**Abstract:**

This project focuses on developing a speech recognition engine that shifts between vocabulary identification and phonemes identification using separate distinct learning algorithm for each. Both the learning algorithm utilizes mel-filter bank features for providing feature vector of the speech signal as an input to the deep neural network model. Vocabulary identification model employs deep residual convolutional network for individual character level classification with connectionist Temporal Classification loss for decoding characters from the model and the phonemes identification model employs two variant neural network that includes convolutional architecture for frame level classification and recurrent architecture with Connectionist Temporal Classification loss for decoding. By making this model more discriminative, the model can also provide promising accuracy in Text to Speech conversion. And by further development this model autonomously can help in the intent parsing process for some embedded system. Combined Dataset of Timit and VCTK-Copus is been used for the training and testing each individual model.

1. **INTRODUCTION:**

Human being are the most advanced breed of primitives living in this planet, not because we possess more physical flexibility or strength but because of our intelligent senses that is been evolved for millions of years. Hearing is one of the sense that update our self with the external environmental we live. This hearing sense connect people and any other living being. This vital sense not just gather sounds in the form as signal but also learns different types of sound signal and classifies them efficiently with proportional to the learning cases. Sound signal became more important to the human being when the language came into the play. Hearing sense added more meaning for communication with these languages. Languages are the most important part of our being, it allow us to communicate, i.e. to convey information from one person to another. It is this ability that permits groups of individuals to transform into an information sharing community with formidable powers. Language is a complex phenomenon, and it can appear in very different forms such as symbols, facial expressions, written letters, spoken languages, etc. When we are speaking about a language, we should not restrict ourselves to think about it as a collection of sound ordered by some grammar. Any language is far more than that.

New age language theories state that what we call languages are the surface manifestations, or exterior representations, of a common inner core, not directly seen from the outside, where thought processes occur (Figure 1-1). In other words, they state that we do think using what we would normally call a language, but internal mental representations that condense all sorts of cognitive process [1]. Under this point of view, the different languages that a person uses are nothing more than a set of different interfaces between the processing results of that inner core and the community.

INNER

CORE

COMMUNITY

LANGUAGES

Figure 1-1. Languages and Inner Core Thought Process.

One form of representing languages which the help of human mouth as a articulators is the speech signal. Speech is a vital form for representing thoughts and gaining knowledge using particular languages. Recognising this speech is more important that producing one. Nowadays electronic machines are becoming as intelligence as humans in some particular domains, hence the right of to speak and to listen have been inherited to those intelligent machines. This facilitates for the development of Speech Recognition System for making machines to understand what we say. Over the years the Speech Recognition Systems have come a long way the process has ensured its presence due to the well-established need of voice operated systems. However, there is a lot to be accomplished. Most of research done so far is attributed to the fact that speech is a very subjective phenomenon [2].

* 1. **Spoken Languages**

The basic building block of any language is a set of sounds names phonemes. As an Example, a condensed list of American English phonemes, in their ARPABET representation, is given in Table 1-1 [3].

|  |  |  |  |
| --- | --- | --- | --- |
| ARPABET | EXAMPLE | ARPABET | EXAMPLE |
| IY  IH  EY  EH  AE  AA  AH  AO  OW  UH  UW  AX  IX  ER  AXR  AW  AY  OY  Y  W  R  L  M  N | beat  bit  bait  bet  bat  bob  but  bought  boat  book  boot  about  roses  bird  butter  down  buy  boy  you  wit  rent  let  met  net | NX  P  T  K  B  D  H  HH  F  TH  S  SH  V  DH  Z  ZH  CH  JH  WH  EL  EM  EN  DX  Q | Sing  Pet  10  Kit  Bet  Debt  Get  Hat  Fat  Thing  Sat  Shut  Vat  That  Zoo  Azure  Church  Judge  Which  Battle  Bottom  Button  Batter  Glottal stop |

Table 1-1. Some examples of American ARPABET phonemes

The phonemes of Table 1-1 comprise all the building blocks needed to produce what is called standard American English. Other versions of English use slightly different sets of phonemes, but similar enough to allow understanding by any English speaker. These differences explain the different accents existing in modern English. As it is in English, all languages follow this structure: they are built upon a basic set of phonemes. It is interesting to note that the number of phonemes in a set of basic phonemes never exceeds seventy. This fact might indicate a limit on the number of sounds a human can produce or distinguish in order to communicate efficiently. The second level of building blocks of a spoken language consist of the words. These building blocks are built from sequentially concatenated phonemes extracted from the basic set of phonemes of the language. An utterance of the word ‘ZERO’ is shown in Figure 1-2. It can be seen in the figure that there are roughly four distinct zones in the utterance, each of them directly related to one of the four phonemes that comprise the utterance of that word [4].

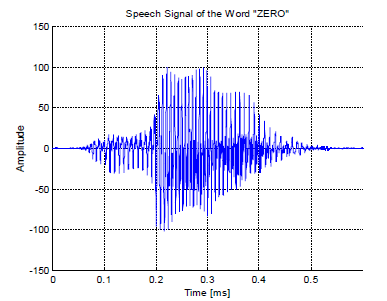


Figure 1-2. Speech Signal for the Word ‘ZERO’

All the words in a language constitute a set called vocabulary. It is generally agreed among experts that a normal person understands and uses an average of 60,000 words [1] of his native language. The next building block assembles the words into phrases and sentences. They are made up from successively braided words, according to the language’s vocabulary, grammar, and the information to be conveyed. Let us consider the following sentences:

* ‘OLD HAPPY NEW TABLE POWERLESS’
* ‘THE ELEPHANT IRONED THE WRINKLES OF HIS WINGS’
* ‘I RECEIVED A PHONE CALL FROM YOUR MOTHER’

The first one is formed by using words from an existing vocabulary, but does not respect a grammar or a context. The second respects a vocabulary and a grammar, but it is not true within a normal context. The third one respects all three aspects

Phonemes combine into words, words combine into phrases and sentences, phrases and sentences combine to form spoken language. This seems to be all, but it is not. Normally, not only what is said conveys information, but also how it is said. The manner in which expressions are uttered is described by its prosody: the characteristic stresses, rhythms, and intonations of a sentence. Each word has its own pattern of stresses that must be precisely kept, *i.e.* the meaning of the word ‘CONTENT’ depends on which vowel is stressed, but the time spanned by the utterances between all vowels is normally kept constant, no matter how many phonemes are between them [5].

**1.2 Speech Recognition:**

Speech recognition is a very popular research goal in the field of machine intelligence. There are many reasons for automatic speech recognition being widely developed by engineers and scientists around the world. Human-machine interaction is one of the most important reasons. We always dream of ordering machines such as the TV to turn itself on and change channels per our orders, thermostats to adjusting the temperature by themselves to adapt to a human’s preferences, or even a robot babysitter to do all the house tasks fast and efficiently. The basic sensory stages of the human-machine interaction are vision recognition and speech recognition. Voice recognition, which is a special kind of speech recognition, is widely used in high security locations. Due to the high demand in the current market, many corporations have already built some Automatic Speech Recognition (ASR) systems: like the dictation system used by IBM and the telephone transaction system used by T-Mobile, AT&T and Philips. Although these systems have been used in commercial area for years already, they still have many problems. First, these systems can only accomplish limited tasks such as recognizing numbers from 0 to 9, or isolated commands (e.g. transfer to customer service, balance request, pay bill, and etc.). Second, they all lack robustness, i.e. these systems have very poor performance in a noisy environment. Some of the “smart” recognition systems can recognize a word, a sentence or even a paragraph but require to be adapted to every new user, so every new user needs to train the system to recognize his/her specific voice. This approach is not suitable or feasible for a commercial use. These problems lead researchers and scientists to improve the speech recognition systems. There are three speech recognition technologies that have been developed over the years:

1. Dynamic time warping: an algorithm for measuring similarity between two sequences which may vary in time or speed. However, this technology has been displaced by the more accurate Hidden Markov Model (HMM).

2. Hidden Markov Model: a statistical model in which the system being modelled is assumed to be a Markov process with unknown parameters. This algorithm is often used due to its simplicity and feasibility of use.

3. Neural Network based approach: an algorithm for training the system to recognize speech using an artificial neural network. This technology is capable of solving much more complicated recognition tasks, and can handle low quality, noisy data, and speaker independence. If properly developed and used it may be more accurate than HMM.

Although, the Hidden Markov Model is the most popular method used in the commercial speech recognition field due to its simplicity and feasibility of use, due to its drawbacks many researchers focuses on neural network based approaches. HMM relies highly on the accuracy of the model phonemes and is state dependent. If the nature of the speech is not the same as the given sample or the next phoneme in the word depends on more than just the previous state, then the recognition rate drops dramatically. In particular, the HMM model cannot properly represent the context of the processed speech, which is an important property of human speech recognition. Another drawback of HMM products is that they are speaker-dependent. The system needs to be trained to create templates of the phonemes and words for each user, i.e. before any user starts to use the recognition system, he/she always needs to train the system with a number of sample words which contains all the phonemes by repeatedly speaking these samples and representing them to the system, then the system can recognize his/her speech by calculating the probability of the current phoneme compared to the database models. This training stage is time consuming. The third disadvantage is that the HMM model always drops the low probability word transitions although they may contain the correct information [6]. . Procedural Diagram of a Conventional Automatic Speech Recognition model is been illustrated in the following figure 1-3.

Feature

Extraction

Model Database

Select

Maximum

M

Probability

Calculation

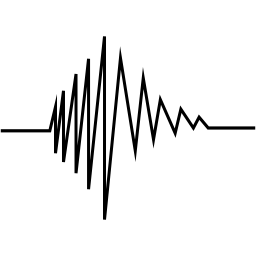


Figure 1-3 Conventional Automatic Speech Recognition

A Neural Network (NN) based approach may present a solution to the above drawbacks. Unlike HMM, the NN approach does not require template creation and to a large degree is speaker independent. The recognition system only needs to be trained once to generate the structure of the network. So the total training time for the system is significantly reduced comparing to the HMM. In the neural network system, all the output neurons, where one neuron represents one recognition category, fire at a certain excitation level all the time, so the system does not lose any useful information even for incorrect outputs. In addition, since NNs are compatible with neural based associative memory structures, they are more likely to use speech context in speech recognition. Previously activated words or concepts may be reused to help recognize a new word or sentence and may remove ambiguity from understanding similarly sounding words. In addition, using speech context will help to overcome the effect of noise or speech interference in case of cocktail party speech recognition.

There are even some Drawbacks faced by normal Neural Network model for Speech Recognition like over fitting, trade-off and also Conventional Neural Network model doesn’t promise with state-of-art accuracy, hence researchers started to push the boundaries of Speech Recognition model with implementing it with Deep Neural Network.

In general, at the present there is no such speech recognition system that is accurate and robust to all conditions and applications in real world speech based communications.

Researchers are looking for a way to simulate human hearing perception, but no one yet is close enough to human ability to understand the speech signals.

1.3. Feature Extraction:

In automatic speech recognition, it is common to extract a set of features from speech signal. Classification is carried out on the set of features instead of the speech signals themselves. The feature extraction stage seeks to provide a compact representation of the speech waveform. This form should minimise the loss of information that discriminates between words, and provide a good match with the distributional assumptions made by the acoustic models [7]. The Speech signal can be represented by a sequence of feature vectors in order for the application of mathematical tools without the loss of generality. Most of these features are also used for speaker dependent Speech Recognition system. In practical real life systems, several of these features are used in combinations. Some of the desirable properties for features sets [8] are as follows.

1. They should preserve or highlight information and variation in the speech that is relevant to the basis being used for the speech recognition and at the same time minimize or eliminate any variation irrelevant to that task.
2. Feature space should be relatively compact in order to enable easier learning of models from finite amounts of data.
3. A feature representation that can be used without much consideration in most circumstances should be used.
4. The process of feature calculation should be computationally inexpensive. Processing delay (i.e. how much of the ’future’ of the signal you have to know before you can emit the features) is a significant factor in some settings, such as real-time recognition.

In the classification problem the speech feature extraction is used for reducing the dimension of the input vector, while maintaining the perceptive power of the signal. Feature extraction is a special form of dataset and it results in extraction of specific features. These features carry the characteristics of the useful information regarding speech. Feature design and selection is the main challenging problem in the speech recognition system development for specific application. For speech identification and verification development, the number of training and testing vector are needed for the classification. The problem grows with the dimension of the given input. The techniques available in enrich literature for speech feature extraction is described in table 3.1, with their properties [9].

|  |  |
| --- | --- |
| **Method** | **Property** |
| Principal Component Analysis  (PCA) | • Eigenvector-based method.  • Nonlinear feature extraction method  • Supported to Linear map.  • Faster than other technique.  • It is good for Gaussian data. |
| Linear Discriminate Analysis (LDA) | • Linear feature extraction method  • Supported to supervised linear map.  • Faster than other technique,  • Better than PCA for classification. |
| Independent Component Analysis (ICA) | • Blind course separation method  • Support to Linear map  • It is iterative in nature  • It is good for non- Gaussian data |
| Linear Predictive Coding (LPC) | • Static feature extraction Method. • It is used for feature Extraction at lower order coefficient |
| Cepstral Analysis  Mel-Frequency Scale Analysis | • Static feature extraction method. • Power spectrum method. • Used to represent spectral envelope  • Static feature extraction method.  • Spectral analysis method.  • Mel scale is calculated. |
| Filter Bank Analysis | • It required frequencies possible • Used for filter based feature extraction |
| Mel-Frequency Cestrum Coefficient (MFCC) | • Power spectrum is computed by performing Fourier Analysis,  • Robust and dynamic method for speech feature extraction |
| Kernel Based Feature Extraction Method | • Nonlinear transformations method |
| Wavelet Technique | • Better time resolution than Fourier Transform, Real time factor is minimum |
| Dynamic Feature Extractions   1. LPC 2. MFCC | • Acceleration and delta coefficients • II and III order derivatives of Normal LPC and MFCCs coefficients |
| Spectral Subtraction | • Robust Feature extraction method |
| Cepstral Mean Subtraction | • Robust Feature extraction method for small vocabulary based system |
| RASTA Filtering | • Used for Noisy speech recognition |
| Integrated Phoneme Subspace Method (Compound Method) | • A transformation based on PCA + LDA + ICA.  • It gives Higher Accuracy than the existing Methods. |

Table 1-2. The Feature Extraction Technique with their comparative properties

In general speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. The main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs spectro temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vector into more compact and robust vectors that are then supplied to the recognizer[10][11].

WINDOWING

DISCRETE FOURIER TRANSFORM

MEL FREQUENCY WARPING

INVERSE DFT

LOG

Figure

**1.4. Deep Neural Network:**

Deep Neural Networks (hereafter, DNNs) are a part of the machine learning field whose success is relatively new. Under this name, a range of structures is included and all of them have something in common: taking classical neural networks (the shallow ones, which are those whose structure is composed of just one hidden layer) as a starting point, hidden layers are added so that they allow dealing with complex problems within the machine learning field, in which traditional structures are limited. Thereby, DNNs try to emulate the complex human learning system, extracting features at multiple levels of abstraction and learning complex functions directly from the input data, without depending completely on human-crafted features. The ability to automatically learn powerful features is becoming increasingly important as the amount of data and range of applications to machine learning methods carries on growing [12]. However, successful experimental results using deep architectures with more than one or two hidden layers were not reported until 2006 [12] (except for the case of convolutional networks, that will be described in section 2.2.5) due to limitations in training this kind of structures. Some of these limitations were not coming from a theoretical point of view, but they were due to limitations in hardware devices. Moreover, although algorithms to train them already existed, the random initialization of the parameters makes these architectures yield poor results [13]. All these limitations were solved with the evolution of hardware devices and with a learning algorithm that greedily trains one layer at a time [14], taking advantage of unsupervised learning algorithms to initialize the parameters. Nevertheless, the huge number of free parameters to train in this kind of neural networks is one of the main disadvantages that they have. Moreover, other of their drawbacks that it can

be highlighted is the computational cost that their training algorithms present and the amount of data that should be used to train the whole architecture.

DNNs are used in classification tasks in which the goal is to assign one of C classes, or labels, to each input. The input is a feature vector that describes some observation. For example, in the MNIST classification task from machine vision, each input is a vector of pixel intensities taken from images of isolated handwritten digits. The classes are the digits 0-9, and the task is to predict the digit present in each image [15]. To solve classification tasks, we build a model, such as the DNN, that can estimate a posterior probability, P (class|input), for each class. For simple classification tasks, such as the digit identification problem described above, the predicted class for an input is simply the class with the highest posterior probability. For a more complicated task, such as speech recognition, the posterior probabilities are used with other systems to solve the task.

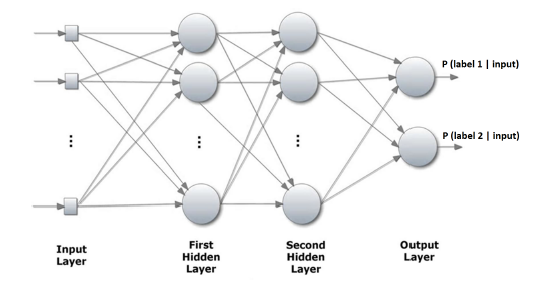


Figure 1-4. A DNN with 2 hidden layers and an output layer [16]. Each hidden layer typically has thousands of neurons, and the output layer has one neuron per class.

Consider a DNN that has input x, which is the feature vector, and output y, which is the vector of class probabilities. The DNN consists of several hidden layers and an output layer. Each hidden layer typically has thousands of neurons. The output layer has C neurons, one for each class. An example DNN is shown in Figure 1-4.

Denote the output of each layer as *yn,* where n is the layer number. The first hidden layer starts *at n = 1*. The output layer has *n = N*, where N is the number of layers in the network. The input can be treated as a pseudo-layer: *y0* = x. Each neuron has a weighted connection from every output in the previous layer; there are no intra-layer connections. An example neuron is shown in Figure 1-5

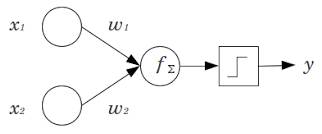
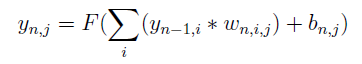


Figure 1.5. An Individual Neuron

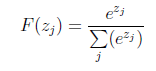
The output of each neuron is a weighted sum of its inputs and a bias term, which is passed through an activation function F. The output of each neuron in the nth layer, *yn;j* ,can be computed as:



Where neurons in layer n are indexed by j and neurons in layer n-1 are indexed by i. The weight from i to j is denoted by *wn;i;j* , and the bias of neuron j is denoted by bn;j. For the hidden layers, common activation functions include the sigmoid, tangent, and rectifier functions. We use the sigmoid activation:



The output layer, also called the softmax layer, uses the softmax activation function:



Since the softmax normalizes to the range [0, 1] and ensures that it sums to 1, the outputs can be treated as posterior probabilities. It is believed that the power of a DNN comes from its depth. Layers closer to the input are able to learn low-level representations of the input, while layers further up are able to learn higher-level representations [17].