SPEECH CONTROLLED ROBOT CAR

USING DEEP NEURAL NETWORK

A project report

Submitted by   
Aayush Kafle (070BEX403)  
Amir Rimal (070BEX405)  
Arun Tamang (070BEX406)  
Pankaj Dhakal (070BEX424)

Of Minor project in Electronics and Communication Engineering

IOE Pulchowk Campus, Pulchowk

Tribhuwan University

August 2016

Tribhuwan University

**Bonafide Certificate**

Certified that this project “SPEECH CONTROLLED ROBOT CAR” is the bonafide work of “Aayush Kafle, Amir Rimal, Arun Tamang, Pankaj Dhakal “ who carried this work under my supervision.

……………………….. ……………………….   
(Signature) (Signature)

Dr. Diwakar Raj Panta Dr. Ram Krishna Maharjan

(Head of Department) (Supervisor)

**Copyright**

Copyright © 2016

All rights reserved. No part of this report may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the author, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law. For permission requests, write to the author address below.

[Aayushkafle@gmail.com](mailto:Aayushkafle@gmail.com)

[Amir.rimal007@gmail.com](mailto:Amir.rimal007@gmail.com)

[Aruntamang1973@gmail.com](mailto:Aruntamang1973@gmail.com)

[Pankrock@gmail.com](mailto:Pankrock@gmail.com)

**ACKNOWLEDGEMENTS**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Department of Electronics and Computer Engineering for providing necessary information regarding the project & also for their support in completing the project.

We would like to express our gratitude towards Dr. Diwakar Raj Pant and Dr. Ram Krishna Maharjan for their kind co-operation and encouragement which helped us in completion of this project.

Our thanks and appreciations also goes to our colleagues in developing the project and people who have willingly helped us out with their abilities.

**Abstract**

This project provides a concept of implementing Neural Network in speech processing, recognizing speech and working with this accordingly. The task designated is to recognize voice commands (start, stop, left, right, back) and drive the car accordingly. We have implemented feature extraction approach from speech to recognize the given signal. The features from speech are extracted as Mel Frequency Cepstrum Coefficients (MFCC). All the computational works are carried out in Computer itself. The car can also be controlled with an android device.

**1. Introduction**

* 1. **Background**

With the constant improvement in the field of science and technology it was only the matter of time before something be created that was once deemed impossible. The revolution in science and technology has changed the way human beings live, it has affected our way of living in every other field. Machines have started mimicking the human way of living, even recognize persons and objects, understand human voice and communicate with human beings. All of these machines have intelligence to recognize and process simple to complex human commands.

A lot of tech giants including Google (Google Voice), Microsoft (Cortana), Apple (Siri) have already stepped in the field of speech processing and provide their user with virtual, voice based assistants.

This project is the implementation of how speech processing and recognition can be implemented in our day to day living.

* 1. **Objectives**

The major objectives of this minor project comprise of following points:

1. Recognize and process human speech in low level.
2. Develop embedded system (car).
3. Interface the processed human speech to drive car.
4. Improve the spirit of team work.
   1. **Scope of project**

Machines have already been inseparable portions of human life and with the advancement in machine learning the dependency of human beings with the machines has only increased. A lot of companies have already developed voice based virtual assistants. Implementing machine learning in it has widened the area of its application.

We have developed a three wheeler with two programmable wheels and one caster wheel. The computer recognizes the speech and gives this car command via Bluetooth and the car moves accordingly.

This system can be used where environment does not favor human beings as command can be given remotely and system works accordingly.

1. **Literature review**
2. A research article submitted by ‘Hasim Sak, Andrew Senior, Francoise Beaufays’ of Google Inc. on Long Short Term Memory (RNN) architecture for large scale acoustic modelling. It showed that the deep LSTM architecture achieved the state of art on large scale acoustic modelling. The training time and the error for speech recognition was drastically minimized with the increasing depth of layer.
3. A research paper published by ‘Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton’ from Department of Computer Science, University of Toronto showed that the combination of deep, bidirectional Long Short Term Memory RNNs with end to end training and weight noise gave state of art results in phoneme recognition on TIMIT database.
4. **Theoretical Background**
   1. **Speech recognition**

Speech based communication is the most common form of communication between human beings. Speech is that form of communication which not only carries linguistic message but carries information about gender, age, race, origin and even human emotions. When it comes to computer based devices the term speech recognition in its primary form means breaking down the speech into component words. In secondary level it not only provides linguistic message but also provides literature, acoustic and cultural message.

The most common speech recognition system can be found in our smart phones whether they are high end or not. Speech recognition can be also used to recognize speaker which in turn can be used for authentication and verifying the identity of speaker.

* 1. **Practical Speech recognition**

The practical speech recognition system involves three primary steps namely

* + 1. **Preprocessing**

In order to increase the efficiency of feature extraction the speech signal fed must first be preprocessed. The output of preprocessing step contains compressed and filtered speech frames which are in turn fed to extract feature. The necessity to preprocess the signal is to generate a signal which is very compact yet contains all the necessary information. The preprocessing process involves three further sub processes namely:

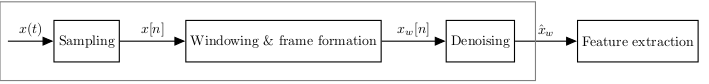


Fig: preprocessing

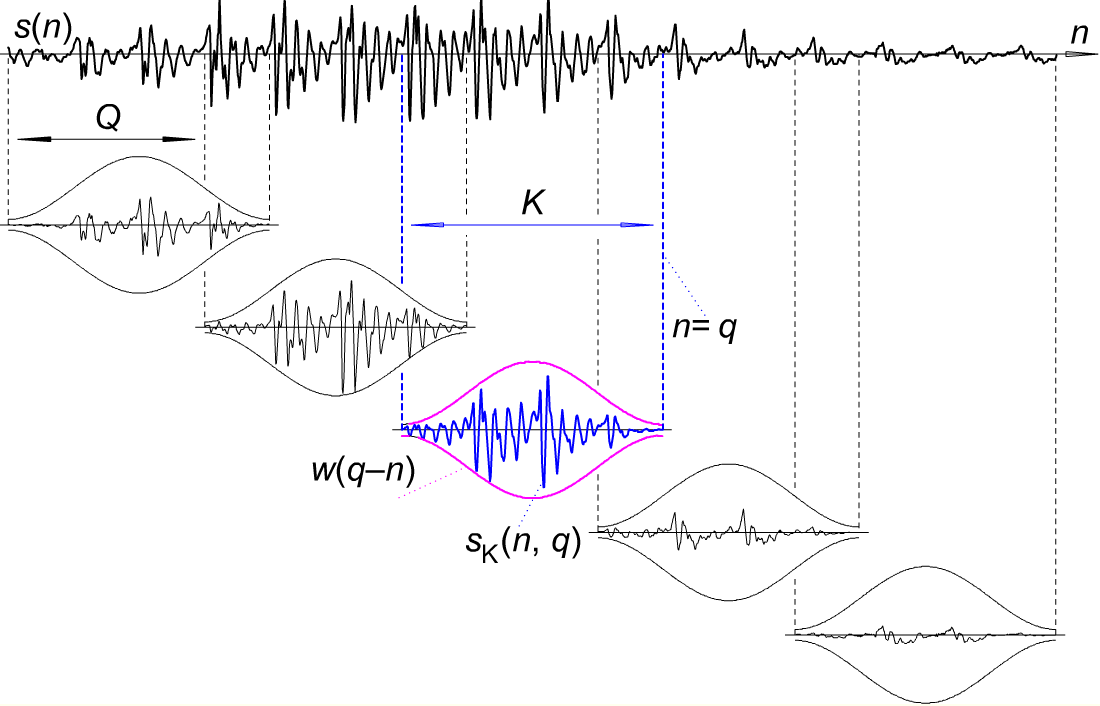
1. **Sampling**

Sampling refers to the digitization of input signal. The time continuous input signal is sampled and quantized. The result is a discrete signal. The BW of human hearing is about 8 KHz (100 Hz- 8 KHz) so according to Nyquist criteria sampling done at 16 KHz is sufficient. For the purpose of having a value discrete signal the sampled values are quantized. This leads to a significant reduction of data. Usually speech recognition systems encode the samples with 8 or 16 bits per sample depending on the available processing power. 8 bit per sample would mean 28= 256 quantization levels, 16 bit per sample provide 216 = 65536 quantization levels. Concluding, if we have enough processing power, a higher bit resolution for the sampled values is preferable.

1. **Windowing**

Speech is time variant and non-stationary signal. Although speech signal in itself is time variant however the fundamental parts of speech, phonemes, remain time invariant for a short period of time (~5-100 ms). So, for this time frame the signal can be assumed stationary.

In order to obtain the frame the signal is multiplied by windowing function.



Frame length (Q) = 5-25 ms

Window length (K)= 20-25 ms

1. **Denoising**

Denoising refers to the process of improving/enhancing the speech signals degraded by noise. The objective is to improve the intelligibility, a measure of how comprehensible speech is.

* + 1. **Feature extraction:**

This component/stage of speech processing is responsible for deriving/extracting the required descriptive features from the windowed and enhanced speech signal to enable the classification of signal. Raw speech signal contains the information beside the linguistic message and has high dimensionality, since both of these characteristics of the raw speech signal would be unfeasible for the classification so feature extraction is required.

The feature extraction algorithm derives a characteristic feature vector with a lower dimensionality, which is used for classification of sounds. Only the linguistic information needs to be emphasized, speaker dependent characteristics, the characteristics of environment like noise should be suppressed. This suppression of random variables is also called dimensionality reduction. It is very essential to choose a right feature extraction method to represent the acoustic signal. Following are the measures of feature extraction:

1. **Linear Predictive Coding (LPC):**

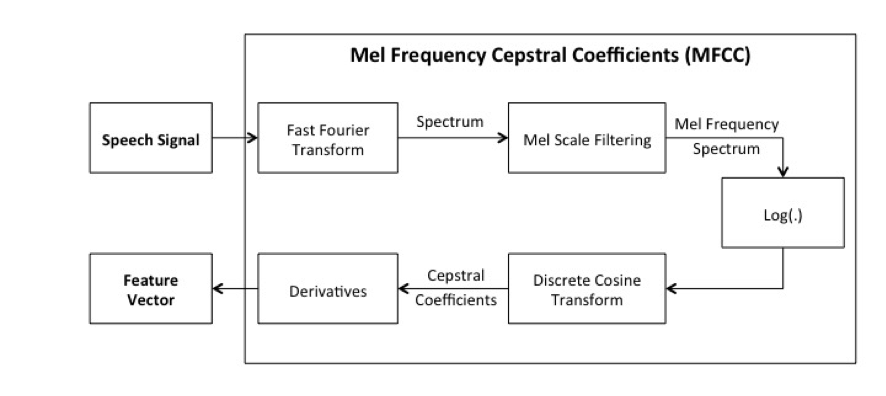
This is a method for signal source modelling in speech it is often used a format extraction method however this is widely used in speech recognition system as well. This is one of the most powerful method for encoding quality speech at a low bit rate. The basic idea is that a specific speech sample at the current time is approximated by the linear combination of past examples.

1. **Pure Fast Fourier Transform (PFTT):**

Direct use of vector coefficients of FFT power-spectrum of the speech signal is also possible as speech feature for speech recognition system. Pure FFT spectrum carries information about the speech signal than mimicking the human auditory system. If higher sampling rates is used, extra information is located in the higher frequency bans which are not considered salient in speech recognition.

1. **Mel Frequency Cepstral Coefficients (MFCC):**

To extract a feature vector containing all information about the linguistic message, MFCC mimics some parts of the human speech production and speech perception. MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. To represent the dynamic nature of speech the MFCC also includes the change of the feature vector over time as part of the feature vector.The implementation of MFCC is explained and shown in figure below:



* **Fourier transform**

The first processing step is the computation of the frequency domain representation of the input signal. This is achieved by computing the Discrete Fourier Transform.



Where N is the number of sampling points within a speech frame and the time frame τ. For implementations the Fast Fourier Transform, which is a variation of the Discrete Fourier Transformation optimized for speed, is used.

* **Mel frequency spectrum**

The second processing step is the computation of the mel-frequency spectrum. Therefore, the spectrum is filtered with Nd different band-pass filters and the power of each frequency band is computed. This filtering mimics the human ear because the human auditory system uses the power over a frequency band as signal for further processing. This processing step can be described by,



where d is the amplitude of the band-pass filter with the index j at the frequency k. The filter bank with the band-pass filters cannot mimic the ear because the ear can use any frequency as center frequency. For ASR Nd equidistant band-pass filters on the mel scale are used. The mel-scale is a non-linear scale that is adapted to the non-linear pitch perception of the human auditory system. The number, the shape (triangular, trapezoidal rectangular) and the center frequency of the band-pass filters can be varied. Figure below shows a typical filter-bank with 25 triangular band-pass filters. Some research suggests that too few and too many band-pass filters have a negative impact on the classification performance and that overlapping rectangular shaped filters achieve a better performance compared to triangular shaped filters.

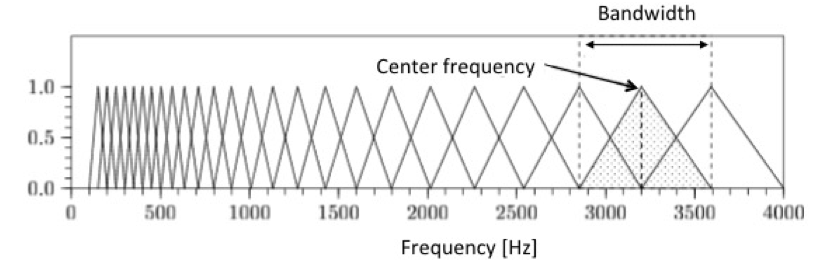


Figure: Filter bank with 25 triangular bandpass filters to compute the mel frequency spectrum

* **Logarithm:**

The third processing step computes the logarithm of the signal, to mimic the human perception of loudness because experiments showed that humans perceive loudness on a logarithmic scale.



* **Cepstral coefficients:**

The fourth processing step tries to eliminate the speaker dependent characteristics by computing the cepstral coefficients. From the Source-Filter model is known, that the signal is the convolution of the speaker dependent source signal and the filter signal. To suppress the source signal the cepstrum is computed. The cepstrum can be interpreted as the spectrum of a spectrum. Therefore, the speaker dependent harmonics of the fundamental frequency are transformed to one higher order cepstral coefficient under ideal conditions. The inverse transformation of the lower cepstral coefficients show the frequency response of the vocal tract and the inverse transformation of the higher order cepstral coefficients show the frequency spectrum of the source signal. Therefore, the speaker dependent harmonics are suppressed by taking the lower order cepstral coefficients for further processing. The cepstrum of a signal is computed by,

http://latex.codecogs.com/gif.latex?\dpi%7b120%7d&space;F%5e%7b-1%7d&space;\left&space;\%7b&space;log(&space;\,&space;F\left&space;\%7b&space;f_%7bn%7d&space;\right&space;\%7d&space;\,&space;)&space;\right&space;\%7d

where f is the input signal and F is the Fourier Transformation. The computation of the logarithm can be omitted because the logarithm of the signal was computed in the previous processing step. Instead of the Fourier Transform the discrete cosine transform can be used because the absolute value of the spectrum, respectively the periodic continuation of the signal, is real and symmetric. The cepstral coefficients are computed by

where Nmc is the number of chosen cepstral coefficients for further processing. Typically Nmcis in the range of thirteen to twenty.

* **Derivatives:**

All previous processing steps included information about the current signal frame. To represent the dynamic nature of speech the first and second order derivatives of the cepstral coefficients extend the feature vector.



The final feature vector is

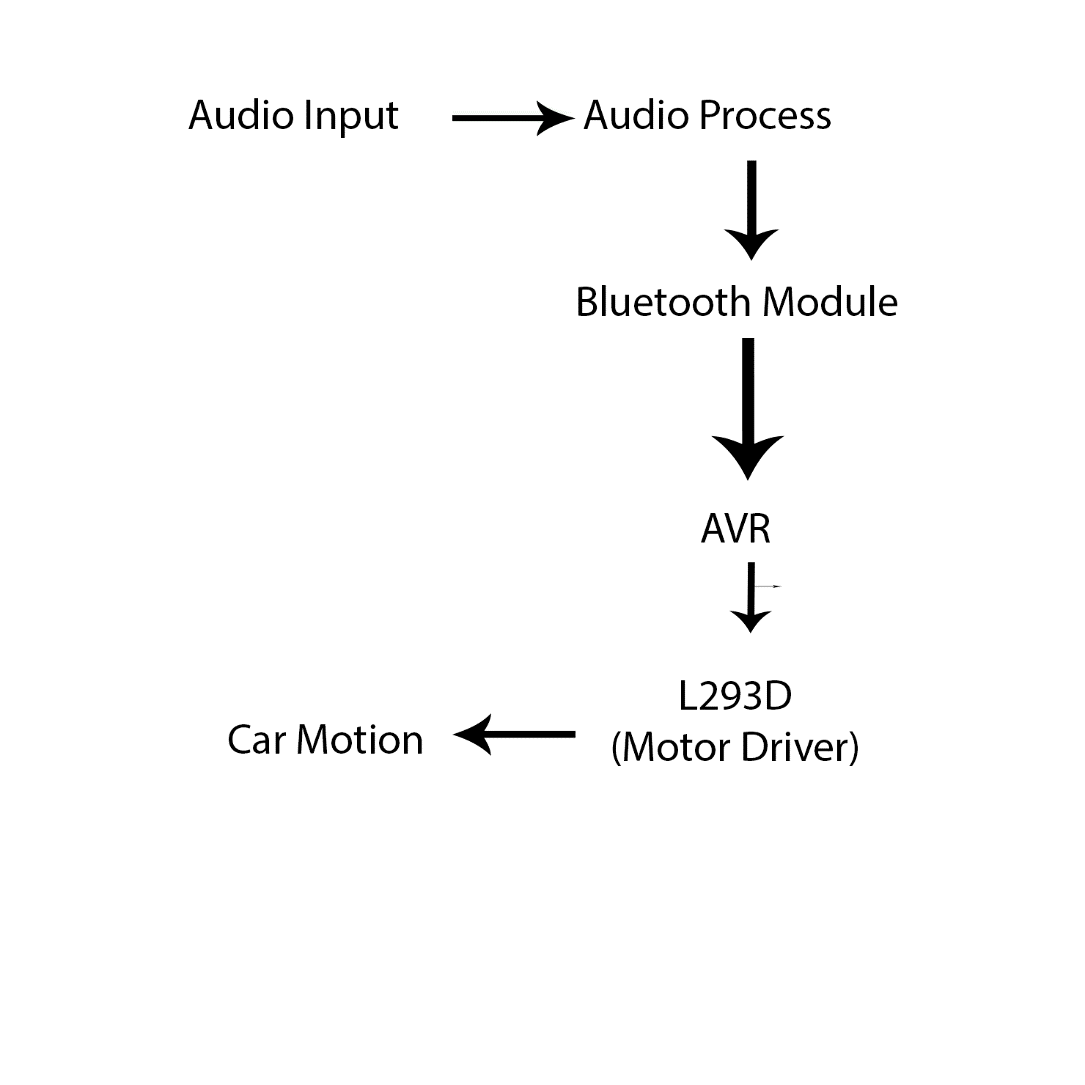


A typical MFCC feature vector would be calculated from a window with 512 sample points and consist of 13 cepstral coefficients, 13 first and 13 second order derivatives. This example would reduce the dimensionality from 512 to 39 dimensions.

1. **Methodology**

In order to meet our objectives we implemented the following methods:

* 1. **System block diagram**



A detailed explanation of above steps in given below:

1. **Audio Input:**

Audio input is taken from the microphone of computer. The no. of channels involved are 1, in case of multiple channels the average is taken.

1. **Audio process:**  
   Audio processing involves the following steps:
2. **Sampling:**

Sampling is done at 16KHz which is sufficient for human voice according to Nyquist Criteria. Sample is taken with duration of 1 sec.

So, 16000 samples are taken in 1 sec.

1. **Windowing and framing:**

Windowing is done using hamming function. The length of each window is 25 ms which contains 400 samples and length of frames is 10 ms. So, 1 sec sample contains about 100 frames.

1. **Cepstral Coefficients:**

//Aayush le lekhne

1. **Neural Network:**

Normal neural network does not depend on the data of past events which is crucial in our case so we have used Recurrent Neural Network. As recurrent neural network in itself is not sufficient we have used LSTM which is capable of remembering data from past events also.

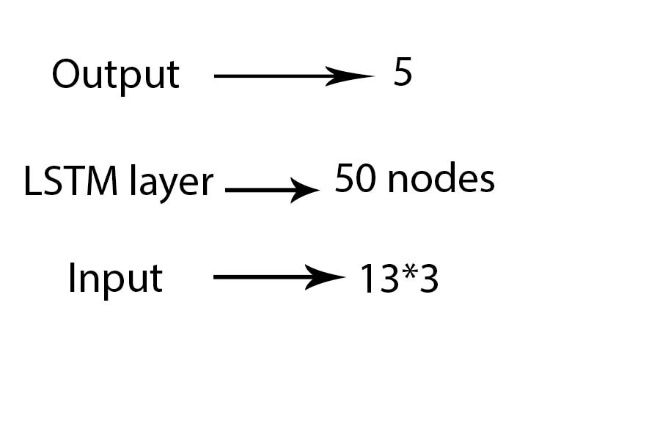


Figure above is the Neural Network layer involved in the audio processing step. Input given are 13 MFCCs, 1st derivative of MFCCs, and 2nd derivative of MFCCs so total inputs involved in input layer are 39. The 1st and 2nd derivative are taken considering the time dependency of the speech signal.

From the MFCCS given as the inputs to the LSTM layer, the feature set for the LSTM layer are determined and stored which depend upon the specific speech signals which in our case are the specific words to be recognized. LSTM were used for the specific purpose of developing feature set which depends upon a long sequence of past data which in our case are the frames of the speech that we passed. A large no of frames were passed sequentially to the network, and the output had dependencies with a large no of frames previously passed and this was modelled using LSTM. The output of the LSTM layer were again passed through a softmax layer of neurons so that the sum of the outputs don’t exceed 1. This made it easy for us to model the probability distribution as the sum of mutually exclusive probabilities also cannot exceed 1. The outputs of the softmax layer were made into an array and sent to a generating function to convert the values into an appropriate form for decision of final output. By training the networks and back-propagating through time for a fixed number of states the network learned the appropriate parameters to recognize the words that were spoken.

The output layer gives 5 possible output. Each output is equivalent to one of the 5 possible commands.

1 0 0 0 0 = Start

0 1 0 0 0 = Stop

0 0 1 0 0 = Left

0 0 0 1 0 = Right

0 0 0 0 1 = Back