DOMAIN SPECIFIC VOICE INTENT CLASSIFICATION WITH BLSTM

Hellarawa Mudiyanselage Madushika Jayani Hellarawa

(199325F)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineer

University of Moratuwa

Sri Lanka

October 2022

DOMAIN SPECIFIC VOICE INTENT CLASSIFICATION WITH BLSTM

Hellarawa Mudiyanselage Madushika Jayani Hellarawa

(199325F)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

October 2022

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: UOM Verified Signature Date: 16.10.2022
Name: H.M.M. Jayani Hellarawa

The above candidate has carried out research for the Master's thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Signature: UOM Verified Signature Date: 16.10.2022
Name:
Dr.T.Uthayasanker

ACKNOWLEDGEMENTS

I would like to express my appreciation and the gratitude for those of who supported me throughout process as this research would not be possible without those. My sincere gratitude goes to my supervisor Dr. Uthayasanker Thayasivam for his continuous support and guidance throughout. And also, for all the academic and non-academic staff, Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa Sri Lanka for their support.

My sincere thanks should go to all my friends and colleagues who helped me to make this a success.

Finally, I must express my profound gratitude to my parents, friends and colleagues who helped me to make this a success by providing me with unfailing support and continuous encouragement. This accomplishment would not have been possible without them. Thank you.

Author
Jayani Hellarawa

ABSTRACT

With the current global pandemic all countries around the globe are facing difficulties managing their healthcare services in a way that ensures the high availability of critical services while maintaining the safety of both the patient and the staff. According to Gartner's top 10 strategic technology trends 2021 [1], it says *"Rather than building a technology stack and then exploring the potential applications, organizations must consider the business and human context first."* where it highlights the need for human centric development while stating that it is the IT leaders that decides what combination of the trends to involve in driving the most innovation and strategy.

A decade ago, simply having a website was enough to impress prospective customers and help them find their way to a service or information need and to establish a brand loyalty or identity. The growth of the technology is demanding more innovative strategies to adopted to every small to large industries that are at any stage of maturity of their roadmap to success. The increasing demands of the clients and the ability to keep a loyal customer base has highlighted the need of having a more natural way of handling a customer's inquiry gives a competitive advantage for any business.

The disappointment due to a customer getting added to a call waiting queues to reach a particular service is very critical and can even cause a loss of business opportunity. Understanding call intents can help a service provider to adapt the business engagement with the outside in a way that customers are positively satisfied which could in return increases the sales revenue. Not only that, but indirectly enables the ability for business to allocation agents or help-desk staff optimally thus avoid understaffing and overstaffing situation, which are indirect costs for any revenue-based figure.

Automation is where the technology is used to automate tasks that once required humans. Here, the menu-based call center automations can be taken as a replacement to the legacy call center agent where the human tasks were replaced by automation. The concept of hyperautomation is where the businesses are rapidly adopting it's revenue-based processes and IT process for automation. And the current state-of-the-art deals with lot of advanced technologies like Machine Learning (ML) and Artificial Intelligence (AI). Where AI and ML are used for extending the capabilities of automations.

The building of a speech recognition (ASR) systems for an open domain has been research for a lone time. Where the most of those are accomplished by collecting the voice corpus, convert them into text and performing a text classification on top of the converted text. However, this comes with lot of limitations, thus is not identified as the most feasible way of deriving intents of a speech query for a specific domain [2]. Therefore, in this research, that is focused on domain specific voice intent classification will be aligned with the healthcare domain for the English language based on a neural network with Bidirectional Long Short-term Memory (BLSTM).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASR	- Automatic Speech Recognition
BLSTM	- Bidirectional Long Short-term Memory
LSTM	- Long Short-term Memory
ROI	- Return on Investment
SVM	- Support Vector Machine
CTC	- Connectionist Temporal Classification
DNN	- Deep Neural Network
DCT	- Discrete Cosine Transform
MFCC	- Mel-Frequency Cepstral Coefficient
RASTA	- Relative Spectral Amplitude Filtering
PLP	- Perceptual Linear Predictive
LPC	- Linear Predictive Coding
WFST	- Weighted Finite State Transducer
AM	- Acoustic Model
LM	- Language Model
SVM	- Support Vector Machine
FFT	- Fast-Fourier-Transformation
TL	- Transfer Learning
ML	- Machine Learning
AI	- Artificial Intelligence
LRL	- Low Resource Languages
HIPAA	- Health Insurance Portability and Accountability Act

1. INTRODUCTION

The concept of speech recognition started back in the early 1950's and it had the focus of voice activation and recognition. Speech, has being the primary way of communication since the early stages of civilization, a great deal of researches have been captivated by both the technological curiosity towards mechanical understanding of human speech, as well as the desire to automate simple human-machine interactions over the past decades. This study area is recognized as the automatic speech recognition (ASR) where the natural language understand and the processing is carried out by machines.

In 1930s, Homer Dudley of Bell Laboratories has proposed a record-breaking system model for speech analysis and synthesis [3], [4]. The evaluation started from the need to develop methodologies that is capable of responding to a small sound to a more sophisticated methods that responds to spoken natural language. With the rise of the statistical modeling of speech around the 1980s, the applications that has a human-machine interaction, like supporting to travel based inquiries, automatic call processing in mobile network and etc. had gained the advantages.

The ability to understanding the intents of call, can help a business give positive experiences and also to excel in the customer support. In return, the positive experiences from the customer can help in boosting sales and revenue while supporting the allocation of staff/agents optimally. This will indirectly decrease the problem raised with situations like understaffed or overstaffed.

But like any research area it has a lot of challenges in collecting and analyzing such a massive dataset of phone calls in order to support caller intent classifications. First, the need to collect good quality speech data for the domain, with confidentiality in mind and then the dataset is too large to analyze manually for preprocessing. Second, different customers often use different words have different approaches in expressing the same intent in phone calls thus finding the best matching machine learning algorithm is really hard.

Research carried out in related to hospital call centers have proven its money making capabilities, despite of been focused on patient the need to attract more and more potential patients while uplifting the quality of the service and customer satisfaction. A

study published in 2003 that is engaged in healthcare [5], has gathered around 2 million phone calls from around 8k customers at 11 different healthcare all centers.

The analysis was focused on revenue gained from the caller followed by the month of their phone interaction with the hospital. This is known as the *downstream revenue,* where the revenue is earned after the immediate sales but by further servicing and supporting. The final output of the study has concluded that callers ROI is of at least $3 in downstream revenue for every $1 that the healthcare center has spend on their call center.

The studies in healthcare field are not only limited to call centers but has expanded over many areas like understanding User Query Intent from Online Health Communities [6], mechanical sensors on the prosthesis and/or electromyography measurements from the residual limb to classify intent of advanced lower limb prostheses [7], where the features extracted from the depth images were classified using SVM to identify user-independent intents and even some studies has already take into account of the labor cost of the the operator, the cost caused by the waiting customers as well as the cost for the lost calls [8]. There is no doubt that every person does not like to wait in a calling queue to get operator assistance or waiting for the second or even third round for the option menu list to be repeated so that you can select the best call destination.

1.1. Research Problem

The area of transcribing sequences of data into related sequences of discrete labels is know as sequence labeling in machine learning. In speech recognition, speech signals that are the inputs of the system is produced by the continuous motion of the vocal tract, while the output that is a sequence of words /labels are mutually created by adhering to the laws of syntax and grammar. Therefore the alignment of inputs and the labels are required and critical in a classification problem.

Recurrent Neural Networks (RNNs) are inspired by the cyclical connectivity of neurons in the brain and use iterative function loops to store information. For a neural network based approach for speech intent classification, it has some unique properties that make those an ideal choice for sequence labeling. This is because the RNNs are flexible towards learning context information. And they can recognize patterns that are sequential patterns [9]. But unfortunately, RNNs also comes with the disadvantage where it is very difficult to make units to store information for long periods of time.

The solution for this was introduced with a new architectural redesign, where the special *memory cell* units called Long Short-term Memory [5]. It has the ability of both processing single data points as well as an entire sequence of data such as video or audio. Even with that, only having one direction access to contextual information (mostly the past for temporal sequences) best suits for time-series predictions. But, it can get the maximum use of information when the context on both sides of the labels are exploited. And some literature presents that the Bidirectional RNNs [10] (BRNNs), overcomes the inability to access previous data by training it simultaneously in positive and negative time directions.

In those networks, they the data is scanned forwards with a forward layer and backwards with another recurrent layers. And therefore, it removes the asymmetry that was there in between input directions and provides access to all surroundings contexts. The structure of the BLSTM combines both benefits from the long-range memory as well the from the two directional processing. The application of BLSTM in regression and classification experiments on artificial data [11] has given better results than other approaches.

There are lot of domains like healthcare that suffers from the low resource availability due to the lack of previous literature, proper corpus of even ethical and legal legislations that challenges the collections of such data in real world. Thus, they require data to be collected synthetically and algorithms and machine learning techniques that can be used to achieve satisfactory amount of accuracy with such a minimum sized corpus.

This paper proposes a domain specific voice intent classification system that directly identifies the intent of a speech utterance, without convert the audio into an intermediate form of text or phonetic representations. Therefore, the research will focus on end-to-end ASR with Connectionist Temporal Classification (CTC) objective function where the model is built with BLTM. And use techniques like transfer learning and data augmentation towards addressing the domains with low resources.

1.2. Motivation

The situations like the Easter Sunday attack raised the problem with supporting the optimum allocation of employees/agents or customer service representatives that avoids understaffing and overstaffing situations for customer support in the hospitality industry. As a professional engaged in the hospitality industry, it was continuously highlighted the need for a better call center support specially when you are providing 24 * 7 service or when you have clients across different time zones. But overstaffing to support for high demanding situations or understaffing to reduce the cost is not the best option.

During this pandemic situation, the healthcare industry became one of the most essential services throughout the world. Considering the previous experience, there is no doubt that healthcare centers are struggling with call center operations now unlike more than ever. Accessing the healthcare services during the pandemic has become critical in such a way that puts both the patient and the service providers at high risk.

Call centers has become the most convenient way of communication to avoid all hazards but most of service providers were not ready for such a critical situation with high availability or they had to prioritize the staff allocation which is very crucial. Most of the prior research in healthcare domain are focused on text based intent classifications and none of the research has published publicly available dataset for healthcare call-centers. The rules and regulations like *HIPAA* Compliance might have a huge impact on such low availability of resources.

Therefore, building a publicly available dataset will give a boost for other researchers who are interested in the same and this initiation has the ability to provide with a better application of technology towards humanity. That motivation has led this study to select the healthcare sector as the domain for intent classification.

2. LITERATURE REVIEW

With the higher attraction towards speech recognition over last decades, there are number of researches in the literature, that studied the areas on building ASR systems for different domains with different approaches. As by the human's nature, they are always fascinated by the new inventions/ technologies that can understand them. And the fact of machines having the basic understandings towards human was really highlighted and has gained lot of popularity. And thus can be used to justify why the world's largest technological companies are very focused on bringing those services to the public.

The literature review was carried out in many different angles to grasp previous workings as over a period of decades, researcher have tried myriad ways to dissect language: by sounds, by structure and with statistics as well. Therefore, the areas related to speech recognition, intent classification, and customer call analysis and also healthcare/medical related studies were observed.

2.1. Speech Recognition

There are several facts that makes speech recognition not easy, such as; by the nature speech is usually continuous and therefore it is hard to define the word boundaries which are not easy to clearly define. The other common challenges that causes the errors in ASR is the lack of the ability to clearly identifying the space between words as speech been continuous, people around the world pronounce the same word with different dialects. Therefore it really easy to misclassify different spoken words that sound similar in expressive and vocabulary rich languages.

The earliest days in speech recognition was focused mainly on the basis of a speech system, creation of vowel sounds which might also learn to interpret phonemes from nearby interlocutors and the inventions like dictation machines by Thomas Edison in the late 19^{th} century, were capable of recording speech and thus grew in popularity among people that deals with taking lots of notes [12].

During the time period from 1950's to the 1960's is considered as the era of 'baby talk' in speech recognition where only numbers and digits; which are less complex than human language, could be comprehended with the invention of Audrey. Audrey (Automatic Digit Recognizer), a machine created by Bell Labs, with the ability to understand the digits 0-9 with an accuracy rate of 90%. But this has a limitation where the accuracy varies on the fact whether the voice belongs to its inventor or not.

This dependency is due to the fact that each individual's voice is different from each other; which is one of the utmost challenges in the study area of speech recognition. In 1962 when the IBM Shoebox was revealed with the ability to understand 16 English spoken words, including 10 digits; 0 - 9 with basic arithmetic commands such as "+", "-" and "total", it returned the final answer for the simple arithmetic problems. IBM with the leadership of Fred Jelinek and his team have researched the areas like Hidden Markov Models, statistical language modeling and etc. for speech recognition system around 1970s [13].

Thereafter, thanks to the improvements of computer processing power and with other related technologies, it grew from continuous speech recognizers, Google voice search, Apple Siri to the voice-controlled speakers like Amazon Alexa and are now embedded into many electronic devices in a user-friendly manner.

2.2. Speech Recognition Techniques

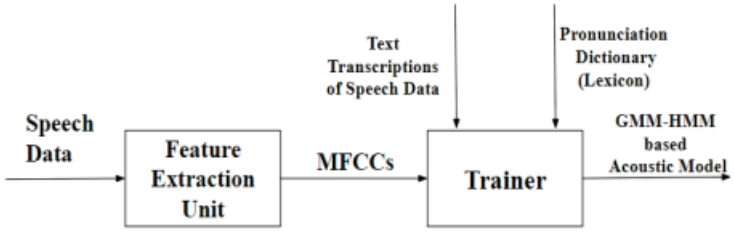

Figure 1: Acoustic model training process

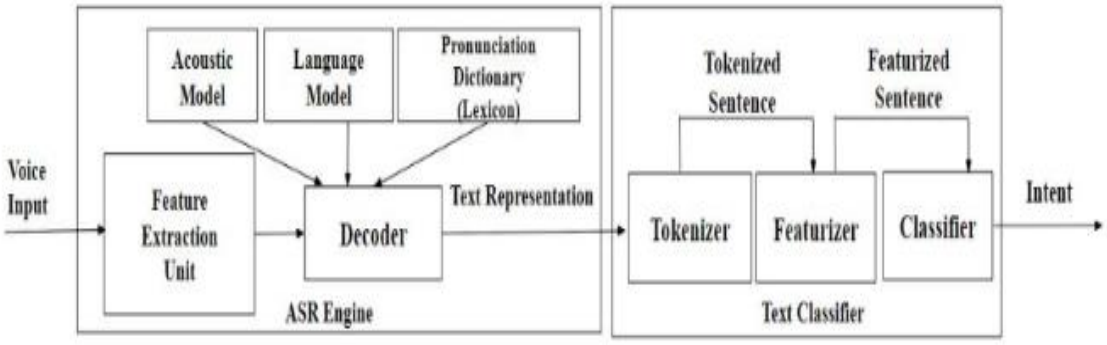

Figure 2: Architecture of a ASR with HMM and GMM

Early stages of the speech recognition techniques were mostly based on text, where the spoken sentences or words are converted into intermediate text and then the classification techniques were applied considering the problem as a regular text classification problem. As shown in Figure 1 and Figure 2 is an example of how ASR used be, where it to consume different resources to get better performing solution. Here,

- Acoustic model: feed for this model is the audio clips and the related text transcriptions
- Language model: Captures the grammar of the language
- Dictionary model: created manually for all words of the selected vocabulary by including the phoneme representation

  There are many speech recognition techniques that are used by many researchers that allow them to adapt to different approaches in speech identification. Hidden markov model (HMM) and DTW (Dynamic Time Wrapping), and also neural network-based approaches are very popular.

a. Hidden Markov Model: a statistical models, the output is a sequence of symbols / quantities.
b. Dynamic Time Wrapping: an algorithm that is capable of measuring the similarity between two sequences. Where those are in different time or speed.
c. Neural Networks: Neural networks have been used in many aspects of speech recognition for a while. Speech recognition with deep neural networks have shown that the combination of deep, BLSTM RNNs with end-to-end training has the state-of-art performance on TIMIT dataset.

2.2.1. Hidden Markov Model

HMM is a statistical model. It is based on the assumption that the specific system is a Markov process and the Forward–backward algorithm used in HMM models was first described by Ruslan L. Stratonovich in 1960. Even though HMM is initially introduced and studied in the late 1960's to 1070's, this statistical modeling method has become increasingly popular in these days as well. And in some studies, related to the application of HMM for speech recognition [14] has mentioned that, when applied properly it can perform well.

The researchers were hold back for close to 80 years by the fact that they were lack of methods for optimizing the parameters of the Markov model to use for matching observed signal patterns when it comes to speech recognition and processing [15]. And after this was addressed in late 1960's, HMM was immediately applied to speech recognition and even in today, it has shown a good accuracy with compared to others in intent classifications for specific domains as well [16].

2.2.2. Neural Networks (NN)

NNs have been used in many aspects of ASR over the last few years. The advances in computer hardware and the availability of GPUs become more affordable while the machine learning algorithms have also advanced, have attracted many research to come up with more efficient methods for training NNs that benefits from more number of layers. This leaded to invent new methods of performing the same machine learning task that outperform the previous.

A shared research of four groups; Microsoft Research (MSR), the University of Toronto, IBM Research and Google [17], has used a training procedure that has two-stages that is used for fitting the deep NNs for acoustic modeling and the research has shown that compared to highly tuned GMM-HMM systems, DNNs can outperformed.

2.3. Bidirectional Long Short-term Memory (BLSTM)

Bidirectional LSTM provides access to long range context in both input directions. This method of model training combines both forward and the backward predictions of the hidden layers. First introduced by Schuster and Paliwal, this architecture gives the network the ability to get the information from both future and past states, which is known as forward and backward passes simultaneously Figure *3*.

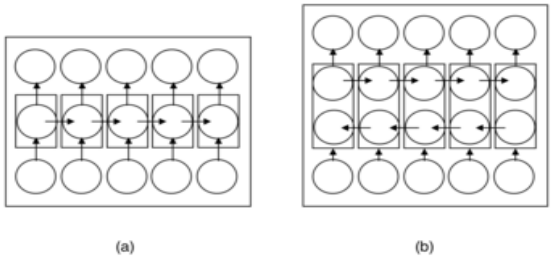

Figure 3: Structure of RNN and BNN

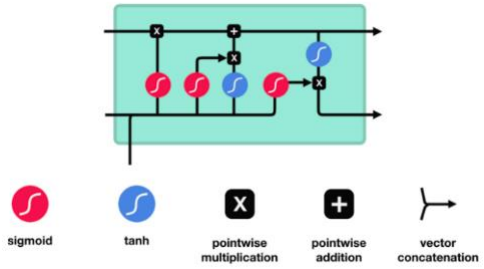

Figure 4*:* LSTM cells and it's operations

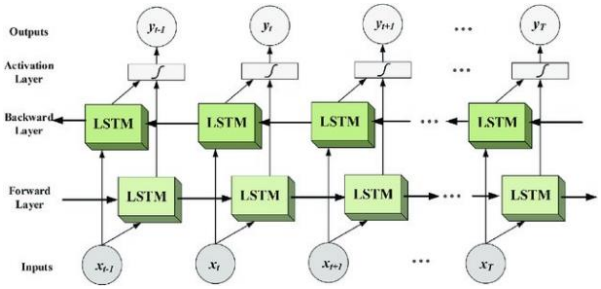

Figure 5*:* BLSTM architecture

The important thing with LSTM as shown in Figure *4*, is that it has an architecture that helps them learn longer-term dependencies. Therefore, BLSTM can get the maximum use of audio signal as they can be fed to the network both in forward and backward passes as shown in Figure *5*.

2.4. Speech Feature extraction

A collection of algorithms that are focused on performing multiple different tasks with different inputs and outputs altogether creates a speech recognition system. Some of those component's purpose may the understanding of the patterns of the speech, statistical analysis for patterns, signal processing and many others [18]. Even though the degree of the effect of those areas differs depending on the type of the recognizer, the part that converts the speech waveform to some type of parametric representation known as the signal-processing front-end; is a common element.

Even though the speech signals are sequential, for human auditory system it is identified in a nonlinear fashion. The feature extraction is the process of compacting the paramedic representation of a speech signal. This extraction process captures the essential elements of a signal while reducing the irrelevant data [2]. And the feature extraction block used in speech recognition needs to be able to reduce the complexity of the problem as soon as possible without dragging the complexities alone the way.

RASTA, PLP, LPC, FFT are most popular feature extraction techniques. And MFCC is known as the state-of-the-art technique for feature extraction.

2.4.1. Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are computed from the mel-scale by using a method called frequency binning that is based on the human ear's frequency resolution. These MFCC features are used to parameterize speech data as the mel-scale is a unit of measurement of perceived frequency (pitch) of a tone [19].

In MFCC, the given audio segment is divided into sliding windows of given size (here in Figure *6*: MFCC windowing it is 25ms) wide to extract audio features. In *Figure 6: MFCC windowing* [20]; shows the widow that is used to separate audio signal to extract features. The size of this window is decided so that it can capture considerable amount of information of the audio but still needs to keep the stationary features inside the frame. In order to capture the proper context by using dynamics among frames, a slide window; which is known as a hop-length (here is it of size 10ms) is used. The F0 feature is known for the pitch and as it has little role in the context of what a speaker said.

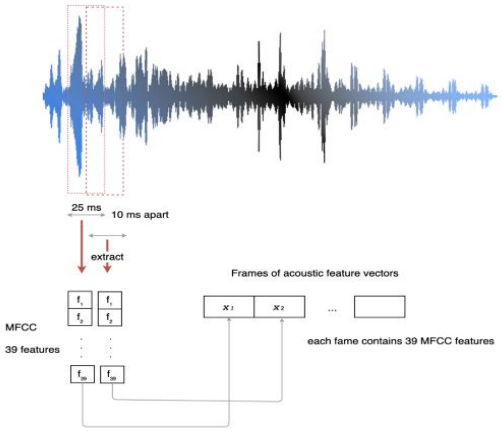

Figure 6: MFCC windowing

The key objectives of the MFCC includes the removal of vocal fold excitation(F0) which is returned by F0 and make the extracted features independent. And at the same time, capturing of dynamics phonemes of the speech signals. And the Figure *7*, shown the block diagram of the MFCC.

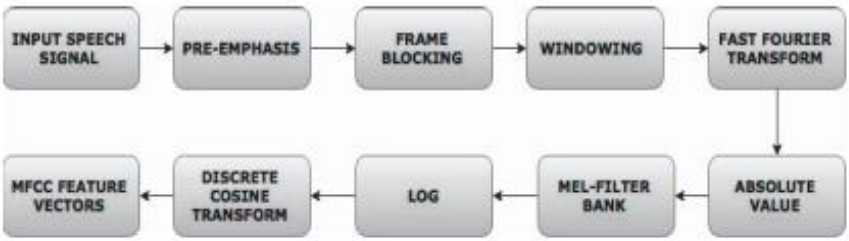

Figure 7: Block diagram of MFCC

The literature that studied on feature extraction techniques [21], has shown that MFCC has better recognition rate compared to other feature extraction techniques like LPC and RASTA.

2.4.2. Connectionist Temporal Classification

For the labeling of unsegmented data, the use of graphical models such as Conditional Random Fields (CRF), Hidden Markov Models (HMM) are prominent, they comes with few limitations as well. The need for a state model for HMMs which requires a huge amount of task / domain specific knowledge to design, or the limitations of CRF like that it requires explicit dependency has highlighted the need for a new way of overcoming those.

On the other hand, other than the input and the output data, Recurrent neural networks require no prior knowledge [9] and also their internal architecture helps in building good solutions for modelling time series data while being robust to temporal and spatial noise. In the research; *Labelling unsegmented sequence data with RNNs* [9], proposes a novel method that can be used for sequence data labeling with RNNs that does not requires above fats and models the sequence within a single network architecture. *Temporal Classification* is where the tasks related to labelling of unsegmented data sequences happens and for the use of RNNs for this purpose it is known as the connectionist temporal classification. In CTC with prefix search decoding on TIMIT dataset, the CTC outperformed a HMM recognizer and an HMM-RNN hybrid [9].

2.5. End-to-end learning

Data processing/learning system that requires multiple steps in processing. In end-to-end learning what happens is that it can take multiple stages in a ML process and replace it with a single NN. In early stages of speech recognition, the goal was to take an audio input X and map it to output Y, which is a transcript of the audio clip. In the traditional way first, the feature extraction using techniques like MFCC is done and then apply ML algorithm to find the intermediate process like find the phonemes (the basic units of the sound) of the audio clip and then, they are put together to form words and words are put together to make the translation of the audio clip.

Unlike those pipelines that has lot of stages to complete, in end-to-end deep learning the network is arranged and trained to accept inputs (audio in this case) and transfer to the relevant transcript directly without intermediate stages. In recent years there has been increasing interest in moving from traditional approach to end-to-end learning as it has less steps and gives better performance.

2.6. Transfer learning

In transfer learning, we reuse the information learned by a previous model and apply that learnt knowledge in a new model. In the current state-of-the-art ASR systems, the methodologies are very data intense thus leads requires a good amount of data to learn from. This makes it very hard to get the advantage of such a methodology without a good amount of data which is quite common for low-resource languages or some uncommon/unique domains. Therefore, addressing such limitations with already available resources has become has captivated the researchers and the outcome of such an attempt is the knowledge transfer or transfer learning between a source domain and a target domain.

Previously the issue with low-resource languages is that many of the existing machine learning methods assumes that the training and the test data are drawn from the same feature space and the same distribution, which was very challenging, impossible or comes with high cost to achieve. And the failure to meet those costs the failure in solutions with poor performances. And the models were built from scratch using those collected dataset. But in contrast transfer learning, allows the source and the target domain to be different.

In the survey on transfer learning [22], it goes into details of different flavors of transfer learning depending on the resources available. Figure 8 and Figure 9 shows the difference between transfer learning and the traditional learning processes.

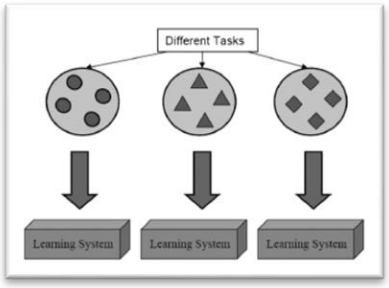

Figure 8 : Learning Process of Traditional ML

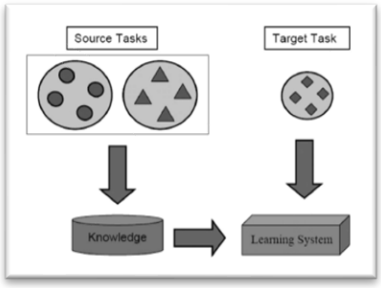

Figure 9 : Learning Process of Transfer Learning

In the context of transfer learning, three main research issues are addressed.

1. *What to transfer:* Answers which part of knowledge can be transferred across domains or tasks. Here the research studies to identify which knowledge is common between the two mains so that it can be used to improve the performance of the target domain or the task.
2. *How to transfer:* Answers how the transfer learning can be done between two domains or tasks.
3. *When to transfer:* Answers in which situations the transfer learning is beneficial. This is to make sure that the purpose of the learning is to transfer positive knowledge to the target domain to improve the performance. In the literature [22], it says that the *brute-force transfer* is not successful when the source and the target domains are not related.

With problems where are two datasets, which is known as cross-dataset recognition; the one dataset is known as the source dataset. This is the dataset used for the training. And the other dataset, which is known as the target dataset is used for the recognition purpose. The characteristics determines what methods best suits for a particular domain or a learning task. And with the help Figure 10; we can identify the most suitable transfer learning method for a given source and the target domain and the task.

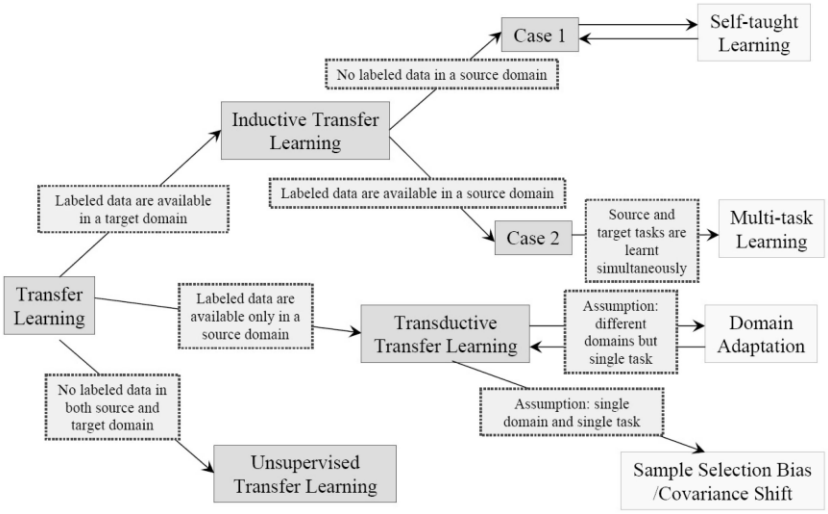

Figure 10: An Overview of Different Settings of Transfer

There are two types of transfer learning approaches practiced, known as

- Fine tuning
- As a feature extractor

Most of the transfer learning approaches found in the literature of ASR, including the benchmark for this research, has focuses the second approach. Where a pre-trained model was used to extract features of the target domain and provided as features to target domain classifier. In contrast, fine tuning takes the weights of the pre-trained model and preserve all or some of the layers and train and tune others [23].

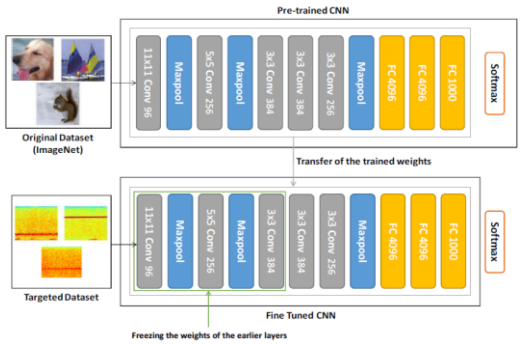

Figure 11: For Fine Tuning

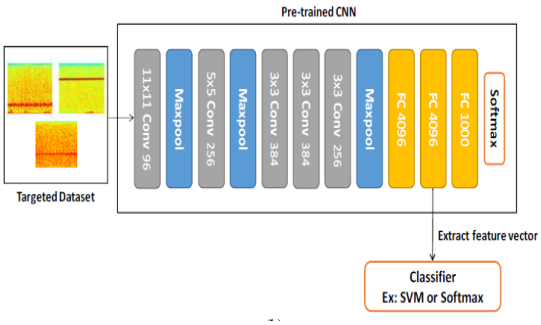

Figure 12: As a Feature Extractor

The research [23] in transfer learning shows the two different approaches as in the Figure 11 and Figure 12. No previous research was found that has used the transfer learning as a fine tuner in the field of ASR intent classification. The main focus in fine tuning is to preserve the featured obtained by the model as those are more generic and applicable to

other tasks of the target domain. And depending on target domain and the classifier performance in the target domain, latter layers are adjusted specifically as those layers provide features specific to the source domain.

2.7. Data augmentation

Throughout the research, one of the main areas of interests is to find better ways of improving the performance of machine learning solutions that suffers from low data / resource availability. Data augmentation has proven its importance in image recognition related ML solutions for a long while. Inspired by the recent success of the speech and vision domains, research like SpecAugment [24] has shown the ability to outperform previous models. This is by avoiding overfitting and improve robustness of the models.

Figure 13, Figure 14, Figure 15 and Figure 16 are some of the popular audio augmentation techniques used by the research [25].

- *Shifting time:* shifts the audio to left/right with a random time period. Shifting to left means it is *fast forwarding* and shifting to right means it is *back forwarding*.

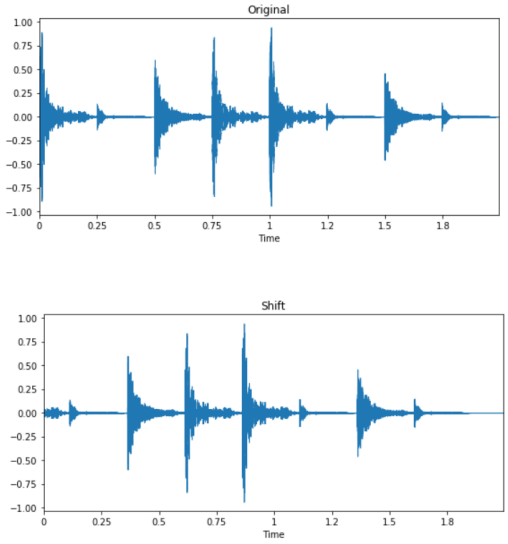

Figure 13: Augmentation with Time Shifting

- *Changing pitch:* Change the pitch randomly.

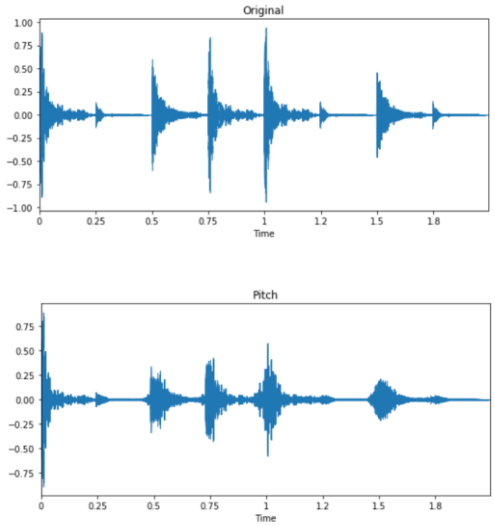

Figure 14: Augmentation with Pitch Changing

- *Changing speed:* same as the changing the pitch, this is performed by stretching the signal in time axis

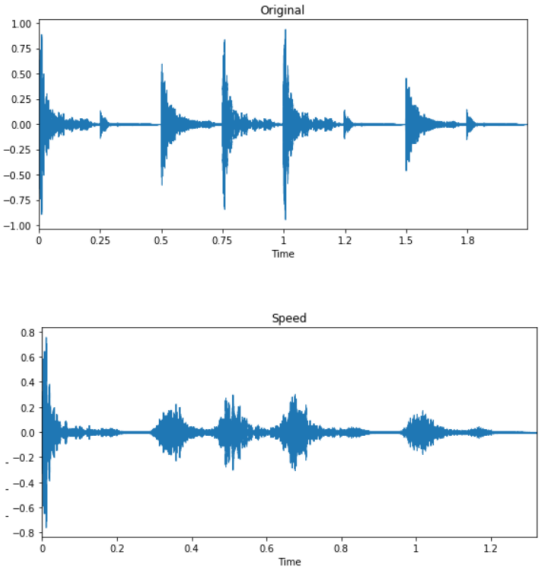

Figure 15: Augmentation by Changing the Speed

- *Noise injection:* injects some random noise to the signal

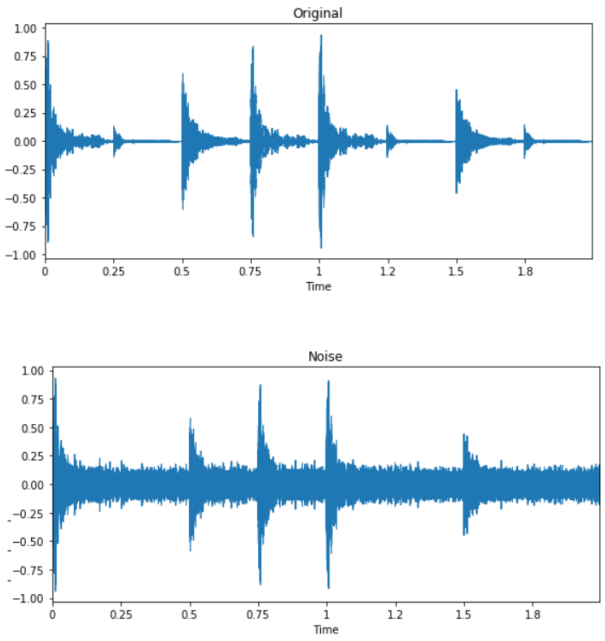

Figure 16: Augmentation by Noise Injection

2.8. Healthcare domain

Medical researchers have done a great deal of work with related to classification and machine learning. Most of the studies related to intent classification are based on textual data as with the development of online health services, the availability of textual data has increased. The internet for sure has changed the way people used to access information related to their day-to-day life and healthcare is one of them. Nowadays lot of websites, blog writers and professionals publish medical related information to patients available online via the web. Same as the marketing strategies followed by business, these non-profit or profitable web resources can add lot of value to any professional body.

Healthcare is complex in nature and the traditional way of information gathering and awareness has since the online health forums have come to the picture. It is a convenient way for patients that are in a hurry to obtain medical information. That large volume of data generated on such forums make it not effective to function its own without a certain level of user friendliness. Therefore, the user intention identification plays a huge role in enabling forums to recommend relevant details to the user by resource filtering.

For user intent understanding in online health forums, [26] tries to derive a taxonomy of intents that helps in capturing user information needs and define a classifier that has learnt on pattern-based features that can identify original thread posts depending on their intents using Support Vector Machine (SVM). And this research has achieved a maximum precision of 75% and has found that the performance can be further improved by integrating data from other forums as to improve the dataset. And same as the above, the CNN-LSTM attention model [10] predicts user intents with an unsupervised clustering method.

Bringing Semantic Structures to User Intent Detection in Online Medical Queries [27], which is based on the concept transition as shown in the Figure 17; the research was focused on capturing the focused picture of medical-related user information search and have a better understanding on their healthcare information access strategies using a graph-based formulation; Figure 18. In the graph representation each node stands for a medical concept and each directed edge stands for a medical concept transition.

○ Medical Text Questions ● Inferred Concept Mentions & Concept Transitions
○ Why do I get dizzy so often? ● *Symptom* → *Cause*
○ My three-year-old child is sick with a temperature of 100 degrees she can't keep anything down including liquids. What kind of medicine should I give my child, and how much? ● Symptom → *Medicine* → *Instruction*
○ Do I have insomnia if I have trouble staying asleep? Any medication is recommended to help me fall asleep easier? ● *Disease* ← *Symptom* → *Medicine*

Figure 17: Query and the Concept transitions.

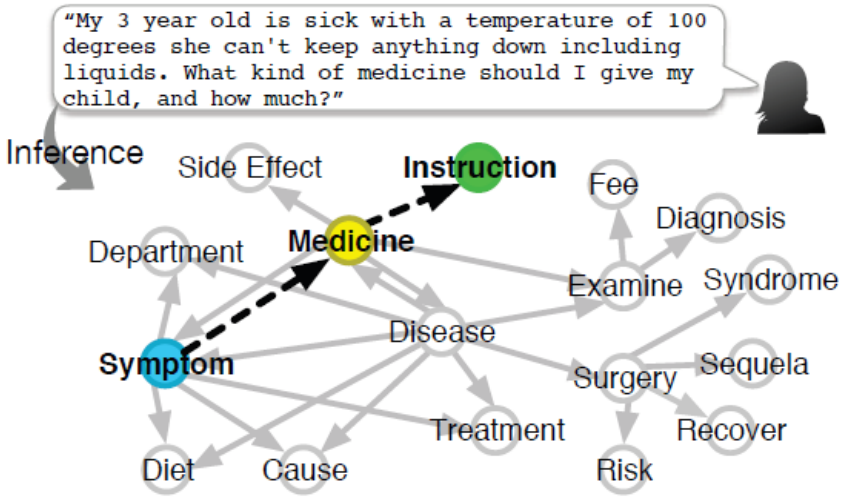

Figure 18: Concept transition inference

There a deep neural model that uses user queries which is based on multi-task learning has used to extract structured semantic transitions where the model extracts word-level medical concept. By examining the transition of the medical concept, the online information provider get the ability to identify what the user wants with the help of user query and provide content rich information that matches information requirement of the user.

Call center customer service gaps can be found in many sectors and with current increased necessity for contactless communication with increased demand, call center automation is a very important feature. No matter whether it is physical or virtual, one of the top reasons for patient dissatisfaction is the longer waiting time. With a situation like Covid-19 pandemic, when the first-call-resolution is low, that puts both the caller and the receiver (here the healthcare provider and the patient) in danger.

With the existing literature, there are lots of research related to intent classification that are based on text with online forum classifications but very few are related to audio-based ASR to the knowledge at the time of this research.

2.9. Data collection methods

Data collection for new research project is very important where the accuracy of such solutions highly depends on the quality of the data. There are many reasons that make the access to such a rich data source is not achievable. This is especially accurate if the research area is uncommon, novel, has security or confidentiality constraints that limits access to outside or even when the resources of such domains are limited. As a solution, recent studies are focused on creating their own set of data using different methods.

In the area of speech recognition, this is one of the challenges faced by most research and therefore comes with the benefits of creating domain specific corpus.

Utilizing a domain-specific data as well as the use of real user speech as data corpus for an ASR solution has the advantage of increasing the performance of ASR systems [28]. By nature, free and rich corpus are really hard to find and when the domains like healthcare cannot get the benefits of most of the feely available corpus directly as there are curtain set of domain specific vocabulary. To comply with the various rules and regulations of patient communications and record-keeping like HIPAA Compliance, leaves such domains to create train and tested on speech data collected under laboratory conditions.

The Health Insurance Portability and Accountability Act (HIPAA) sets the standard for sensitive patient data protection [29] where any company providing treatment, payment and operations in healthcare with protected health information must have necessary security measures and precautions in place to ensure HIPAA compliance. Those datasets can only be collected after getting the approval of the participants to be on site or remote data collections where there’s no requirement of the speaker to present in a particular place. The first approach has high cost and tends to differentiate actual user spoken data as there the environmental conditions are controlled. But on the other hand, latter solution allows a large number of users to participate despite of their geographical location with low cost as there are no specific environment required. Even though it comes with initial low cost, it has hidden cost as such uncontrolled environment, unsupervised collection methods can add noise, quality depends on the recording device as well as requires to manually process later for quality check and cleaning. Even with the inability to control environment, the new emerging trend is to techniques in collecting speech data and applications focuses on either web-based or smartphone-based approaches.

Web based approaches can either use existing resources to build the speech corpus and harness the web resources from selected contents like YouTube, social media, Online Radio as well as on the TV and etc [30] or use a customized web application and collect data over the web. The first approach it is not easy to find relevant queries to the domain among the huge available data. But the second approach has shown significant advantages over other methods and already used by the leading players of the speech recognition study like Mozilla [31] and had moved it to a next level by not just collecting data with public contributions but also for the data validation.

In contrast, mobile-based approach has gained the recent attention but as it requires additional knowledge and steps on installing the app compared to web-based approaches. In these setups a mobile application is developed and available in app stores for participants to install and once the app is installed and opened, users can record the given text query. In some of the researches, they have given the ability to record the audio offline without connected to internet and then upload recorded clips once the device is connected to the internet [10], [32]. As both of the app-based methods require an internet connection to communicate with the server, web-based applications seem more user friendly.

3. PROPOSED SOLUTION

In this paper, it proposes a domain specific voice intent classification system that uses end-to-end learning and directly transcribes input into intents without text, or any other intermediate representation. This will reduce the number of layers / models that is requested for a traditional machine learning solution and thus reduces the complexity.

As one of the identified limitations of domain like healthcare is the limitation of available good quality dataset. Therefore, this research studies the areas like transfer learning that has gained attention in recent years in order to improve the performance of a solution, where the knowledge from another domain can be utilized to learn common features of the targeting task. As almost all of the research in intent classification research on audio has used the transfer learning as a feature extractor, instead this research has focused on studying the behavior of transfer learning as a fine-tuner.

Here with fine-tuning, first a source domain with good amount of quality data will be identified. Then the features of that domain will be used to build a BLSTM model with end-to-end learning. Then this model weights will be used as the initial weights for the next model with exactly same architecture to train on the target domain, which is the healthcare domain with the limited amount of data. As of the research, all or some of the layers that is been initialized with the weights can be freeze or unfreeze. Therefore, find out the best combination that will provide a better solution.

In low resource languages or domains the synthetically collected data is limited as well as lacks in generalization. Therefore, the techniques that can be employed to reduced model overfitting or underfitting plays a huge role. The data augmentation techniques proved to increase the accuracy of a solution by increasing the data available for the model to train and learn from and at the same time generalizes the features by injecting random noises, change of speech speed, pitch of the speaker and etc.

The model architecture is based on a combination of the Bidirectional Long-Short Term and the Connectionist Temporal Classification (CTC) objective function. Furthermore, study the data augmentation techniques in the field of audio datasets and research how it can be used to improve the accuracy to address the low resource availability in low resource languages and domains.

3.1. Data Collection

Speech corpora does not exist for most of the domains. And it is one of the most challenging tasks in any study of this kind. First, a small research will be carried out to get an overall idea about healthcare call centers. This will include the questions related to the interaction of the participants with hospitals over a call. Then the results from this will be used for intent identification for the study.

Then for the data collection related to the above identified intents will be carried out using open-source tools with crowdsourcing with lesser effort. As to focus on the primary goals of the project, an easy language like NodeJS will be used with publicly available Javascript plugins will be used to collect user audio clips. Here the open-source tool "Voicer" [10], will be used as a guideline that can be accessed over any device with stable internet connection without depending on the language of the intents or the domain been selected. The collected data will be stored with privacy of the users in concern.

4. IMPLEMENTATION

4.1. Domain Identification

One of the main contributions of this research is to create a healthcare speech corpus. The task comes with lots of challenges and thus required lots of effort during this phase of the project. During the collection of the dataset below are the steps followed.

1. Identification of most common queries/ intents of the domain
2. Identify of most common inflection/ of each intent.

4.1.1. Identification of Intents

One of the leading healthcare solution providers [33] has listed top six reasons for a patient to call a healthcare center which are as follows.

- Appointments and directions
- Medical emergency
- Nurse advice
- Prescription refills
- Health insurance
- Marketing

And to get more details about healthcare call centers, multiple healthcare providers were observed and with those and above information, a survey was created to improve the quality of the domain knowledge and distributed among friends, family, colleges and open invitations for all those are interested in. One of the main goals was to identify the priority of intents that were identified during the initial stage and order those based on their frequency and urgency. 68 people has contributed to the survey and below are the related details.

The survey had 4 sections to be filled by the user.

1. User bio information and overall experience in healthcare systems.

Basic information as shown in Figure 19; like user age range: Figure 21, gender: Figure 20, highest qualification, how often the person uses healthcare service, what is the normal method used to access information when required, the overall experience: Figure 23 and etc. are some of the data collected under this section.

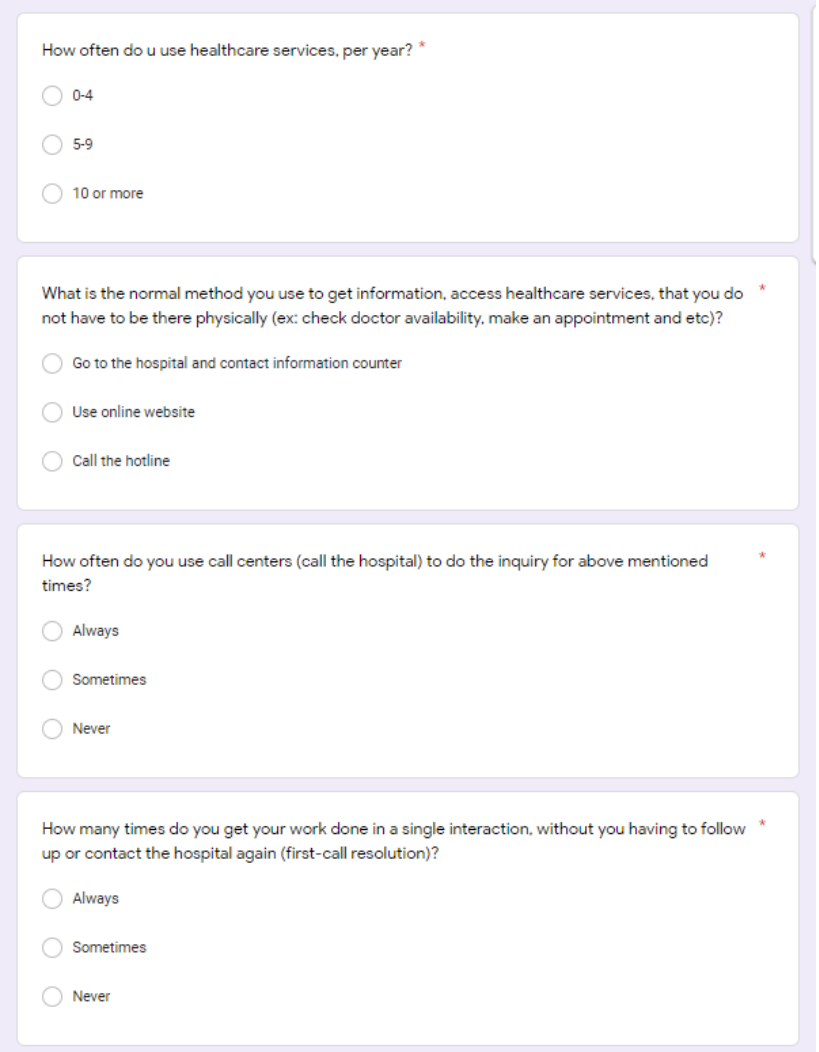

Figure 19: User basic information and overall experience

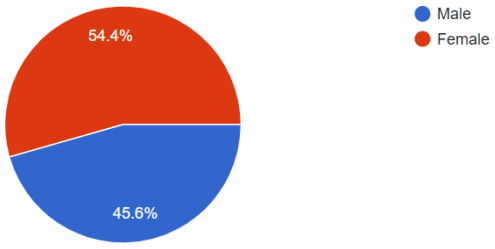

Figure 20: Gender distribution - participants

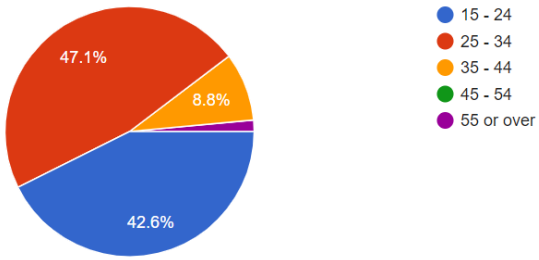

Figure 21: Age distribution - participants

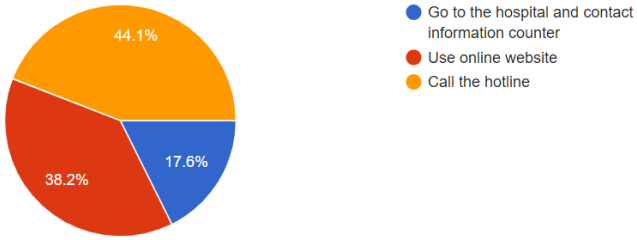

Figure 22: Healthcare access method distribution of the survey

And in the Figure 22; it shows that the majority of the people call the hotline to get necessary information.

2. Observe existing ‘Menu based’ call centers.
   Almost all of the existing call centers observed during the initial stage was based on menu-based option list which was played once a user dialed the healthcare center. To better identify the need for call center automation, couple of questions were given to the user based on the menu-based call handling systems.

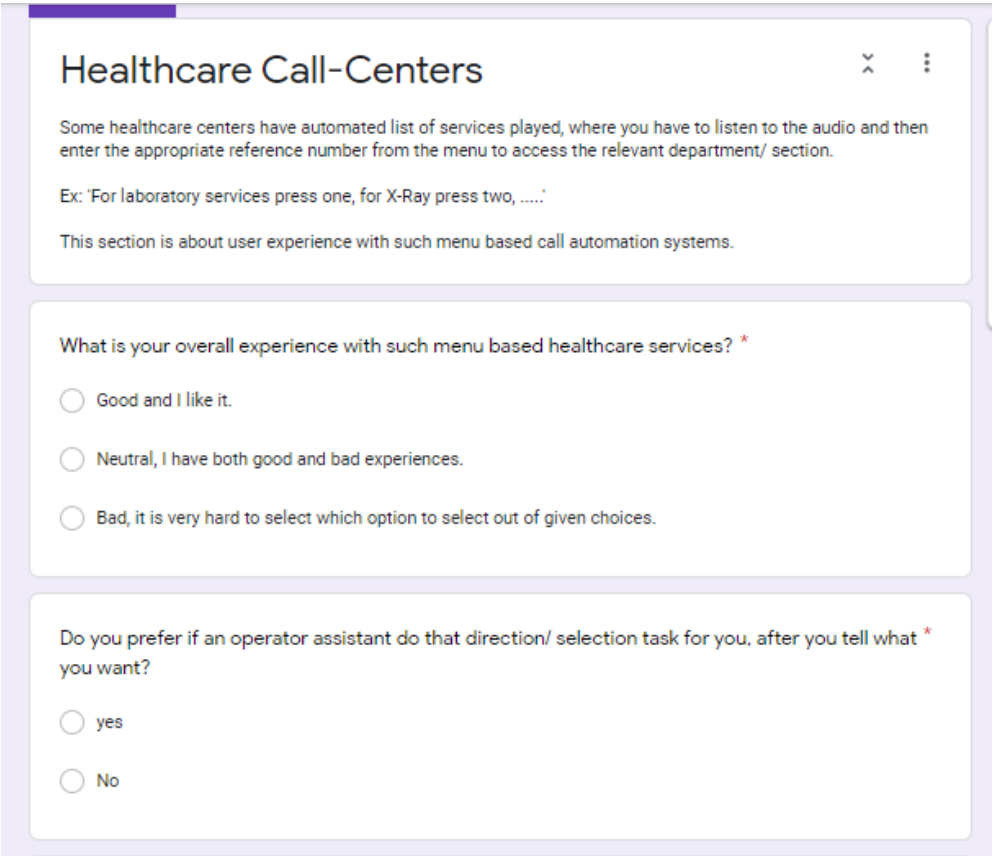

Figure 23: User experience on menu-base systems

And according to the survey it shows that the majority of the healthcare users has neutral experience towards the menu-based caller redirection. Ans this can be due to the fact that users prefer to get the information over the phone rather than physically presenting themselves at the location. The assumption is due to the fact that they have voted to have a operator assistant to redirect their calls rather than them selecting the direction path as shown in Figure 24.

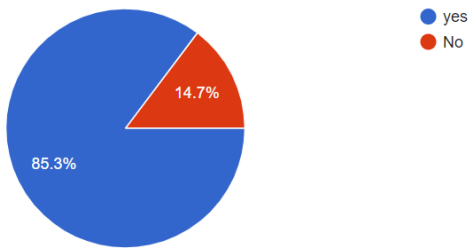

Figure 24: User preference for an operator assistant over the menu-based selection

3. Frequent inquiries/ intents

   Questionnaire asks the user to rate top 3 inquiries based on the frequency and the importance. Below are the intents given to rank.

I. Medical Emergency (call the hospital expecting a quick answer and the appropriate action including call for an ambulance)
II. General appointments and directions to get doctor availability and make a booking.
III. Appointment related changes (Cancellation, modification)
IV. Appointments for Radiology, X-ray, CT, MRI and etc
V. Laboratory services and results
VI. Seek for a 'nurse advice' / operator assistance.

According to the survey results, the most frequent calls that a healthcare centers receive is for general appointments, laboratory services and then appointment related changes; Figure 25. But when it comes to the most important ones to attend that needs immediate attention, the sequence has changed into Emergency calls, general appointments, appointment related changes and then laboratory services as top 4 intentions; Figure 26.

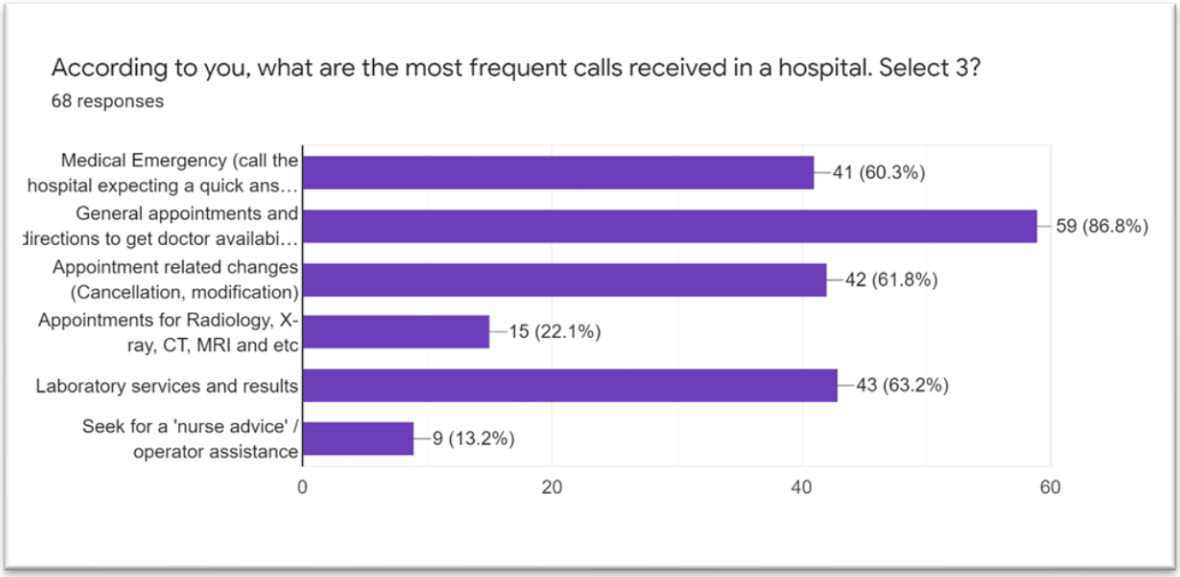

Figure 25: Most frequent queries received at a call-center.

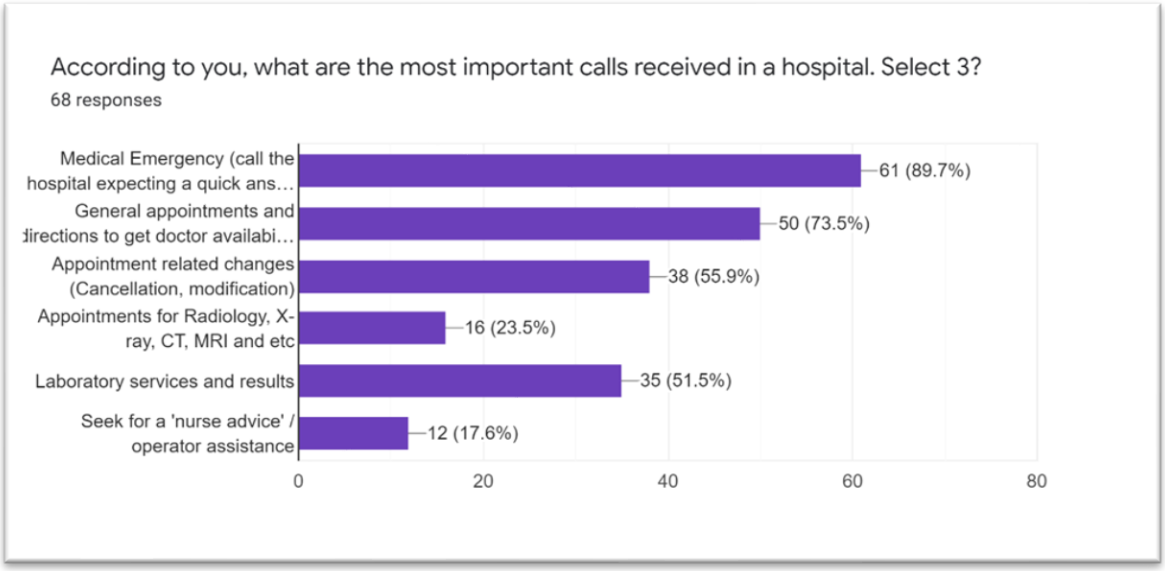

Figure 26: Most important queries received at a call-center.

Other than the queries listed above users were asked to give any other scenarios that they would contact healthcare centers. But there was no specific scenario highlighted and therefore, it was assumed that the survey has done a satisfactory job in identifying the intents of the healthcare users.

4. User inquiry variations / inflections

   In this section questionnaire asks user to give their way of asking a particular query (using 2, 3 small sentences) of a given a scenario.

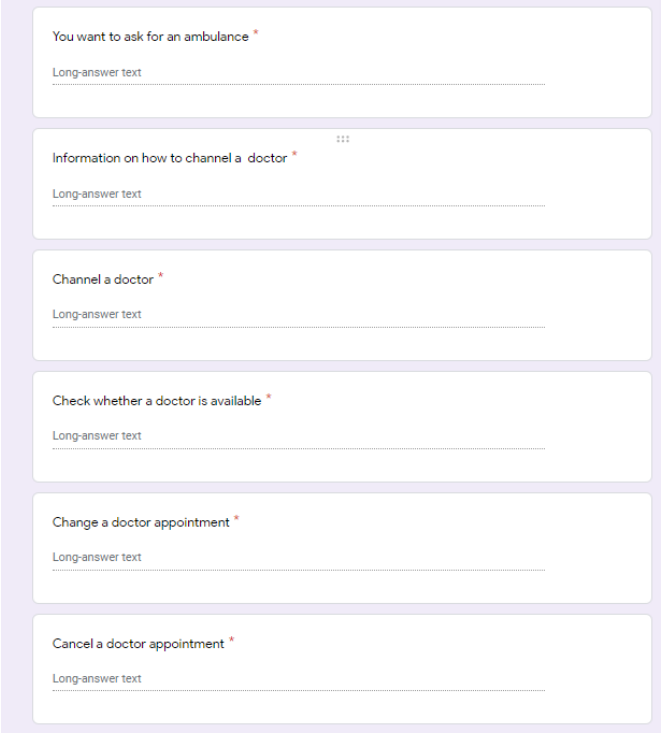

Figure 27: User inputs for individual intents

The user inputs from this section are used in identifying the user inflections for a given utterance. Figure 27, shows an outline of the form.

4.1.2. Identification of Inflections

According to the user priority identification of the user intents, it shows that the lease interested feature is ‘Seek for a nurse advice’ and therefore removed from the intent list and the rest were included as intents as moving forward from this point for the healthcare domain. By taking an average of user votes for intents for most frequent and most important, below is the intents in their significance in the domain in descending order and the number of inflections selected depending on the significance. And the total number of inflections to be 20 as shown below.

Table 1: Most common intents

Intent	Avg. Votes	Significance	# Inflections	Intent Class
General appointment related queries	55	1	5	1
Medical emergency	51	2	5	0
Appointment related changes	40	3	4	4
Laboratory services	39	4	4	2
Radiology, XRAY, MRI related services	15	5	2	3

After going through the user feedback for given queries Table 2 has the list of inflections selected for each intent.

Table 2: List of inflections for each intent

#	Inflection	Intent
1	Can I get an ambulance	0
2	I want to get an ambulance; may I know the details	0
3	Can you send me an ambulance to this address	0
4	There is an emergency so can I get an ambulance immediately	0
5	I am looking to arrange an ambulance to transfer a patient. Can you please help me with that	0
6	I want to get some information on how to channel a doctor, can you help me	1
7	How can I channel an eyes specialist from your medical center	1
8	I want to make an appointment to an eye specialist. May I know the chargers	1
9	I want to channel Dr. Pathirana. Can I have some details about how to channel him please.	1
10	Can you please check whether Dr. Pathirana is available	1

	today for an appointment	
11	I took some tests in the morning. And I would like to know whether my lab results are available	2
12	Could you please tell me, whether my laboratory report is available or not	2
13	I need to do a blood test how many hours should I be fasting	2
14	I want to get some tests done. So can I contact someone from the laboratory services	2
15	I want to cancel my doctor's appointment scheduled tomorrow. Can I get a refund for that	4
16	I am unable to come to the appointment today. Can you please cancel it	4
17	Can I change my appointment this Tuesday? And do I have to pay any additional charges to to reschedule it	4
18	I want to reschedule my appointment to see Dr. Pathirana. Can you give me any other day	4
19	I want to make an appointment to get an MRI done. What are the available days during this week	3
20	Do you have radiology facilities at your hospital	3

4.2. Data collection

After identifying the intents of the domain, the next step is to build a speech corpus for the development. Previous works to something similar was found for Banking domain [10]. For that, web-based approach was selected as it is more convenient for most of the people compared to some of the previous search attempts from the literature such as mobile applications as those are having additional steps for the participants. Node JS open-source server environment with EJS, bootstrap 4 front end with a licensed template was used for the web application. Third-party plugin [34] for the recorder function was used rather than building all the functions from the scratch.

The audio recorder has below attributes

- Bit depth – 16 bits
- Number of channels – 1
- Sampling rate – 48000 Hz

Users are given with the options to record and listen the recorded audio clip to listen before submitting. Once the user clicks to submit the current page, then the data get submitted to the server and processed relatedly rather than waiting to complete all. This was done to minimize the data loss that can happen if a user leaves the survey in the middle of the process. Therefore, with this method individual queries are shown to the user to record and those individual clips get saved once the user clicks next button.

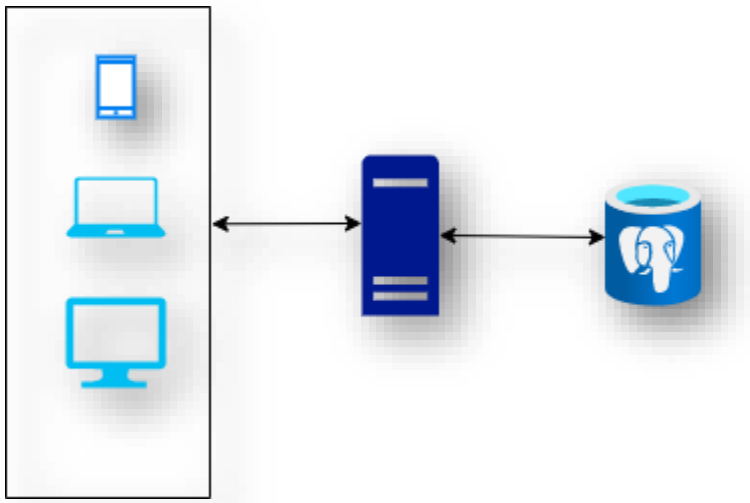

Figure 28: Data Collecting Web Application Architecture

Figure 28, shows the simple client-server architecture of the web application developed to collect voice clips with EJS, Boostrap, HTML 5 front end with NodeJS server and a PostgreSQL database to store user profile details alone with the clips recorded by the user. The choice of the technology was decided only depending on the individual expertise and with the previous working experience. After a successful completion of all the inflections, the user is given a thank you message with the ability to submit another set of clips.

In the welcome page of the application, a brief introduction to the project was given to the user in order to give participants an overall ideal of the project as this kind of data collection might be very new to non-technical people. The developed web application was shared among friends, family, and colleagues and other professional platforms for other interested parties to contribute. And only the participant's name, age and the gender were collected as additional details as shown in Figure 29. And the participant can record the audio by clicking the button to start and then stop and then "Next" to submit and continue to the next audio as shown in Figure 30.

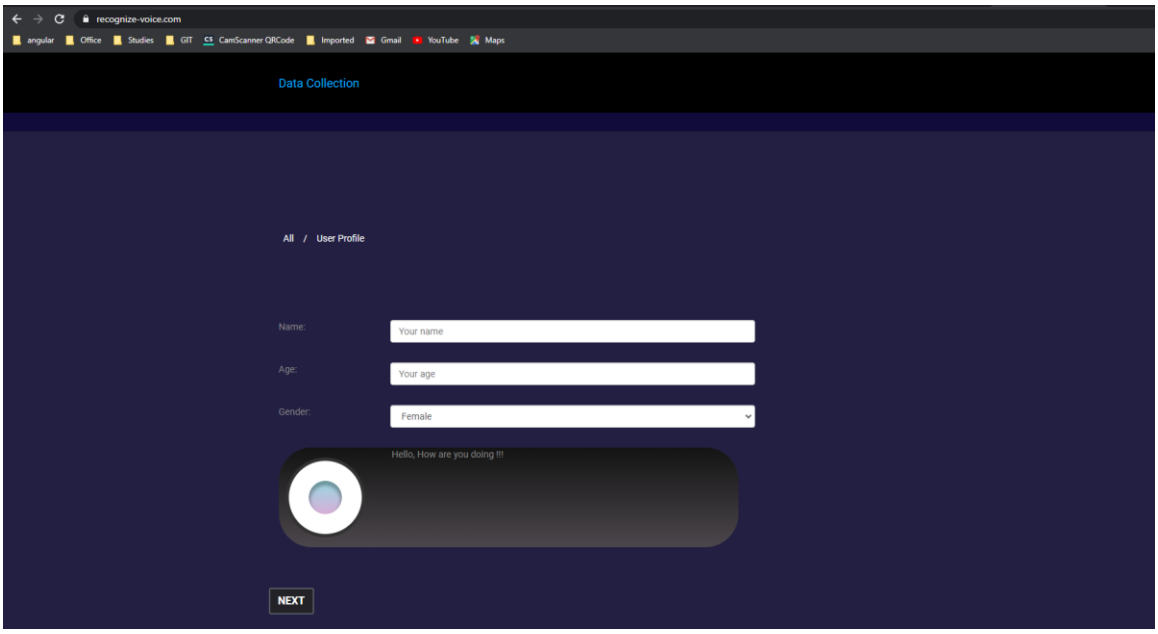

Figure 29: Web application for data collection

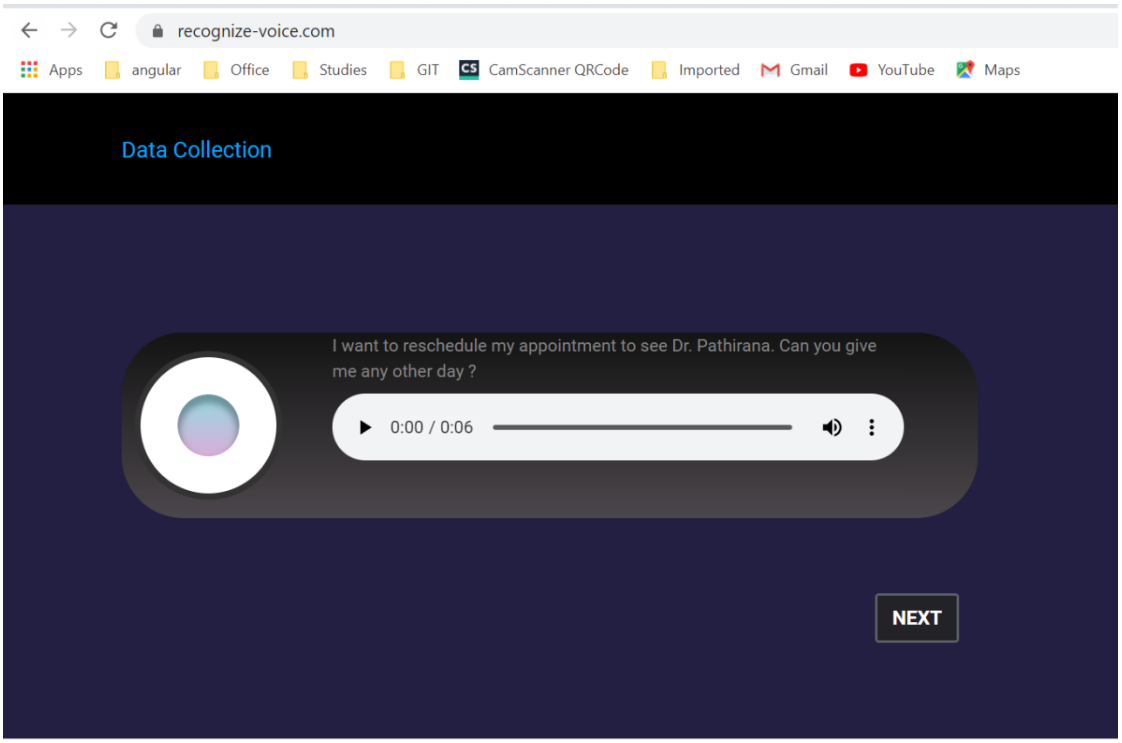

Figure 30: Record Inflections

Once a user completes the profile details a unique user ID is created to track the user throughout the rest of the process. For the first user input of audio clip of the first inflection a folder is created for the user ID and all the audio clips are saved under that directory so that later identification takes less effort for the identification without using details like the name of the user. The server, the collected data and the database was subjected to frequent backups to avoid any data loss due to unexpected crashes. The order of the intents was rotated multiple times during the period of data collection to avoid giving preference to one inflection over another.

4.3. Data preprocessing

Collected data was first visualized in different perspectives to get an overall picture of the distribution. And the gender distribution; Figure 31, of the female to male participants are 54 : 83 percent which implies that the contribution of the male is almost twice as of the female and the total number of participants are 137, where the total number of clips submitted are 2249.

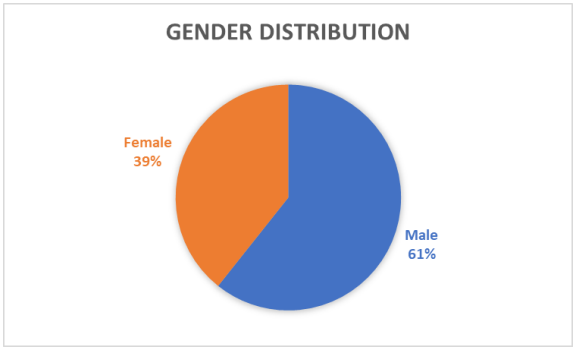

Figure 31: Gender distribution of the Audio clips

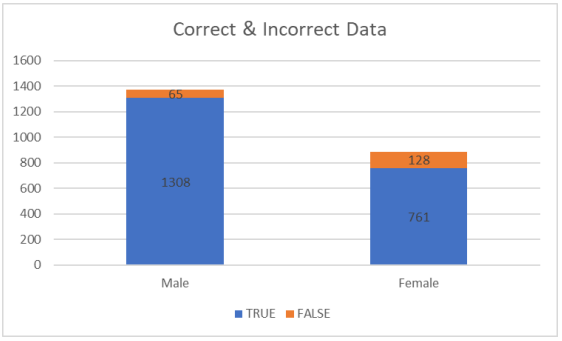

Figure 32: Audio clips by Gender and Validity

After the clips are downloaded from the server, each clip was manually listened to identify the quality. First those were categorized under the attribute valid where the valid audios met the qualities of an accurate inflections that can be used to train the model. According to the Figure 32, out of 2249 clips only 193 clips were incorrectly recorded, and the rest 2069 clips are correct. Further the incorrect clips were divided into below categories.

- NOT_CLEAR
- INCORRECT
- INCOMPLETE

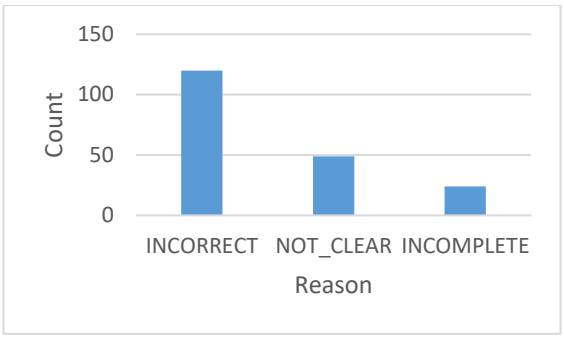

Figure 33: Invalid Clips by the Reason

With the data from *Figure 33*, the most common reason for the invalid audio clip getting recorded is that when the participant has incorrectly recorded the given text. But as a ratio the number of invalid clips are not very large and is an acceptable.

Then, the next step is to visualize the audio distribution for each intent class of the domain. As shown in the *Figure 34*; the number of audio clips for first two intent classes are almost equal and 2 and 4 are also almost equal. But the intent 3, which is the least significant intent out of 5 intents; *Table 1*, identified for the domain, has only 223 audio clips. This is less than the half of the most significant intent.

Figure 34: Audio clips distribution by the Intent

The collected dataset was around 2GB and therefore the decision to use Google Colab [35] and Kaggle; tool was decided where the GPU and TPU computing runtimes are available for free with combination of AzureML based pipelines. As the processing times of the above platforms are limited. For Colab based development, all the dataset was uploaded to the Google drive and the drive was mounted to each colab instance used.

Before any other steps, the duration of each audio clip was extracted. For that ***librosa;*** the python package for music and audio analysis was used and below are the steps followed.

1. Mounted the Google drive
2. The destination folder is opened and enumerated for all folders under the root where the individual folders belong to an individual participant and the folder name is the user ID that is unique for the user.
3. And the resulting data was saved to a csv format with rounded duration values.

According to the Figure 35, the durations of the audios rang from 0s to 27s. But the valid durations are only present from 3s to 11s and the majority of those are from 5s – 8s range.

Table 3: Dataset Duration

	# Clips	**Duration**
Valid	2069	3h 37m
Invalid	193	18m
Total	2249	3h 55m

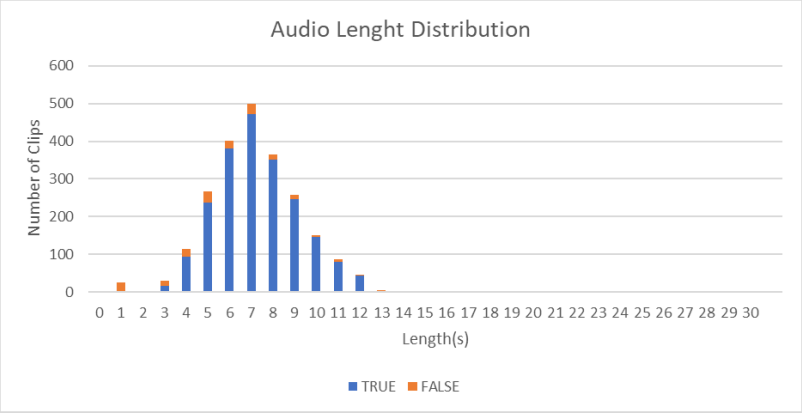

Figure 35: Number of Clips by Duration

Table 3, shows that the collected dataset is around 3h 18m and the length of the audio to be used for the model training is 11s. In the next step is to extract the MFCC features from the data alone with the class labels for model training. Same as the duration extraction the dataset was iterated from the root directory and for each audio clip the in the folder path below steps were followed.

1. A mapping function *get_class()* to get the intent class for a given inflection was created.

```
def get_class(x):
    return {
        '1' : 0,
        '2' : 1,
        '3' : 2,
        …….
        '17' : 1,
        '18' : 3,
        '19' : 4,
        '20' : 4,
    }[x]
```

2. Two json objects were created to store MFCCs and for meta data including labels, file location of the processed audio.
3. Load audio file with *librosa* using the default sample rate which is 22050Hz. This is because the fact that the human audible sounds frequency is from about 20Hz to 20kHz. And the most of the previous literature has also used this sampling rate. Therefore, the recorded audio is down sampled from 48kHz to 22.05kHz.
4. After that, 13 MFCC features from the audio is extracted with 2048 as the number of FFTs, and 512 as the hop-length.
5. The number of MFCC features for an audio with the length of 11s with the sampling rate at 22.05kHz is 474. Therefore, in order to make the length of the MFCC features with same dimension, the extracted MFCC is padded for fixed length.
6. As the file names are of <<inflectionNumber>>.wav form, it was split using the '.'delimiter and the inflection number was passed to *get_class()* function to get the intent class of the audio.
7. Features were standardized and class labels were encoded.
8. Those two attributes were added to two json files that stores MFCCs and class labels separately. And then saved as json files.
9. Then another set of MFCCs were created including only the audio clips with valid audios and saved to a different location.
10. Furthermore, to experiment the ability of augmentation in improving the model performance, a separate dataset was created with speed changed and time shifted MFCCs.

4.4. Benchmark And Proposed Architecture

The research on speech intent classification on Sinhala and Tamil dataset for Low Resource Languages (LRL) [36] has the most similar features as the research problem of this thesis and therefore taken as the benchmark for the research. The research presents speech intent classification methodology for LRLs and with 1D CNN it has achieved 97.31% accuracy for 7.5hr of dataset. The architecture is shown in Figure 37.

The benchmark model has,

- Used pre-trained DeepSpeech model to get character probability features.
- Used pre-trained model on the LibrishSpeech to get the phoneme-based probability values.
- Use those extracted features to identify intents of speech in LRL Sinhala language dataset.
- Employed 5-fold cross-validation to measure the overall accuracy with 500 epochs

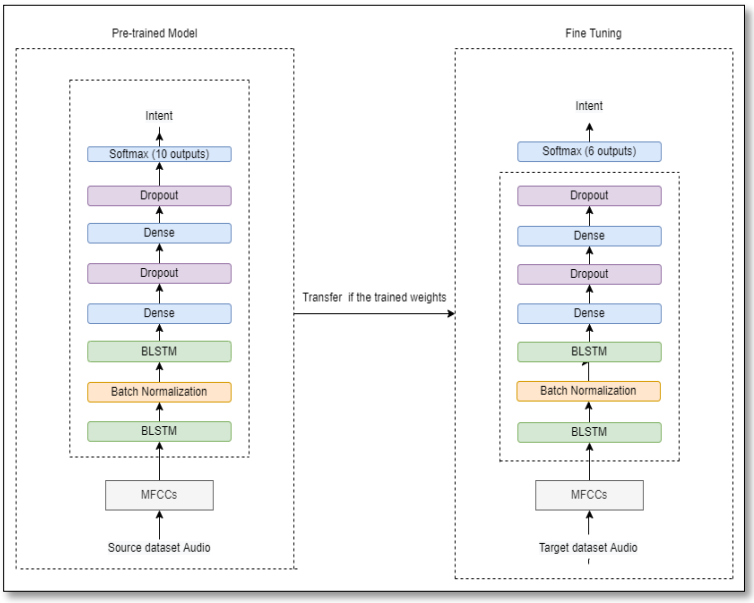

Figure 36: Proposed Model Architecture

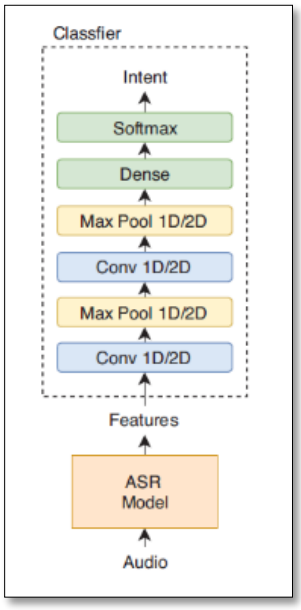

Figure 37: Benchmark Architecture

4.5. Model Training

Following parameters were used during the model building throughout the implementation.
Loss function: sparse_categorical_crossentropy
Optimizer: Adam
Learning rate: 0.0001

First the network was built by using the MFCC features of the speech-command dataset, and the model was finetuned to get the best possible performance for BLSTM classifier. Then with the knowledge gained from the pre-trained model, the experimental models were created with different batch sizes, epochs and etc to find the best model for the target dataset that can surpass benchmark performance.

As the dataset is limited, 5-fold cross-validation was employed to measure the overall accuracy of the model. And the mean accuracy and the standard deviation over the splits were calculated alone with the F1-score. Table lists the final results and Figure 38; shows that how the variance and the accuracy changed over the size of the training batch. The proposed model outperformed the benchmark with 250 iterations with 32 batch size.

Table 4: Accuracy Over Batch Size

Batch Size	Accuracy	Variance
32	97.83	0.404
64	96.24	0.5795
128	87.29	18.4905
256	84.01	15.0818
512	81.62	17.8442

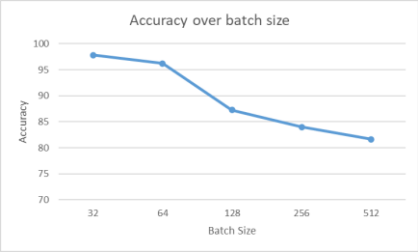

Figure 38: Accuracy Over Batch Size

The performance of the proposed solution was evaluated for the number of time / resources it takes to outperform the benchmark.

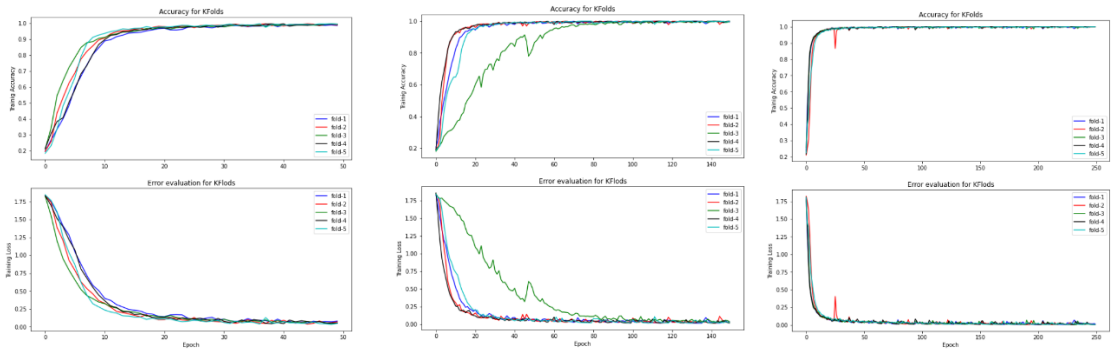

Figure 39: Training history of different epochs; 50, 150 and 250

The Figure 39, shows the behavior of the model in related to accuracy of the couple of models that were tried out during the research to decide the best model. The benchmark model is highlighted [36].

Table 5: Model Results Summary

Model	**Accuracy**	**F1 Score**	**Variance**
Without transfer learning	96.64	96.53	1.9419
Phoneme based 1D CNN – benchmark	97.31	-	-
With transfer learning	97.83	97.80	0.4840
With transfer learning and 0.6 augmentation	98.53	98.52	0.1889

Out of the first model that surpassed the benchmark was tried with different variations where it was trained with augmented data with different percentages. The summary of the final models was listed in Table . Couple of audio augmentation techniques that were identified during the literature review, were used to augment the data. Where the original audio clips were changed in speed Figure 15, and time shifting Figure 13; in time axis randomly.

In order to overcome the overfitting of the model, dropout layers and the use of L2 regularization was added for all of the models that were tried out. And Figure 40, shows the summary of the final model trained. There are 2,203,206 trainable parameters that were used to learn the dataset, which was initialized with weights from the pre-trained model on Google speech-commands dataset [37].

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional (Bidirectiona  (None, 430, 512)         552960
 l)

 batch_normalization (BatchN  (None, 430, 512)         2048
 ormalization)

 bidirectional_1 (Bidirectio  (None, 512)              1574912
 nal)

 dense (Dense)               (None, 128)               65664

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 pretrained_output (Dense)   (None, 6)                 390

=================================================================
Total params: 2,204,230
Trainable params: 2,203,206
Non-trainable params: 1,024
_________________________________________________________________
```

Figure 40: Final Model Architecture

The same model has applied to the healthcare dataset without augmentation and summary is listed in Table 6.

Table 6: Results Summary in Healthcare Dataset

Model	**Batch Size**	**Accuracy**	**F1 Score**
Model 1	32	79.16	78.84
Model 2	16	84.01	84.22
Model 3	8	85.15	85.40

Compared to the healthcare dataset, the distribution of the dataset is really high in banking domain dataset as shown and the number of intents is 6 compared to this research where there are only 5 intents. The Table 7; shows the summary of the difference of the two domains.

Table 7: Comparison of dataset of Banking domain and Healthcare domain

Dataset	Total hours	# intents	# inflections	# clips	# participants
Banking	10h	6	38	7624	215
Healthcare	4h	5	20	2249	137

Apart from the benchmark dataset, the proposed solution was applied to another dataset Free Spoken Digit Dataset (FSDD) [38], that was publicly available. Similar to the MNIST dataset available for image digit identification purpose, this dataset is created for the identification of audio digits and contains around 30,000 samples of spoken digits from (0-9) of 60 different speakers.

Table 8: Performance Comparison on AudioMnist dataset

Model	**Accuracy**	**Std**
AlexNet based - identification of audio digits, based on Spectrograms of audio	95.82	1.49
Proposed model	98.54	0.12

5. DISCUSSION

The proposed model surpassed the benchmark model for the selected Low Resource dataset for Sinhala Language. It has managed to overcome not only the performance but also the number of additional pre-trained models thus reduces the complexity and will be in capable of lowering the propagation of the errors from one model from affecting the performance of the next or the final results. And the lesser number of epochs in training means the reduced model training time, which is very crucial for both technical and business feasibility of a ML solution.

As shown in Table 5: Model Results Summary; the proposed model without augmentation has achieved equal performance of the benchmark just with 250 epochs and has very low difference in F1-score and variance. As the model keeps improving, the data shows that it also improves not only in the accuracy, but also in F1-score and the variance. Which means the model is getting more generalized. Having a higher F1-score is better as the distribution of the dataset in banking domain has class imbalance. But both the F1-score and the variance were not reported for the benchmark model. Therefore, comparison based on those parameters were not possible in evaluating the proposed model.

Even with a size of one third of the benchmark dataset, the BLSTM model achieved around 85.15% as shown in Table 6, with overall accuracy with 5-fold cross-validation. This is a very good performance. Compared to the benchmark model that is based on CNN, BLSTM learns from both positive and negative timestamps of the input data so that the features available for learning are higher.

The ability to benefit from transfer learning by fine tuning the already available is shown with the results available in Table 5: Model Results Summary. Different combinations to transfer learning which was mentioned in the literature was tried. As the top layers of the pre-trained model has the capability in learning the common features of audio and the later layers has more ability to learn the features specific to the domain, it is recommended to freeze the top layers and train only the latter layers without or without adding additional layers to optimize the learning capabilities.

When the some of the layers were freeze, the model showed very little learning capability. This can be due to the fact that there are only two BLSTM layers and are in

the top used in the model, Figure 36: Proposed Model Architecture. And there need to be more layers to learn even in the latter layers for such an approach to work. But with training all the layers of the model, it gets to generalize the knowledge gained from the source domain with pre-trained model while optimizing the available data of the target domain.

The source domain has 10 classes and contains 1s average sound clips while target domain has only 6 classes with around 11s average length audio clips. This shows that, even though the source domain the target domain feature spaces and the output is different, transfer learning with fine-tuning can be used.

As shown in the Table 8, the proposed model works well with the publicly available AudioMnist Dataset. And has outperformed the highest known accuracy. Which means the proposed model is generalized to work on other domains.

The class imbalance is one of the other things that might have affected the accuracy of the model as one intent has less than the half of the number of the audio clips of the highest. One of the reasons for that is there was lesser number of inflections of the intent and therefore that class was appeared lesser number of times during the data collection. This imbalance can be handled by having equal number of inflections for a given intent or by giving a priority for the classes that has less recorded clips by the user. But here also if the attempts have incorrect, invalid or incomplete recordings this won't be very effective.

With related to the dataset size, the banking domain has almost three times data than the healthcare domain and the number of clips is almost 3.5 times larger. Compared to the successful contribution, this value is very high. This can be due to the reason that most of the people attempted comes from non-IT background and therefore the concepts might be quite new and that might have stopped them from contributing to research. This might have been avoided if the data collection was done while the user is present under a guidance of a knowledgeable person.

For data collection applications, there were some lessons learnt and below are some of them.

- Rather than using mobile apps, the web applications help more user involvement as it is more common, but it is expected to have mobile users who are participating web-app needs be cross browser compatible in order to avoid user dissatisfaction.

- The application was developed in a way that only one clip is recorded at once and upon moving to next clip, the recorded audio file getting persisted. And this avoided loss of multiple clips if user discontinue before the completion.
- Each audio clip was first stored under the folder of the user ID and files are renamed with the inflection number. This saved lots of post-processing time and robust architecture.
- Make inflections easy and short so that the percentage of users dropping in the middle can be reduced.
- Development stack involved ExpressJs and required the web-application to host on a NodeJS supported hosting provider which came with higher cost than a normal simple application.
- Collected data is mostly from friends, family and colleagues and therefore is biased.
- Most of the users start attempting the audio recordings but didn't continue either the instructions were not clear or had doubts as the concept is novel.
- Even though a survey was carried out to identify the intents and the inflections of the domain, the collected data can be improved with the help of a language or domain specialist.
- Approaches like open data contribution as a data collection method like Mozilla [31], is useful and saves lot of manual processing and thus saves cost.

6. CONCLUSION

During the data collection there are some lessons learnt with related to sourcing data for machine learning projects. With the existing literature, a grate interest is shown towards web-based data sourcing approaches more than ever.

In this research, a BLSTM model with transfer learning with end-to-end learning was proposed to address the limitations faced by low-resource domains. The model uses MFCC features for domain specific intent classification of healthcare domains. And one of the main contributions is the collected dataset for healthcare related research and the other is a Bidirectional Long Short-Term Memory with transfer learned with fine-tuning from a different domain. It was identified that the model outperforms the benchmark and can be improved further with the quality and the quantity of the dataset and the pretrained model. The highest accuracy of the aforementioned BLSTM-TL approach is 98.53% and can be used for low-resource domain ASR systems for intent classification.

To build an intent classification system, majority of the existing literature tries to use an intermediate status, a language model to convert the voice sample into a textual representation and then build ML models to work on the text classification. Just collecting the dataset and preprocessing being costly and time consuming, as it takes lots of human involvement and the domain expertise traditional approach of intent classification seems really impossible for domains like healthcare. In order to avoid overwhelming approaches to ASR, now-a-days end-to-end classification and transfer learning types of technologies can be involved.

When selecting intents for a particular machine learning problem, it is very important to have highly discriminating intents/inflections while keeping the length of the inquiries simple and precise as the model feature space is highly depends on the length of the audios as well as the interests of the participants during corpus creation.

Furthermore, the model performance improvement with a good quality with similar feature space as a pretrained model with more target domain related speech dataset is one of the future works identified. And techniques like silence removal of audio clips to remove unwanted complexity due to the large dimensions and sounds below the human understandable range (caused by the equal length requirement of all the audio clips) are some of the areas of future research to improve the quality of the classifier.

With the related to augmentation, research the adaptability of techniques like SpecAugment [24] for the MFCCs as it can reduce the time consuming pre-processing to create augmented data directly out of raw audio, the cost associated with the storage and the maintenance. And also keep improving the corpus that can be used for both healthcare related and ASR related research.

7. REFERENCES

[1] K. Panetta, "Smarter With Gartner," Gartner, 21 October 2019. [Online]. Available: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020/. [Accessed May 2020].

[2] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Javasena and U. Thayasivam, "Domain Specific Intent Classification of Sinhala Speech Data," in *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, 2018.

[3] H. Dudley, "The Vocoder—Electrical Re-Creation of Speech," *Journal of the Society of Motion Picture Engineers*, vol. 34, pp. 272-278, Mrach 1940.

[4] H. Dudley, "The carrier nature of speech," *The Bell System Technical Journal*, vol. 19, pp. 495-515, October 1940.

[5] L. Page, "The transformation of the hospital call center," *COR Healthcare Market Strategist,*, November 2004.

[6] R. Cai, B. Zhu, L. Ji, T. Hao, J. Yan and W. Liu, "An CNN-LSTM Attention Approach to Understanding User Query Intent from Online Health Communities," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, 2017.

[7] M. A. a. H. A. V. Y. Massalin, "User-Independent Intent Recognition for Lower Limb Prostheses Using Depth Sensing," *IEEE Transactions on Biomedical Engineering*, vol. 65, pp. 1759-1770, Aug. 2018.

[8] S Kim, J.E Nah, "Workforce planning and deployment for a hospital reservation call center with abandonment cost and multiple tasks," *Computers & Industrial Engineering*, pp. 297-309, June 2013.

[9] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks.

[10] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Jayasena and U. Thayasivam, "Voicer: A Crowd Sourcing Tool for Speech Data Collection," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2018.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673-2681, Nov. 1997.

[12] L. J. Newville, "Development of the Phonograph at Alexander Graham Bell's Volta Laboratory," [Online]. Available: http://www.gutenberg.org/files/30112/30112-h/30112-h.htm. [Accessed 05 August 2020].

[13] IBM, "Pioneering Speech Recognition," IBM Laboratory, [Online]. Available: https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/breakthroughs/. [Accessed 07 08 2020].

[14] L. Rabiner, "A Tutorial on Hidden Markov Models and," *Proceedings of the IEEE*, vol. 77,

pp. 257-286, Feb. 1989.

[15] L. Rabiner, B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, pp. 4-16, Jan. 1986.

[16] T. Dinushika, L. Kavmini, P. Abeyawardhana, U. Thayasivam and S. Jayasena, "Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition," in *2019 International Conference on Asian Language Processing (IALP)*, Shanghai, Singapore, 2019.

[17] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, pp. 82-97, Nov. 2012.

[18] L. Rabiner, B.H Juang, Fundamentals of speech recognition, Division of Simon and Schuster One Lake Street Upper Saddle, River, NJ,United States: Prentice-Hall, Inc., 1993.

[19] C. K. On, P. M. Pandiyan, S. Yaacob and A. Saudi, "Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition," in *2006 International Conference on Computing & Informatics*, Kuala Lumpur, 2006.

[20] J. Hui, "Speech Recognition — Feature Extraction MFCC & PLP," Medium.com, 29 Aug. 2019. [Online]. Available: https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9. [Accessed 2020 08 10].

[21] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering*, Noida, 2016.

[22] S. J. P. a. Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. vol. 22, pp. 1345-1359, 2010.

[23] M. A. M. B. T. a. N. W. E. A. Hadhrami, "Transfer learning with convolutional neural networks for moving target classification with micro-Doppler radar spectrograms," in *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2018.

[24] W. C. Y. Z. C.-C. C. B. Z. E. D. C. Q. V. L. Daniel S. Park, "SpecAugment: A Simple Data Augmentation Method," *Proc. Interspeech*, 2019.

[25] E. Ma, "Data Augmentation for Audio," Medium.com, 1 Jun 2019. [Online]. Available: https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6. [Accessed 2021].

[26] T. Zhang, J. H. D. Cho and C. Zhai, "Understanding User Intents in Online Health Forums," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, vol. 19, 2015.

[27] C. Zhang, N. Du, W. Fan, Y. Li, C. Lu, P. S. Yu, "Bringing Semantic Structures to User Intent Detection in Online Medical Queries," in *2017 IEEE International Conference on Big Data (BIGDATA)*, 2017.

[28] J. Spitz, "Collection and analysis of data from real users: implications for speech

recognition/understanding systems," in *Speech and Natural Language: Proceedings of a Workshop*, Pacific Grove, California, 1991.

[29] J. D. Groot, "What is HIPAA Compliance?," DIGITAL GUARDIAN, December 2020. [Online]. Available: https://digitalguardian.com/blog/what-hipaa-compliance. [Accessed 06 2021].

[30] A. C. A. L. a. H. C. S. Bougrine, "Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties," in *Third Arabic Natural Language Processing Workshop,*, 2017.

[31] Mozilla, "CommonVoice," Mozilla, [Online]. Available: https://commonvoice.mozilla.org/en. [Accessed 05 2021].

[32] A. W. M. E. a. K. R. I. Lane, "Tools for Collecting Speech Corpora via Mechanical-Turk," in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, June 2010.

[33] C. Healthcare, "CHANGE Healthcare," CHANGE Healthcare, [Online]. Available: https://www.changehealthcare.com/insights/6-reasons-patients-call-hospitals. [Accessed 08 2020].

[34] M. Diamond, "github.com," [Online]. Available: https://github.com/mattdiamond/Recorderjs. [Accessed Feb 2021].

[35] Google, "Welcome To Colaboratory," Google, [Online]. Available: https://colab.research.google.com/notebooks/intro.ipynb. [Accessed 05 2021].

[36] Y. Karunanayake, U. Thayasivam and S. Ranathunga, "Sinhala and Tamil Speech Intent Identification From English Phoneme Based ASR," in *2019 International Conference on Asian Language Processing (IALP)*, Shanghai, Singapore, 2019.

[37] "https://ai.googleblog.com/," Google, 08 2017. [Online]. Available: https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html. [Accessed 03 2021].

[38] M. A. S. L. K.-R. M. W. S. Sören Becker, "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals," 2019.

[39] J Zhong, W. Li, "Predicting Customer Call Intent by Analyzing Phone Call Transcripts based on CNN for Multi-Class Classification," arXiv:1907.03715, 2019.

[40] B. H. Juang , Lawrence R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development," 2015.

[41] Müller, Meinard, Information Retrieval for Music and Motion, 2007.

[42] S. Kyaagba, "Dynamic Time Warping with Time Series," medium.com, Sep 2018. [Online]. Available: https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-with-time-series-1f5c05fb8950. [Accessed 15 08 2020].

[43] C. Kim et al., "End-to-End Training of a Large Vocabulary End-to-End Speech Recognition System," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop*

(ASRU), SG, Singapore,, 2019.

[44] J. Hu, G. Wang, F. Lochovsky, J. Sun, Z. Chen, "Understanding user's query intent with wikipedia," in *18th international conference on World wide web*, 2009.

[45] C. Clerke, E. Agichtein, Q. Guo, "Estimating Ad Clickthrough Rate through Query Intent Analysis," January 2009.

[46] S. Agarwal, A. Sureka, "But I did not Mean It!- Intent Classification of Racist Posts on Tumblr," in *2016 European Intelligence and Security Informatics Conference*, 2016.

[47] S. Team, "Deep learning for siri’s voice: On-device deep mixture density networks for hybrid unit selection synthesis," *Apple Machine Learning Journal*, vol. q, 2017.

[48] "How amazon Alexa works," channels.theinnovationenterprise.com, [Online]. Available: https://channels.theinnovationenterprise.com/articles/how-amazon-alexa-works. [Accessed 05 09 2020].

[49] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin et al., "Acoustic modeling for google home,," *INTERSPEECH-2017*, 2017.

[50] W. L. P. O. J. ZHANG, "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective," *ACM Computing Surveys*, 2019.

[51] Y. Karunanayake, U. Thayasivam, S. Ranathunga, "Transfer Learning Based Free-Form Speech Command Classification for Low-Resource Languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy, 2019.