Speech Recognition System with NN-Classifiers and Feature Extraction

Akash AJ [1], Gautam Vignesh [2], Himanshu Sood [3], Vikram Jain [4]
[1, 2, 3, 4] SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abstract – Speech Recognition is a technology that allow machine to identify and transcribe the spoken language. Via feature extraction, pattern are highlighted in the waveform using speech analysis techniques. By incorporating a Hidden Markov Model(HMM) into the system, classification of characteristics is possible as it is capable to encode spectral vectors. However this model is unable to percept verbal cues, such as intonation which are used in human interaction. This can be identified by way of deep neural networks through training in pitch and context analysis.

Index Terms – Speech Recognition, Deep neural networks, Feature Extraction.

1. INTRODUCTION

The functionality of a machine has usually been impaired for layman due to complicated lingo and technical indisposition. Thus for the integration of artificial intelligence into society, it is mandatory that a much more efficient way to communication with these intelligent systems apart from manual coding and inanimate interactions is available for the general public. This heeds the endeavor of speech recognition.

While speech recognition has been a daunting task for machines to replicate from the inception of the idea, its inability to be completely effective is due to the lack of intuition. This is because high inconsistency found in speech signal due to multiple variables such as differentiating various speakers, irregularities in speech rate and verbal cues required to understand the intonations and undertone of a speech.[1] Minor variations in speech pattern are found lead to semantic changes in its context through intonation for interrogation and sarcasm.

This variation in pitch can be observed by standard speech analysis techniques which can be used for developing patterns. The computational analysis of a speech signal for sound to text conversion is handled by feature extraction techniques such as Mel Frequency Cepstrum Coefficient(MFCC) which is discussed in detail later on.

Deep Learning is a fragment of machine learning that allows systems to automatically discern patterns needed for feature identification. Deep Neural Network are capable of assimilating a function without a function specific programming by means of multiple hidden layers from input to output with no backward looping[2]. Such deep learning architectures are motivated by existing biological neural networks.

2. RELATED WORKS

The earliest speech recognition systems were mere number recognition systems only capable of detecting string of digits with a vocabulary size of not more than ten[3]. These used spectral peaks in the power spectrum of the respective speech signal.

Although current methodology has the potential to determine speech signal from a digitized waveform, it is preferable to reduce the variability by pre-pending constraints on known factors such as noise, pitch, periodicity, [2]

Feature Extraction computes the distribution of power across the short term power spectrum which allows for key characteristics of the subject's speech to be observed through the noise.

With a wide range of speech analysis techniques currently available such as Linear Predictive Analysis(LPC), Perceptual Linear Prediction(PLP), First Order Derivative(DELTA)etc.[2], extraction of acoustical features is first processed then classified using Support Vector Machines(SVM) but are unable to grasp the context of the subjects speech.

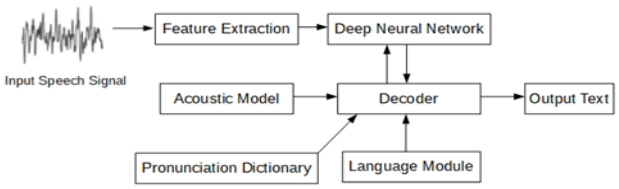
Gaussian Mixture Models(GMM) or Hidden Markov Models(HMM) provide a framework for the distribution of low level features where the hidden variables rather than being independent of each other, are interdependent through a Markov process[7].

Speech emotion recognition at the utterance level has allowed us to analyse each phoneme through deep neural networks and extreme machine learning[4]. This allows us to recognize pattern for pitch analysis multiple times in an utterance.

3. PROPOSED SYSTEM

The front end consists Recording with mic and can be processed using a standard computer sound card with sampling frequencies of about 8000Hz It is fed into a

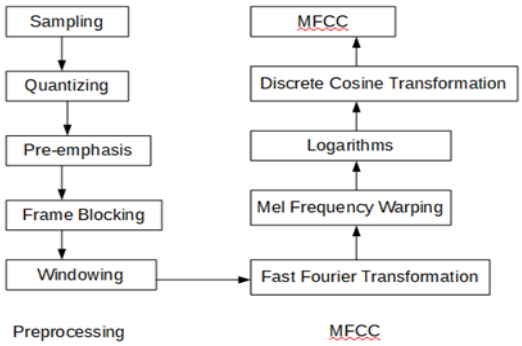
1. Architecture:



feature extractor where it is pre-processed and characterized through cepstum coefficient. The output is fed into a DNN before it is passed on to the decoder. The back end consists of matlab programming using FFT and DCT. These results are then fed to a decoder with matches feature classifiers through pattern recognition and is selected through distribution of the HMM for the state of each phoneme in the acoustic model. The decoder attempts to sequence the fabricated acoustic model with it respective phoneme. After matching, phoneme are sorted through professionally engineered pronunciation dictionary libraries from language experts. Finally it is processed by the language module for contextual meaning usually assorted using hidden layer networks. Before output of text the decoder verifies the final package through the DNN before displaying

2. Modules:

Preprocessing:



Feature extraction takes place at the utterance level, here the primary objective is to extract feature in each segment within an utterance and allows for integration into acoustic models after pre-processing. Thus a Segment-level Feature Extraction is designed as it helps better interpret speech signal by effectively collecting information from the waveform with minimum loss of data using smaller time frames. As speech signal deviates rather slowly, a short time spectral analysis technique(MFCC) is used. First, the input speech signal is sampled to reduce it to a discrete-time signal. This signal is mapped into smaller sets from a large number of input values by quantization. Noise removal is handled through a high frequency filter fitted through which the input signal must pass through. The signal is the divided into segments. Blocking arranges ten segments by overlapping to form a frame. These frames are windowed simultaneously to minimize spectral distortion due to discontinuities.

Hidden Markov Model:

Modelling the sequence of the short time power spectrum by using probabilistic functions to assign features through pattern recognition and identifying its acoustical equivalent is done in the DNN with HMM state output. A Markov chain of the distinct states in the input waveform is described using state transition probability matrix as such.

$$a_{ij} = \Pr(q_t = j \mid q_{t-1} = i), \qquad 1 \le i, j \le N \qquad a_{ij} \ge 0$$
$$\sum_{j=1}^{N} a_{ij} = 1$$

To avoid overloading the processor with the need to test too many test cases, we use Baum forward-backward technique to improve the computational efficiency to a more linear time scale. A training sequence is obtained through recorded simulations providing us the critical parameters for assortment and selection. Sound classes modelled for the phonemes and words are then optimized by evaluation of probability of the unknown utterance. The language and context and manufactured using Probabilistic Context Free Grammar(PCFG) and Chain rule

Deep Neural Networks

The DNN framework works layer by layer by beginning at the input layer, using logistic functions to map the state value.[5] Deviation of the target output from the actual output is measured using trained DNN classifiers. A directed graph using a rapid fine-tuning through weights by fitting a generative model to a single layer of feature detection. Large weights are used to prioritize smaller process paths and reduce over fitting. Unlike traditional speech recognition systems, MFCC is favourable compared to filter banks coefficients from the Mel frequency scale as they are mostly independent.

$$E(\mathbf{v},\mathbf{h}) = \sum_{i \in \text{vis}} \frac{(v_i - a_i)^2}{2\sigma_i^2} \quad - \sum_{j \in \text{hid}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}$$

The DNN must first be trained using a set of practice modules containing the various instances and test cases that the system is probable of encountering during its trial and actual testing to create a capable generative model in the hidden layer based on predictor variables from the training sets. The output provided by the DNN is the probability of the respective HMM relative to that of its corresponding acoustical input, p(HMM state/Acoustic Input) but before must first the be divided by the frequencies of the HMM state that are used in its fine tuning

before it can be acceptable to the forward-backward algorithm used in the HMM.

3. Techniques

Mel Frequency Cepstrum Coefficient:

The process of Feature Extraction is carried out using MFCC, imitating the cochlea of the human ear by filtering linearly below 1000KHz and logarithmic above 1000Hz as humans get less frequency selective above 1000hz[6]. Fast Fourier Transformations are performed on the signal for wavelet transformations to convert it from time domain to frequency domain. A Mel frequency scale is used for the non-linear pitch perception of the speech signal by mapping the spectrum through band pass filters prioritizing the power of each band over frequency band. The amplitude is measure using logarithms of the power band. Finally the log power bank is treated with Discrete Cosine Transformations, computing the Mel Frequency Cepstrum Coefficients respective to the required number, this is usually 13 coefficients.

$$c_{\tau,j}^{(4)} = \sum_{j=1}^{N_d} c_{\tau,j}^{(3)} \cos\left[\frac{k\,(2j-1)\pi}{2N_d}\right] \qquad k = 0, 1, ..., N_{mc} < N_d$$

Context Manipulation:

By using the various layers of the DNN, the PCFG[8] can be modified using the MFCCs through a type of neural network known as Back Propagation Network(BPN). It consists of 4 layers, the input layer uses predefined predictor variables for the pitch and periodicity required to sequence the spectra, describing the way one sound changes to another. The hidden layer applies the training data to match the value of this predictor variable to its respective test cases. The pattern layer uses the weight of targeted output type of each test case to calculate the total weighted vote for the pattern. The decision layer selects the state in the distribution by calculating the maximum weighted votes through comparison with weights of all necessary states for preferred pathway. Once a decision is made, the state of the PCFG is altered as per the training DNN receives in the practice module

4. Discussion

As the system is an intelligence agent, the practice sets and training modules fed into the classifier must better befit the criteria for pattern recognition in the hidden layer. This helps create for accurate and effective weighted graphs that are vital for efficient processing of the speech signal to their respective acoustic models. The identification of the various verbal cues that are related to alteration of context is also required to tabulate through numerous instance of their occurrence. It is necessary for the classifier that Mel scale warping is closely monitored so has to not hamper the pitch variance of the speech waveform as it will affect the acoustic characteristics extracted for feature classification.

4. CONCLUSION

Speech Recognition paves the way to a more convenient for Human Computer Interaction. By using such an expressive for of communication, there are numerous methods to describe a particular function or task. This is will greatly improve efficiency and can also provide various option from the systems database to choose from if the data is accordingly optimized. The possibility of embedding speech recognition systems in our everyday electronics can prove to set a better accessibility rate for most equipment, through better understanding of language and processing. This can lead to complete integration of Internet of Things in today's society without the need of coding and other technical lingo that is shrouded in a complicated interface to access its features. Although we are able to make machines better understand human intentions we are still missing all key features required to determine context such as general knowledge, facial expression and body language etc. and are yet to figure out how each aspects influences one or the other.

REFERENCES

[1] Urmila Shrawankar and Dr. Vilas Thakare "Techniques for feature extraction in speech recognition system: a comparative study" 2013
[2] Dhavale Dhanashri and S.B. Dhonde "Isolated Word Speech Recognition System Using Deep Neural Networks" 2017
[3] William C. Dersch "IBM Shoebox" 1960
[4] K Han, D Yu and I Tashev "Speech emotion recognition using deep neural networks and extreme machine learning" 2014
[5] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury "Deep Neural Networks for Acoustic Modelling in Speech Recognition" 2012.
[6] Anagha Sonawane, M.U.Inamdar and Kishor B. Bhangale "Sound based Human Emotion Recognition using MFCC & Multiple SVM" 2017
[7] B. H. Juang a and L. R. Rabiner "Hidden Markov Models for Speech Recognition" 2012
[8] Kristian Kersting, Luc De Raedt and Tapani Raiko "Logical Hidden Markov Models"2006

Authors

Akash James

UG Scholar

SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Vikram Jain

UG Scholar

SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Gautam Vignesh

UG Scholar

SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Himanshu sood

UG Scholar

SRM Institute of Science and Technology, Chennai, Tamil Nadu, India