and voice changes since that is the nature of human speech, and it is how the users of the system will naturally interact with it.

Considering all of this information, we also understand that both linguistic and non-linguistic features can present a challenge for many ASR systems. Non-linguistic features like coughing, sneezing, lip-smacking, and laughing need to be recognized by the system in some way so it can function properly. This also applies to linguistic features like word boundary ambiguity, homophones, names, and numbers. Uncertainty where one sentence ends and another starts could prevent recognition, just as well as homophones, or an ambiguity related to numbers or names of people. The system needs to understand when the user is using the name of a person, a number, or one word over another when they sound similar. Other non-linguistic sounds can be external noise. Background chatter, traffic noise, beeping appliances, location echo and many other noise factors can get in the way of ASR success. To reduce errors in the system, it must be able to filter out the noise to understand the current speaker (Petkar, 2016).

Speakers can also jump between languages in a phenomenon known as code-switching. The speaker will use words or phrases from more than one language in the same sentence or phrase (Diwan et al., 2021, Yılmaz et al., 2018). This could be due to the multilingual nature of many languages, using names of places, people, street names, or simply from the comfort level of the speaker from one language to the next. Multilingual or code-switching speech can pose difficulties for ASR systems because they need to be trained to determine this change in languages. Even trying to solve this by using multilingual data can pose more problems. Using the wrong language in tandem with our target language could actually degrade the performance of the system, so choosing another language is a crucial decision. Thus code-switching is a common challenge in modern ASR systems (Diwan et al., 2021).

While we often focus on the ASR system itself, it is just as important to pay attention to the tools that we use to create it and gather the data. If building a dataset is also a part of creating the ASR system, there are many factors to account for. Microphone quality, recording equipment, number of speakers, sample rate, noise levels and types, consistency of procedure between speakers, and privacy rights of speakers are all crucial factors in creating a dataset (Chitu & Rothkrantz, 2012, *Data Protection under GDPR*, Petkar, 2016). If the system will run on an already existing dataset, then it must comply with the data license and be given whatever credit is necessary (*Common License Types for Datasets*). In addition, the system can face everyday technical problems during any stage. Coding is an exercise of trial and error and ASR systems are no exception. Taking the time to solve any technical problems that arise at any point will be a natural part of building a new system or even re-working an existing one.

As previously stated, automatic speech recognition systems for under-resourced languages carry a particular challenge due to the lack of linguistic data available. Many state of the art models require a significant amount of data to achieve an impressive result. While this is not necessarily a problem for high-resource languages, it presents a problem for the rest that don't have the data available to take advantage of these models. This is a significant challenge because it excludes the use of many different languages and limits these state-of-the-art models to a small number of languages. With this challenge, there are a few different methods to address under-resourced languages in particular. Data augmentation allows us to create new data and mimic a larger dataset by changing different aspects of the audio. In addition, cross-linguistic and multilingual approaches use high-resource languages or groups of languages to train systems in addition to the under-resourced languages. Both of these methods act as a sort of supplement for the lack of data in URLs and we will go into further depth of these methods later on (Sailor et al., 2018).

Earlier in this section, we discussed human variation of speech, noise, code-switching, and technical equipment as challenges of ASR. While these factors present a challenge for high-resource ASR, they present an even bigger challenge for URLs. Human variation is needed in our research to improve these systems, but URL data is very limited, and often does not have a large enough sample size to achieve this variation. This includes non-linguistic features like coughing or sneezing, and training the system to recognize and filter out noise. Without this data, URL systems have less of a chance for accuracy and high performance. Code-switching is very important for under-resourced languages because it is so often used by people who speak these languages in tandem with a majority language. If a community uses code-switching very frequently, an accurate URL system will not function if it cannot understand both languages. Finally, URL communities (more than high-resource ones) do not have the money and resources to purchase state-of-the-art equipment for their research. While we can use basic supplies to record audio in cases such as crowd-sourced datasets like Common Voice, it is still limited to one type of data available. In summation, these challenges are generally more difficult for under-resourced language ASR systems and offer additional difficulties to reach the standard of current state-of-the-art ASR systems. To build an ASR model for under-resourced languages, a background knowledge of these particular challenges is important for understanding the problems that the model may face. By understanding the context, we can create better performing models and conduct more efficient research for URL automatic speech recognition.

2.3 Understanding Automatic Speech Recognition

To understand the structure of an ASR system, we must first understand the types of speech associated with speech recognition, and how these have changed over time with developing technology. Isolated words or utterances, connected word, and continuous speech are the different types associated with ASR. While isolated and connected word have trouble distinguishing boundaries and are more limited to individual utterances, continuous speech is closer to how humans actually speak. It is the most difficult to achieve because the system needs to be trained to recognize the boundaries between words, but it is the most advanced and widely applicable type of ASR today (S. Arora & Singh, 2012).

Automatic speech recognition systems are also categorized by two different types of speaker systems. Speaker dependent systems were often used in early speech recognition systems and require that each utterance is recorded in order to be recognized. In contrast, speaker independent systems need a variety of speaker recordings to function with a larger range of users in mind. While these systems are more difficult and less accurate, they offer a much wider variety of ASR to a wider range of people. Through the years, ASR engineers and researchers have been able to make many improvements and speaker independent systems are the primary form of speech recognition today (S. Arora & Singh, 2012, Besacier et al., 2014).

Automatic speech recognition is also characterized by different modeling techniques. The acoustic-phonetic approach, pattern recognition approach, and artificial intelligence approaches are the three primary approaches. The acoustic-phonetic approach is based around finding acoustic properties or labels for phonemes. Theoretically, with this approach we can label each possible phoneme in a language with acoustic properties and build a system with that information. However, while the earliest approaches used this method, it is not a popular method today when compared with more modern approaches due to the tedious and often inefficient nature of labeling each phoneme without any pattern recognition to assist (Arora & Singh, 2012, Gaikwad et al., 2010).

The pattern recognition or pattern-matching approach has four phases: feature measurement, pattern training, pattern classification, and decision logic. Based on labeled training samples, it uses a training algorithm to create speech pattern representations. The speech input and the new patterns are then compared to determine the output as recognized speech. In the last 60 years, the pattern recognition approach has become the most widely used method. Stochastic models are very common and based on probability. Popular statistical learning methods include Hidden Markov Models (HMM). HMM-based statistical methods are varied and often used for modeling time series data and speech classification. This stochastic method using HMMs revolutionized ASR systems in the 80's, and has consistently contributed to new methods through the present day (Arora & Singh, 2012, Gaikwad et al., 2010). While it has its own difficulties such as making incorrect assumptions and hurting system performance, the inclusion of a statistical model can improve efficiency from template based methods.

Finally, the artificial intelligence approach to ASR is basically a combination of the acoustic-phonetic and pattern recognition approaches. It is considered to be a knowledge based approach, which means that an expert level knowledge of the language is needed in order to build. This linguistic knowledge is coded into the system and is highly accurate with modeling speech variations. So it takes the linguistic properties used in acoustic-phonetic ASR and the analytic and decision making aspects of pattern recognition. However, this method is difficult to achieve because of scarcity or difficulty to obtain linguistic experts for the task. This makes it impractical for many modern systems that do not have the budget or ability to hire these experts. Thus, pattern recognition remains the most widely used with many efforts to improve throughout the years (Arora & Singh, 2012, Gaikwad et al., 2010).

The basic architecture of a traditional ASR system is categorized by stages. These are signal processing, feature extraction, acoustic modeling, language (pronunciation or lexical) modeling, and decoding (hypothesis or recognition). This classic architecture is based on the stochastic HMM method (Besacier et al., 2014, Saini & Kaur, 2013, S. Arora & Singh, 2012, Yu & Deng, 2015). Basically, the system takes in a raw waveform and removes the noise and any channel distortions from the audio signal. It then extracts the feature vectors or phonemes from the audio, breaking down the basic sound units so the acoustic model can understand them. The acoustic model takes the knowledge of the feature vectors and generates a score for the sequence of features. The language model then estimates the statistical probability of the sequences of words based on the knowledge fed about the phoneme sequences and linguistic information about the language. Finally, the decoder turns the estimation of the acoustic and language models and turns it into output. The best scoring estimation of word sequences is presented and we get the text as output (Yu & Deng, 2015).

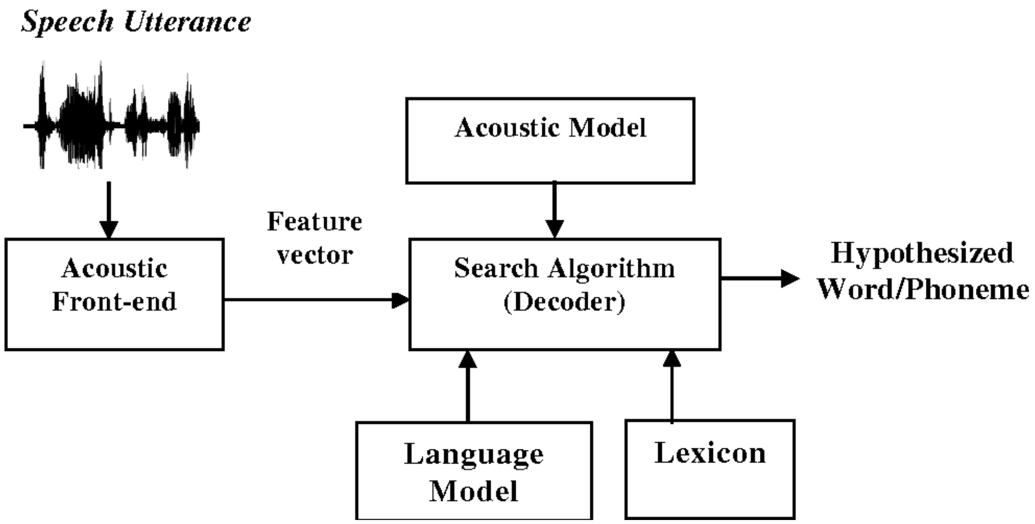

Fig. 2: Speech Recognition Architecture
Image from *A Review on Automatic Speech Recognition Architecture and Approaches*

In order to determine how accurate the system is, there is an evaluation stage after the model has been trained. There are many different methods of evaluation for ASR systems including Single Word Error Rate (SWER), Command Success Rate (CSR), Phoneme Error Rate (PER), Character Error Rate (CER), and Word Error Rate (WER). Word Error Rate (WER) is a common metric for evaluation and works well at the word level versus phoneme level for evaluation of ASR systems (Arora & Singh, 2012, Besacier et al., 2006, Gaikwad et al., 2010). Character Error Rate (CER) is another common evaluation technique which focuses on specific character errors versus full words. It is useful alongside WER and is also useful for alphabets that use characters, such as Mandarin Chinese. In this case, we will focus on WER for evaluation since it is a very widely used metric and comparable to many other models. It is calculated using this equation:

$$WER = \frac{S + D + I}{N}$$

S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

The equation is implemented in the code during the evaluation portion once the fine-tuning has been completed. The word error rate describes the basic percentage of how many words are incorrectly deciphered by the system. With this method, we can evaluate and determine the accuracy of our system in order to understand it better and to make further improvements going forward.

2.4 Multilingual Automatic Speech Recognition

Multilingual ASR is a method of automatic speech recognition that uses a cross-lingual approach. Multilingual models are able to share linguistic information across different languages. The goal is to create a model that can learn from multiple languages in order to outperform bilingual and monolingual models. This can be a great asset to under-resourced languages by using high-resource languages or several URLs to create a larger set of data. However, it is possible to encounter reductive learning with these systems, so researchers must be aware of this when using and creating multilingual models.

There are many different models used for multilingual ASR, and they are often chosen based on the specific research topic. Under-resourced ASR research often uses the multilingual technique to improve upon a small dataset, or uses several small datasets to create a system. Previous experiments have used a variety of these models including Deep Neural Network based models, the multilingual BERT model, and Wav2Vec 2.0 model versions. In the section below, we will briefly discuss some previous experiments with multilingual methods and further discuss the model chosen for this experiment.

2.4.1 Deep-Neural Networks (DNN)

Deep neural networks are a great asset to ASR, and are used often in modern systems. However, they require a large amount of data to achieve the best accuracy and this becomes a problem for under-resourced languages. Many researchers, such as those explored here, have still used DNNs to research URLs and created new methods to compensate for the data issue. We will explore some past research to understand the options available, and briefly discuss why these methods were not chosen for this research.

Diwan et al. (2021) explore the use of deep neural networks for under-resourced Indian languages in code-switching and multilingual situations. In this research, one of the models used was an DNN-HMM or Deep Neural Network and Hidden Markov Model with a time delay neural network (TDNN). Using the Kaldi toolkit, they build the model to compare with a classic HMM-GMM model for their experiment along with an end-to-end model. They used approximately 600 hours of data from 6 different Indian languages including Hindi, Marathi, Odia, Telugu, Tamil, and Gujarati as well as code-switched data in Hindi-English and Bengali-English. The authors take care to, "1) consider the influences of three major language families – Indo-Aryan, Dravidian and Austro-Asiatic, which influences most of the Indian languages [4], 2) cover four demographic regions of India – East, West, South and North, and, 3) ensure continuum across languages."

(Diwan et al., 2021). The results show that WER improved from multilingual to monolingual in Tamil, but not for the other languages (Diwan et al., 2021) This could be the result of overfitting with the data, but could be further discussed in future research.

Miao & Metze (2013) conduct research on under-resourced languages using dropout and multilingual methods in CD-DNN-HMM models. This context dependent, deep neural network and hidden markov based model has limited function with less than 10 hours of data. In general, DNN based architectures achieve impressive results, but need a larger amount of data to run. Their proposal uses dropout to avoid overfitting, and multilingual data to increase data abilities for low resource datasets. Like all DNN-HMM systems, the model is different from a traditional HMM-GMM system in that it does not use gaussian mixture models. These are replaced with deep neural networks. For their experiment, Miao and Metze use 3 languages from the GlobalPhone corpus: German, Spanish, and Portuguese. The Spanish and Portuguese data is used for multilingual training, and German is used for evaluation. Their results suggest that using both dropout and multilingual methods collectively improves WER by 11.6% with their 2 hour data and 6.2% with their 5 hour data (German evaluation data) compared with other models (Miao & Metze, 2013).

A shared-hidden-layer multilingual DNN based architecture makes use of many layers in order to share information across languages for a multilingual approach. In research conducted by Huang et al. in 2013, the authors explore the use of this model when compared to monolingual counterparts. The SHL-MDNN uses multitask learning to perform tasks simultaneously to increase learning efficiency, and can be pre-trained in both supervised or unsupervised methods depending on whether the softmax layer is language-specific or not. However, the authors studied this difference and recommend using labeled data as it performs better compared to the unlabeled data. They used multilingual training data of French (138 hr), German (195 hr), Spanish (63 hr), and Italian (63 hr). The results of the SHL-MDNN show that the model can reduce WER for all languages involved by 3-5% when compared with the monolingual model. In order to test if there is a difference in results for very different source and target languages (cross-lingual transfer) the model was tested on 139 hours of Mandarin Chinese and American English data. As opposed to WER used with the other languages, Mandarin was evaluated with CER and the results found an 8.3% improvement in Mandarin and a 4.6% WER improvement in English. While this model shows improvement in all languages including Mandarin, which is significantly different linguistically, the amount of data is important to achieve a better WER. The research tested data from 3 hours through 100+ hours and the error rate significantly improved by up to nearly 15% with more data (Huang et al., 2013).

2.4.2 Multilingual BERT

There are other methods besides DNNs that use pre-training, then fine-tuning stages. Wang et al. (2020) conducted research for extending the use of the multilingual BERT (M-BERT) with languages outside of the 104 previously studied with the model. Since the M-BERT has performed well with zero-shot cross-lingual tasks, the goal of the research is to extend this to under-resourced languages. They propose a method to "Extend" the model to create E-MBERT instead of training individually with each language, as that is expensive and time consuming. Their solution is better fitted to URLs in that it is more efficient monetarily and time-wise. They experimented with the model to fine-tune with 27 languages in total (11 of which are new to M-BERT model research) from LORELEI corpus. Each language is evaluated one at a time. Their results found that their extended version E-MBERT outperformed the original M-BERT on almost every language, even those that were already included in the original dataset. In addition, it improves upon efficiency as the BERT base trains over 4 days using 4 cloud TPUs (Tensor Processing Units), and the E-MBERT trains over 7 hours using only one cloud TPU, making it ideal for URL research (Wang et al., 2020).

2.4.3 Wav2Vec 2.0

The introduction of the Wav2Vec 2.0 model in late 2020 by Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli provided a significant asset to the field of automatic speech recognition as a whole. This model is self-supervised, which means it is trained on unlabeled data. Labeled data has corresponding transcribed text, where unlabeled data does not. The model itself has two primary steps of training. The first step is self-supervised and is trained on unlabeled data, and the second step is supervised fine-tuning which uses labeled data. Once the data has been fine-tuned, the system can make phoneme or word predictions to create text (Baevski et al., 2020). The system also takes after the BERT system, where it uses the same form of masked sequence modeling, use of transformers, and split pre-training and fine-tuning (Yi et al., 2021). This methodology is significant for under-resourced ASR in particular, because it offers solutions to the unique challenges faced by these systems. It offers the opportunity to use less data and unlabeled data that can previously be a significant challenge for many under-resourced languages. We will discuss further the details of this significance and the WER results of the system later in this section.

In the original proposal of Wav2Vec 2.0 in 2020, the authors concluded that "on the clean 100 hour Librispeech setup, wav2vec 2.0 outperforms the previous best result while using 100 times less labeled data." (Baevski et al., 2020). The model has also been used in previous under-resourced language research with different methods. Cheng Yi et al. (2021) study the use of Wav2Vec 2.0 with various different under-resourced languages. They use both the wav2vec2.0-base and wav2vec2.0-large models to fine-tune with the CALLHOME corpus of Mandarin Chinese, English, Japanese, Arabic, German, and Spanish. Each has about 15 hours of transcribed speech. They compare their results with previous experiments and find an improvement in WER and CER in every language, but English far more than the others due to Wav2Vec 2.0 being trained with English data (Yi et al., 2021). Based on this previous study, we can conclude that Wav2Vec 2.0 has impressive results with small amounts of data and is a good candidate for URL research.

The Wav2Vec 2.0 model also has different versions available for use. One large version is the XLSR-53, which is trained with a combination of 3 datasets: Common Voice, BABEL, and Multilingual Librispeech (MLS). This new large dataset consists of 53 languages and is used to train the Wav2Vec 2.0 XLSR-53 system before it is fine-tuned with more data (Conneau et al., 2021). For our research purposes, this model offers efficiency with the pre-trained method, ability to work with very little data, and available resources based on ASR expertise and will be used in our experiment.

The first self-supervised training step begins with a multi-layer convolutional neural network (CNN) that encodes the raw speech waveforms into latent speech representations. This CNN is a feature encoder, in which the first convolutional layer is normalized so that the channel sequences have zero mean and unit variance, and is then followed by a GELU activation function. The resulting latent speech representation spans are then masked before being fed to the Transformer (Baevski et al., 2020, Vaessen & van Leeuwen, 2022).

The feature extraction output of latent representations are then given to a Transformer network. In this stage it uses relative positional embedding via a convolutional layer. According to the authors they "add the output of the convolution followed by a GELU to the inputs and then apply layer normalization" (Baevski et al., 2020). This results in contextualized representations from the generated sequence. From these speech representations produced by the feature encoder, the system then uses product quantization to choose a certain number of representations, then connects them together. Finally, a Gumbel softmax layer is implemented at the end of the process as the activation function (Baevski et al., 2020, Vaessen & van Leeuwen, 2022).

The initial training is a pre-training stage, because the system will be trained (or fine-tuned) again with labeled data. According to Baevski et al, "The training objective requires identifying the correct quantized latent audio representation in a set of distractors for each masked time step." Which accurately sums up the training sequence outlined previously.

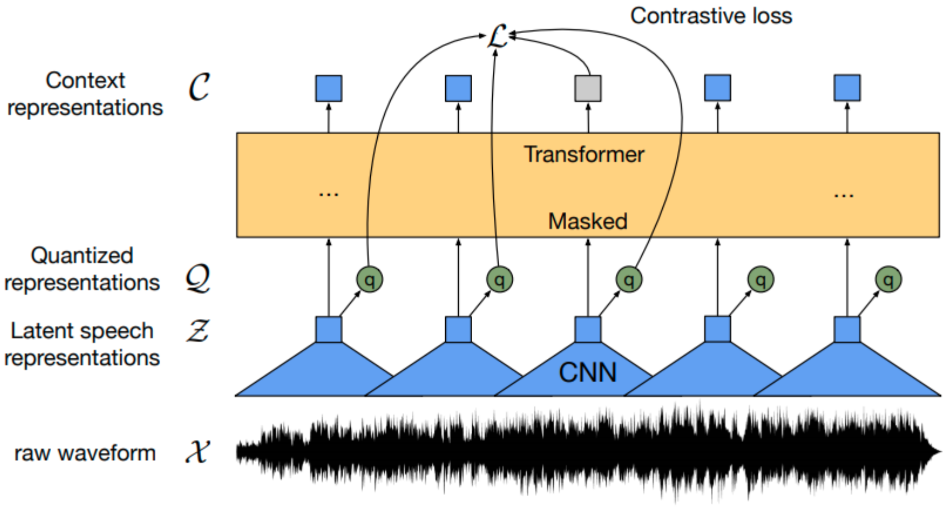

Fig. #: Framework of Wav2Vec 2.0
Image from *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*

In automatic speech recognition, the latent speech representations generated in pre-training can be shared across languages. In this case, we can use multilingual data to train systems, then fine-tune them with a specific language. A version of the Wav2Vec 2.0 model XLSR-53 functions in this same way. According to the official Wav2Vec 2.0 github page, the dataset used to train the large XLSR-53 model consists of 53 languages, and is a combination of 36 languages and 3.6k hours of Common Voice, 17 languages and 1.7k hours of BABEL, and 8 languages and 50.7k hours of Multilingual Librispeech (MLS) data. The self-supervised model is trained on the large dataset without corresponding transcriptions, then fine-tuned with a different dataset. The model makes use of cross-linguistic methods to supplement for a very small dataset which makes it ideal for under-resourced languages. Even further, the multilingual model has significant advantages with sharing linguistic information across languages. While the cross-linguistic approach can be effective with monolingual models such as the XLSR-English, the XLSR-53 multilingual model outperforms the monolingual model significantly (Conneau et al., 2021).

In their research, Conneau at al. (2021) compared the XLSR-53 models with previous work, as well as other XLSR models including XLSR-English, XLSR-monolingual, and XLSR-10 (a multilingual model with 10 languages). The purpose of the research is to test the effectiveness of these models on unseen languages, or languages that differ from the languages used in training. Using both BABEL and Common Voice data to fine-tune the models, the results find that "on the Common Voice benchmark, XLSR shows a relative phoneme error rate reduction of 72% compared to the best known results. On BABEL, our approach improves word error rate by 16% relative compared to a comparable system." (Conneau et al., 2021). Based on these results, the authors released the large XLSR-53 multilingual model since it showed potential for future cross-lingual and multilingual research, especially for under-resourced languages. In a part of their Common Voice fine-tuning results, they found that using high-resource languages in multilingual models may actually increase error rates "due to interference." (Conneau et al., 2021). So the models may be uniquely suited towards under-resourced language research.

In another Facebook AI study, Xu et al. (2021) study zero-shot cross-lingual methods with the XLSR-53 model for phoneme recognition research. As part of the study, they compared non-pre-trained, monolingual (wav2vec2-LV60K), and multilingual (XLSR-53) models by fine-tuning each with 13 different languages using Common Voice test data. As a result, every single language outperformed the other models using the multilingual XLSR-53 model (Xu et al., 2021). While it is possible for multilingual models to be reductive in learning, this previous research shows positive results using the XLSR-53 model with small amounts of data and is a good fit for this particular research for Common Voice's 4 hour Irish dataset.

3 Research Question and Hypothesis

With the consistent need for under-resourced language models and the implementation of new methods, past research has shown that we can use modern models with cross-linguistic and multilingual methods to supplement the lack of data in smaller datasets. A particular challenge for small datasets is those with an extremely small amount of data. Some corpus are low on resources, but can still offer hundreds of hours in data. A prevalent question in under-resourced language research is whether or not we can create systems with almost no data. This leads to the question and hypothesis presented in this research.

Research Question: Can we use multilingual data in automatic speech recognition systems to train models with less than 5 hours of data?

Hypothesis: Using a multilingual pre-trained wav2vec 2.0 model fine-tuned with less than 4 hours of data, we can achieve a WER of less than 50% based on previous models.

Based on research by Conneau et al, the Wav2Vec 2.0 XLSR-53 model shows improvement from monolingual models, and further improvement particularly for under-resourced models due to its cross-linguistic and self-supervised approach (Conneau et al., 2021). Xu et al. (2021) also shows an improvement upon monolingual, bilingual, and smaller multilingual models with the XLSR-53 model (Xu et al., 2021). Furthermore, WER achieved on similar Wav2Vec 2.0 models by Abid Ali Awan and Manan Dey, we can predict a WER of 50% or less. Abid Ali Awan's Wav2Vec 2.0 model (version xls-r-1b) fine tuned with Common Voice Irish data[1] achieved a WER of 38.45% and Manan Dey's Wav2Vec 2.0 (version large-xlsr-53) fine tuned with the Common Voice Irish data[2] achieved a WER of 42.34%. Based on these results, we can hypothesize that using Wav2Vec 2.0 XLSR-53 fine tuned with the Common Voice Irish dataset will achieve a WER under 50%.

The intention of this research is not to create a state of the art system, but to contribute a working model for an under-resourced language for future research and experimentation with very small datasets. Using already created open source models is a great opportunity for under-resourced language communities to use ASR systems without large budgets or teams of experts. In this case, the model functions with Irish and can be adapted in the future for other languages as well. This model was adapted from open source models and can be further adapted for future research.

[1] https://huggingface.co/kingabzpro/wav2vec2-large-xls-r-1b-Irish

[2] https://huggingface.co/manandey/wav2vec2-large-xlsr-_irish

4 Methodology

The methodology is organized by two main sections: data and model. In data, the choices made, and all the relevant information about the dataset will be outlined. The model section also details the choices and relevant details about the research model. With this information, the details are outlined, reasoned, and presented to understand the methods while performing the experiment.

4.1 Data

The Dataset used to train the large XLSR-53 model is a combination of Common Voice, BABEL, and Multilingual Librispeech (MLS) that results in a large corpus of 53 languages (Conneau et al., 2021). The Wav2Vec 2.0 model does not require labeled data in this training process, but requires labeled data for the fine-tuning train process later on.

The dataset for fine-tuning in this research is the Mozilla Common Voice Irish dataset. Last updated July 4, 2022 the dataset has 4 hours of validated data and operates under a CC-0 (creative commons) open source license. It includes information on the sex, age, and "accent" (dialect) of each recording. The primary demographic of the dataset is young males under 40. All three Irish dialects (Ulster, Connacht, and Munster) are present in the dataset. Common Voice is also a crowd sourced dataset, meaning that the audio is recorded by individuals who volunteer to send in their voices and help to validate the voices of others' recordings.

The Irish dataset was chosen specifically for the very low data available. Using a dataset with only 4 hours of data presents the opportunity to do research on extremely small datasets to see what kind of results I could achieve, and offers the opportunity to try new methods to supplement the lack of data. In addition, it provided transcriptions alongside the audio of the recordings. This is necessary for the Wav2Vec 2.0 format as the pre-training is unsupervised and does not require transcriptions, but the fine-tuning stage requires labeled data. The dataset also included both male and female data. Some datasets I considered were separated by sex, and I preferred to use one with both male and female data in order to include a wider range of voice frequency. Finally, the dataset was compatible with my computer. Other datasets I considered were difficult to obtain or were incompatible with download. The Common Voice dataset could be viewed by my computer so I could study the format and apply that to my code.

4.2 Model

The model used for this experiment was chosen for its cross-linguistic aspects using multilingual datasets. The Wav2Vec 2.0 large XLSR-53 model is trained on 53 different languages from 3 different large datasets. Facebook AI's Wav2Vec 2.0 is open source code for public use and available on github[3]. In addition, the past research

[3] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md

conducted by Conneau et al. (2021) and Xu et al. (2021) on the model was convincing. The impressive results combined with the accessibility to the base code made it a good match for the research I wanted to perform. Already, there are several models built from this base code available on HuggingFace with different datasets and purposes. In this experiment, 3 models were used as a base code, then adapted for the Common Voice Irish dataset, computational abilities in google colab from disk space and RAM, and any other necessary details for proper functionality. While many other models showed promise such as the hybrid DNN-HMM models by Diwan et al. (2021) and Miao & Metze (2013), and the SHL-MDNN by Huang et al. in 2013, they were not the best fit for my research. In general, the models required more data than I had with the 4 hour Irish corpus. While they are equipped for under-resourced language research, they did not offer the same flexibility as Wav2Vec 2.0 for my personal research interests. In addition, the E-MBERT model by Wang et al. (2020) offers impressive results, but was not as accessible as the Wav2Vec 2.0 model. Conclusively, the Wav2Vec 2.0 is friendly for beginner users, has a significant amount of information and models available online, is open-source code, and works flexibly and efficiently with very small amounts of data. For all these reasons, it was chosen for this experiment.

The code for my model is currently available on github for public viewing[4]. For the process, the data was split for training and testing by combining the train and validation files, then the test files were used for evaluation. Training this model again on the Irish dataset is also referred to as fine-tuning. The model is pre-trained on the large multilingual data, then can be fine-tuned on a different dataset. In this case, I am fine-tuning the Wav2Vec 2.0 model with Irish data. It is important to note that the first training of this model is self-supervised, so the large multilingual dataset does not need corresponding transcriptions. However, the second part involving fine-tuning does require labeled data so the dataset used (in this case Irish) will require corresponding transcriptions.

In order to use this model, the dataset is loaded and prepared for the model, then using transformers we can discover the tokens and extract the features. Before training, all the audio is checked at the same sample rate and matched with what the model requires. In this case, all audio is sampled at 16kHz. The model is loaded with hyperparameters defined, then training begins. After completing these steps, the model is trained on the Irish training data then pushed to the HuggingFace hub. We can call models from the hub using a specific path, so with a simple request the trained model is called and evaluated using the test data. The model is evaluated using the word error rate (WER) metric. The goal of the evaluation is to reach a WER of less than 50% as stated in the hypothesis.

[4] https://github.com/SJFaste/ASR--Fine-tuning-Multilingual-XLSR-Wav2Vec2-with-Irish-Common-Voice

5 Results and Discussion

After the model was trained and evaluated, I was able to get a WER result. My hypothesis states that using a multilingual pre-trained wav2vec 2.0 model fine-tuned with less than 4 hours of data, my model can achieve a WER of less than 50%. The hypothesis is validated since the model achieved a 46.88% WER.

The training stage provided data on the training loss, validation loss, and WER in each epoch and step. The training results are as follows:

Training Loss	Epoch	Step	Validation Loss	Wer
0.0812	28.57	400	1.3952	0.6447
0.0692	57.14	800	1.5167	0.6315
0.034	85.71	1200	1.585	0.6155

The training presented a WER of 64% at step 400 and was lessened to 61% at the end of the training, with a validation loss of 0.034 and a training loss of 1.585 at step 1200. With the initial training finished, the model was then evaluated with the test data for a final WER. This is where I got the final result of 46.88%.

Achieving these results is significant because they are comparable with the other similar models. The model has achieved a WER of less than 50% based on the hypothesis, and that is a successful result for what I aimed to achieve in this experiment. State-of-the-art systems generally require a much lower WER, but for our research purposes this is a fully sufficient outcome. The purpose is not only to achieve a WER of below 50%, but to do so with open source materials and under-resourced language data. Since under-resourced language communities often lack the time, expertise, and resources to produce these systems, the achievement of a WER below 50% with all accessible and beginner-friendly materials is meaningful. This research is an example that novice ASR users can find these resources available to them, and contribute to URL systems for their communities. Furthermore, the 46.88% WER is a beneficial starting point for further changes to the system. The purpose of this study was not to create a state-of-the-art system, but to provide an open-source, accessible model that achieves a less than 50% WER on very small datasets under 5 hours. This study shows the results of the system without added data augmentation or several hyperparameter changes. This leaves room for growth in the future, and a base from which to compare with future research. With the WER already below 50%, there are significant possibilities to improve this with further experimentation.

5.1 Challenges

There were many difficulties encountered during this experiment. As an absolute beginner with automatic speech recognition, the technical application was a massive challenge. However, due to my limited experience with ASR we can conclude that it is possible for beginners to apply this technology at a basic level with some base knowledge. This could allow speakers of under-resourced languages with little to no experience to experiment with open source ASR systems and contribute to language preservation.

There were many issues surrounding the code in this experiment since open source code still needs to be adapted to the individual's needs. Using Google Colab for the experiment was the first choice, but there were some concerns. Mainly, the free version of Google Colab offers a limited amount of RAM and disk space. Even simple ASR systems take a long time to run, and in longer sessions I ran out of space to train the system. When run efficiently, my model will run to completion with the RAM and disk space allocated with the free version. For less concern about these issues, Google Colab Pro can be used to increase the RAM and disk space while training the system. For students with access to the Peregrine HPC cluster with their university, this code could be adapted for that format. I attempted to use this method to get the model running while encountering errors in Colab. Ultimately the errors were resolved and Colab was used due to its functionality and easier access for other users. Using a Jupyter Notebook format with a student Peregrine account will also suffice when the code is adapted properly.

One of the main issues encountered in the code was uploading the audio. The rest of the dataset with transcriptions was usable in the code with no issues, but the audio was not attached. Ultimately, the issue was with torchaudio. It is important to note that this model currently requires an earlier version of torchaudio such as 0.9.0 which is used here and written in the code. Using the current 0.12.0 version will result in errors when calling the audio.

Another main issue was evaluation. The evaluation methods included with Patrick Von Platen's code worked just fine for his code, but did not function with my data. After experimenting with several different methods, I discovered that Manan Dey's method of evaluation with his Irish system worked once I made the necessary changes to suit my code. The resulting WER is 8.43% higher than the best performing Irish system in HuggingFace, but it comes below the 50% mark presented in the hypothesis. The training hyperparameters of the system are different to mine, and in future experiments it is possible that changing these could improve my model WER.

5.2 Limitations and Recommendations

While the models and data chosen for this experiment are suited for research, it is important to understand that they have limitations. As previously mentioned, the hyperparameters for the model may need to be optimized for the best possible results. This requires more time and research to experiment with the best possible interactions. In addition, the model could be run for a longer period of time with the hyperparameters. When running the system, the first few training sessions had a far higher WER of 95% at step 400. When I ran it for the final time, the WER at step 400 was 64%. It is possible that training longer could continue to lower this WER, and would make an interesting experiment in the future. To add more variety to the model, we could also add a character error rate (CER) to the evaluation process. While WER is an important metric for this model, adding CER to the evaluation could help give extra information to future experiments.

The WER prediction for this experiment was based on the results of previous models by Abid Ali Awan and Manan Dey. The attempt to achieve a WER below 50% was functional for this experiment, but is not a universal metric for measuring the WER of all multilingual under-resourced ASR systems. In future experimentation we can make comparisons between this result and new results from added variables, but the metric will change with each model.

The dataset also has many limitations that should be considered. Datasets used in automatic speech recognition should have a wide range of voices in order to offer a proper variety for the system to recognize. All ages should be represented including elderly voices and children. It can be difficult for ASR systems to recognize these voices if there is no data for it to learn from, since their voices have unique characteristics such as high pitch or croakiness. The dataset should also include those with speech impediments as well as a relatively equal male to female ratio of audio. A variety of unique and varied voices will make the best dataset so the system can train to recognize different voices. This is a big limitation for under-resourced language datasets (small datasets) because they often lack these attributes. The Common Voice Irish dataset has a much larger percentage of male voices, most of whom are under 40 years of age, so it has quite a limited pool of voices to work with. While a crowd-sourced dataset is a great opportunity to gather data, it does not always allow for control of the demographics. Combining this issue with the small size of the dataset, it is likely that the models created with this set will have trouble recognizing people outside of the young male demographic. The goal here is that the multilingual data used to pre-train the system will be able to compensate for this demographic issue, and offer a wider range of voices to the system. In order to test for this issue, future research could evaluate the system functionality on a range of Irish speakers.

The crowd-sourced format also offers other limitations since the naturalness of the speech might not be present in the dataset. The volunteers are reading sentences out loud, so it might not necessarily sound like casual spoken conversation. It is possible that the lack of natural or organic speech in the dataset might affect how the system reacts to natural speech in recognition. Patterns and intonation of this speech often differ from read speech, and it would be far better for a system to have both options as input in the dataset so it has more to train with. Another issue with the crowd-sourced format is noise. This is often an issue with ASR systems and dealing with noise is a big part of experimentation in modern ASR systems. As previously discussed in the challenges section, the system needs to be prepared to deal with noise during recognition. If it cannot determine the voice through the noise, the system will not function. Adding more noise to the dataset can be rectified by expanding the data, but also using data augmentation to add noise to the existing audio.

To solve dataset issues, the best option is to use a dataset with all of these features included. However, creating a large and varied dataset requires time, money, and expertise which most language communities do not possess. The Common Voice Irish dataset was the only open source one I could find, although there may be payable versions available. Regardless, it is difficult to find any Irish datasets available other than Common Voice.

Another option for some of the issues is data augmentation, which can expand the dataset by raising pitch, time, or adding noise to mimic more varied audio.

Using data augmentation to further this study would be highly beneficial to improving WER along with experimentation with the hyperparameters. If we can expand the dataset further to mimic a larger version, our system would have much more to work with than just 4 hours. Because of this experiment, we now know that a WER under 50% can be achieved with just the 4 hours of data. Now, we can expand on this research and try implementing these methods to refine the system for the lowest possible WER. However, like the previous study, this requires a significant amount of time, effort, and expertise. These all served as limitations during the research process due to the limited time of the program and the thesis window. In future study, more time and expertise would be required to implement these solutions.

Future research could also benefit from comparing the Wav2Vec 2.0 XLSR-53 to other Wav2Vec 2.0 models. Similar to the research conducted by Conneau et al. in 2021, it would be an asset to the model to test it against the XLSR-English, XLSR-monolingual, and XLSR-10 models to compare the results. While the XLSR-53 is a very large and effective model, it is possible we could get a better result from one of the other models. This way, we can compare monolingual, bilingual, and multilingual methods with this particular dataset while still using the same Wav2Vec 2.0 base model.

6 Conclusion

Automatic speech recognition offers a wide range of applications to its users in medicine, security, accessibility, and general daily use. Unfortunately, speakers of most languages in the world are not able to benefit from these innovations. More than 90% of the world's languages are in danger of extinction by the end of the century, and ASR technology can help to combat this problem (Prud'hommeaux & Jimerson, 2018). By offering opportunities to use their language in daily life with technology, and improving quality of life by increasing accessibility and services for everyone, ASR can help sustain languages through continued use. Despite the unique challenges of low recorded data in under-resourced languages, researchers have studied multilingual methods to combat this issue. Facebook AI has developed the Wav2Vec 2.0 XLSR-53 model in order to train in a self-supervised manner with large amounts of multilingual data, then fine-tune with smaller datasets. Using the Irish Common Voice dataset, I predicted that we could achieve a WER of less than 50% using this model with less than 5 hours of data. After fine-tuning the model, my version achieved a WER of 46.88%. This result concludes that it is possible to achieve a WER below 50% with completely open-source materials on a very small amount of data.

This experiment could benefit from a variety of extensions in the future. As a basic extension, changing the hyperparameters during training could test the viability of the current model. If the WER shifts based on these changes, we can optimize the model for future use. Furthermore, we can expand the research by comparing it with the XLSR-English, XLSR-monolingual, and XLSR-10 models. We can fine-tune the models with the

Common Voice Irish dataset and compare results to determine the best WER in monolingual, bilingual, and different sized multilingual models with the same base of Wav2Vec 2.0 XLSR. To further this experiment, data augmentation can also be implemented to determine the effectiveness in these particular models. Doing a test both with and without data augmentation with the 4 models, we can find the best model for the Irish data.

In conclusion, this experiment shows that novice ASR users can achieve results on under-resourced language systems using multilingual methods, and completely open-source model and dataset options. Since language is such a significant part of culture and identity, these results are very meaningful. Using accessible materials, we can now create ASR systems for under-resourced language communities so that more people may benefit from the applications of automatic speech recognition systems. The possibilities are improving through more research, and will hopefully continue to improve further in the future with more focus on these languages as an important part of people's lives.

References

About the Language (Ireland). (n.d.). An Coimisinéir Teanga. Retrieved July 17, 2022, from https://www.coimisineir.ie/faoin-teanga?lang=EN

Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., & Hill, N. (2021). User-friendly Automatic Transcription of Low-resource Languages: Plugging ESPnet into Elpis. *COMPUTEL*. https://doi.org/10.33011/COMPUTEL.V1I.969

Arora, S., & Singh, R. (2012). Automatic Speech Recognition: A Review. *International Journal of Computer Applications*, *60*, 34–44. https://doi.org/10.5120/9722-4190

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Benkerzaz, S. (2019). A Study on Automatic Speech Recognition. *Journal of Information Technology Review Volume 10 Number 3 August 2019*, *10*(3). https://www.academia.edu/43773934/A_Study_on_Automatic_Speech_Recognition

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Carrier, M. (2017). Automated Speech Recognition in language learning: Potential models, benefits and impact. *Training Language and Culture*, *1*(1), 46–61. https://doi.org/10.29366/2017tlc.1.1.3

Chitu, A., & Rothkrantz, L. (2012). Building a Data Corpus for Audio-Visual Speech Recognition. *Man-Machine Interaction Group Delft University of Technology*.

Common license types for datasets. (n.d.). Retrieved July 6, 2022, from https://docs.data.world/en/59261-59714-2--Common-license-types-for-datasets.html

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016). Selection Criteria for Low Resource Language Programs. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4543–4549. https://aclanthology.org/L16-1720

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *Interspeech 2021*, 2426–2430. https://doi.org/10.21437/Interspeech.2021-329

Diwan, A., Vaideeswaran, R., Shah, S., Singh, A., Raghavan, S., Khare, S., Unni, V., Vyas, S., Rajpuria, A., Yarra, C., Mittal, A., Ghosh, P. K., Jyothi, P., Bali, K., Seshadri, V., Sitaram, S., Bharadwaj, S., Nanavati, J., Nanavati, R., … Abraham, B. (2021). Multilingual and code-switching ASR challenges for low resource Indian languages. *Interspeech 2021*, 2446–2450. https://doi.org/10.21437/Interspeech.2021-1339

Dodiya, T., & Jain, S. (2016). Speech Recognition System for Medical Domain. *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1)*, *7*, 5.

Gaelic vs. Irish: What's the Difference? (n.d.). Retrieved July 17, 2022, from https://www.unitedlanguagegroup.com/blog/gaelic-irish-differences

Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications*, *10*(3), 16–24. https://doi.org/10.5120/1462-1976

Huang, J.-T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7304–7308. https://doi.org/10.1109/ICASSP.2013.6639081

Irish is now at the same level as the other official EU languages. (2022, January 3). [Text]. European Commission - European Commission. Retrieved July 19, 2022, from https://ec.europa.eu/info/news/irish-now-same-level-other-official-eu-languages-2022-jan-03_en

Irish Language and the Gaeltacht—CSO - Central Statistics Office. (2016). CSO. Retrieved July 17, 2022, from https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/ilg/

Miao, Y., & Metze, F. (2013). Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. *Interspeech 2013*, 2237–2241. https://doi.org/10.21437/Interspeech.2013-526

Petkar, H. (2016). A Review of Challenges in Automatic Speech Recognition. *International Journal of Computer Applications*, *151*(3), 23–26. https://doi.org/10.5120/ijca2016911706

Prud'hommeaux, E., & Jimerson, R. (2018, May). ASR for Documenting Acutely Under-Resourced Indigenous Languages. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1657

Raut, P. C., & Deoghare, S. U. (2016). Automatic Speech Recognition and its Applications. *International Research Journal of Engineering and Technology (IRJET)*, *03*(05), 4.

S, K., & Chandra, E. (2016). A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*, 393–404. https://doi.org/10.14257/ijsip.2016.9.4.34

Sailor, H., Patil, A., & Patil, H. (2018). Advances in Low Resource ASR: A Deep Learning Perspective. *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 15–19. https://doi.org/10.21437/SLTU.2018-4

Saini, P., & Kaur, P. (2013). Automatic Speech Recognition: A Review. *International Journal of Engineering Trends and Technology*, 5.

Singh, N., Khan, Prof. R., & Pandey, R. S. (2012). Applications of Speaker Recognition. *Procedia Engineering*, *38*, 3122–3126. https://doi.org/10.1016/j.proeng.2012.06.363

The Irish Language in Ireland. (2021). De Gruyter. Retrieved July 26, 2022, from https://www.degruyter.com/database/LME/entry/lme.15385116/html

Vaessen, N., & van Leeuwen, D. A. (2022). Fine-tuning wav2vec2 for speaker recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7967–7971. https://doi.org/10.1109/ICASSP43922.2022.9746952

Wang, Z., K, K., Mayhew, S., & Roth, D. (2020). *Extending Multilingual BERT to Low-Resource Languages* (arXiv:2004.13640). arXiv. http://arxiv.org/abs/2004.13640

Xu, Q., Baevski, A., & Auli, M. (2021). *Simple and Effective Zero-shot Cross-lingual Phoneme Recognition* (arXiv:2109.11680). arXiv. http://arxiv.org/abs/2109.11680

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). *Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages* (arXiv:2012.12121). arXiv. https://doi.org/10.48550/arXiv.2012.12121

Yılmaz, E., Biswas, A., van der Westhuizen, E., de Wet, F., & Niesler, T. (2018). *Building a Unified Code-Switching ASR System for South African Languages* (arXiv:1807.10949). arXiv. https://doi.org/10.48550/arXiv.1807.10949

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition*. Springer London. https://doi.org/10.1007/978-1-4471-5779-3