Deep E.A.R.S.

Swaroop Mahadeva
ARRIS Group Inc

***Abstract*—** *This paper presents a **Deep** Learning based **E**nhanced **A**uditory **R**ecognition **S**ystem for a CPE device such as a Set top for a Television. A novel technique is proposed to automatically control the audio volume output of a device by enabling it to be contextually aware of its environmental surroundings using the acoustic information sensed through a microphone. The application of the proposed technique is envisioned for a smart home environment or for an ambient assisted living home like environment. This paper presents the implementation of the proposed technique with the deep learning model of Convolutional Neural Networks (CNN). The architecture wis an adaptation of an image processing CNN, programmed in Python and TensorFlow backend. The theoretical background that lays the foundation of the classification problem is briefly presented. Accordingly, from the experimented results it is shown that the model achieves a decent accuracy with the current state of art technology and promises much more for the near future.*

Keywords: *- Deep Learning, Convolutional Neural Networks, Artificial Intelligence, Human Activity Recognition, TensorFlow, Big Data, Discrete Fourier Transform.*

I. INTRODUCTION

Voice-user interface, or VUI, has exploded in popularity over recent years. They have encompassed every other mode of machine interaction, from those things that we tend to take for granted, like keyboards and the screens of our desktop computers, to technologies that are more complex, like the movement-based UI the Xbox Kinect is built upon. We're not in "Westworld" just yet but it's clear that voice actuated machine interactions are here to stay. The new generation of smart assistants – Alexa, Google, Watson and others – have their roots and learning algorithms tied into the present hotly pursued field known as "Deep Learning". The main research activity associated in the field of deep learning involves the advanced training of convolutional neural networks (CNN). CNN are one of the most impressive forms of Artificial Neural Network (ANN) architecture. Fig 1 shows a basic building block of an ANN. These biologically inspired computational

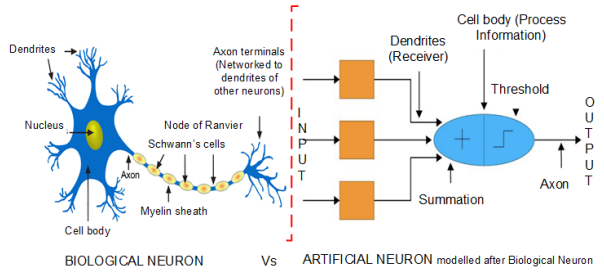

Figure 1: Basic building block of neural networks

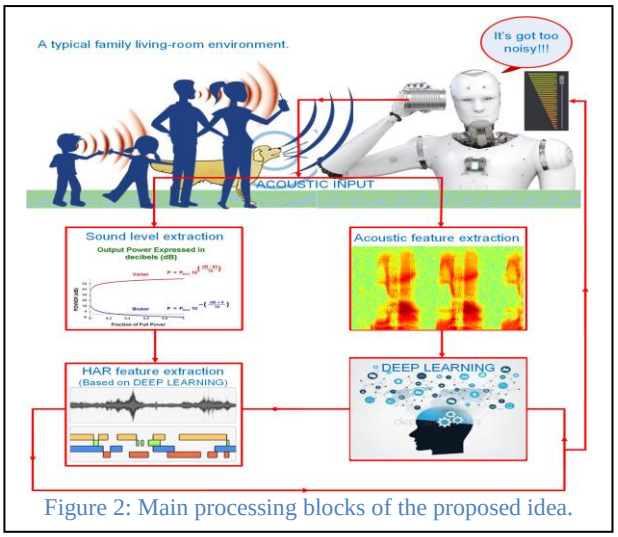

Figure 2: Main processing blocks of the proposed idea.

models can far exceed the performance of previous forms of artificial intelligence in common machine learning tasks. This has been triggered by a combination of the availability of significantly larger training datasets for Human Activity Recognition (HAR), thanks in part to a corresponding growth in "Big Data", and the arrival of new GPU-based hardware that enables these large data-sets to be processed in reasonable time-scales. Suddenly a myriad of long-standing problems in machine learning, artificial intelligence and computer vision have seen significant improvements, often sufficient to break through long-standing performance thresholds. HAR is the bases of ubiquitous computing in smart environments and a topic undergoing intense research in the field of ambient assisted living. HAR training data set includes acoustic data sources such as human voice, mobile ringtones, pet dog barking sound etc., associated with a typical living room setup. There has perhaps never been a better time to take advantage of the power of deep learning into the consumer products with the help of these training data sets and the ANN architecture.

In this paper, I present one such an opportunity for an enhanced living in a smart home setup and more importantly to help in solving the crucial problem for the assisted home living like environment to enable the elderly or physically disabled to live independently for a longer time. The envisioned idea is to address the below contextual scenarios: (a) Auto TV volume control by sensing the surrounding noise level to sustain the perceived TV audio output quality. (b) Perform video play/pause control or Audio mute control by recognizing that the TV viewer has shifted their attention away from TV to answer the phone call or having a conversation with other family members. In this paper, we introduce a recurrent neural network model for HAR. Figure 2 depicts the envisioned use-case scenario and the important processing blocks involved.

The paper is organized as follows. Section 2 briefs about a related work. Section 3 defines the abbreviations or acronyms used in this paper. Section 4 provides the background and motivation. Section 5 details the algorithm. Section 6 mentions the applicable use-case scenarios. Section 7 describes the proof of concept and experiment trials. Finally, Section 8 discusses the future scope and the conclusion.

II. Related Work

HCMLAB/VADNET - Real-time Voice Activity Detection in Noisy Environments using Deep Neural Networks (http://openssi.net)

VadNet is a real-time voice activity detector for noisy enviroments. It implements an end-to-end learning approach based on Deep Neural Networks.

III. Terminology

Term	Description
Deep learning	Also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms
Convolutional Neural Network (CNN)	In machine learning, a network is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptron designed to require minimal preprocessing.
Biological Neural Network (BNN)	A neural circuit, is a population of neurons interconnected by synapses to carry out a specific function when activated. Neural circuits interconnect to one another to form large scale brain networks.
TensorFlow	is an open-source software library for dataflow programming across a range of tasks. It is also used for machine learning applications such as neural networks.
Human Activity Recognition (HAR)	HAR involves continuous monitoring of human behaviors in the area of ambient assisted living, sports injury detection, elderly care, rehabilitation, and entertainment and surveillance in smart home environments.
Big Data	is a term used to refer to data sets that are too large or complex for traditional data-processing software to handle.
Mel-frequency cepstral coefficients	MFCC are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of audio clip.
Spectogram	is a visual representation of the spectrum of frequencies in a sound as a function of time.
Graphics Processing Unit (GPU)	A GPU is a computer chip that performs rapid mathematical calculations, primarily for the purpose of rendering images.
Westworld	Westworld is an American television fictional series about a technologically advanced Wild-West-themed amusement park populated by android "hosts".
Discrete Fourier Transform (DFT)	In mathematics, the DFT converts a finite sequence of equally-spaced samples of a function into a same-length sequence of equally-spaced samples of the transform, which is a complex-valued function of frequency.

IV. Background and Motivation

A. *The Confession*

Let me begin with a confession – there was a time when I didn't really understand deep learning. I would look at the research papers and articles on the topic and feel overwhelming. You can basically learn and practice a concept in two ways:
Option 1: You can learn the entire theory on a particular subject and then look for ways to apply those concepts. Robust but time taking approach.
Option 2: Start with simple basics and develop an intuition on the subject. Next, pick a problem and start solving it. Learn the concepts while you are solving the problem. Keep tweaking and improving your understanding. So, you read up how to apply an algorithm – go out and apply it. Once you know how to apply it, try it around with different parameters, values, limits and develop an understanding of the algorithm. I prefer Option 2 and take that approach to learning any new topic. I might not be able to tell you the entire math behind an algorithm, but I can tell you the intuition. I can tell you the best scenarios to apply an algorithm based on my experiments and understanding.

B. *The Goal*

The main objective of the proposed algorithm is to identify and detect simple and complex activities in real world settings using acoustic data captured through a microphone device. Noise in the data can be caused by humans or by other sources associated with a typical smart home environment, for example start and stop of human conversation, a person talking on a phone, whizzing noise from a vacuum cleaner, crying of a baby or simple barking of your pet dog. Such real-world settings are full of uncertainties and calls for methods to learn from data, to extract knowledge from it that helps in making decisions. The goal of this paper is to design an algorithm that can perform a real time human activity recognition based on the live acoustic information.

V. Algorithm

A. *Overview*

The proposed algorithm comprises of four main stages:
Stage 1: Acoustic Feature extraction.
Stage 2: HAR feature extraction.
Stage 3: Sound Level Extraction.
Stage 4: Decision tree classifier.

Firstly, the raw sound waves are sampled and grouped in small chunks of 23 milli-seconds time samples. These samples go through Discrete Fourier Transform (DFT) to convert itself to frequency domain. The frequency domain signal is then subjected to an acoustic feature extraction process. The log-scaled mel-spectogram output is the passed to the second stage of the algorithm that involves "Deep Learning". Once HAR feature is detected and extracted from the deep learning process, it is further passed on to the stage 3 to extract the sound level in DB value. Finally, in the stage 4 the algorithm is set for the decision tree classification. In this stage, a decision is made whether the sound level of the extracted feature exceeds the

decibel threshold level set by the algorithm for the decision making. If yes, then the mapped decision branch is executed and thus completes one cycle of the algorithm flow.

B. *HAR acoustic feature training data set*

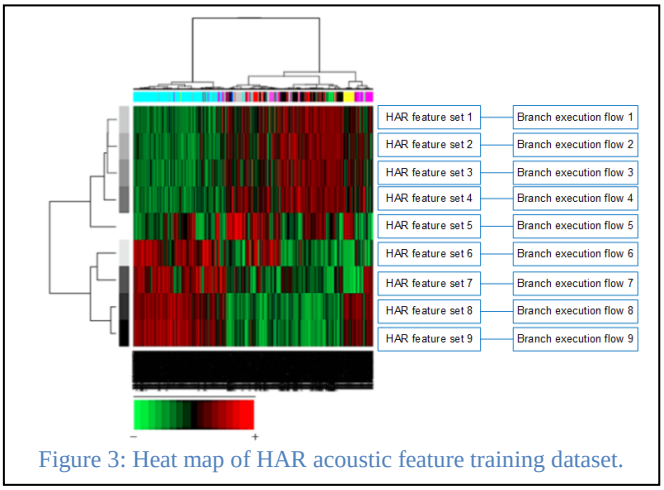

Figure 3: Heat map of HAR acoustic feature training dataset.

Our model is only as complex as our data, thus getting labelled 'data is very important in machine learning'. The complexity of the Machine Learning systems arise from the data itself and not from the algorithms. In the past few years we've seen deep learning systems take over the field of image recognition and captioning, with architectures like ResNet, GoogleNet shattering benchmarks in the ImageNet competition with 1000 categories of images, classified at above 95% accuracy (top 5 accuracy). This was due to a large amount of labelled dataset that were available for the Models to train on and also faster computers with GPU acceleration which makes it easier to train Deep Models. The problem we face with building a noise robust acoustic classifier is the lack of a large dataset, but Google recently launched the AudioSet - which is a large collection of labelled audio taken from Youtube videos (10s excerpts). Earlier, we had the ESC-50 dataset with 2000 recordings, 40 from each class covering many everyday sounds. The heat map of one such open frame work for HAR recognition model is shown in figure 3 with 9 features classification power.

C. *Design*

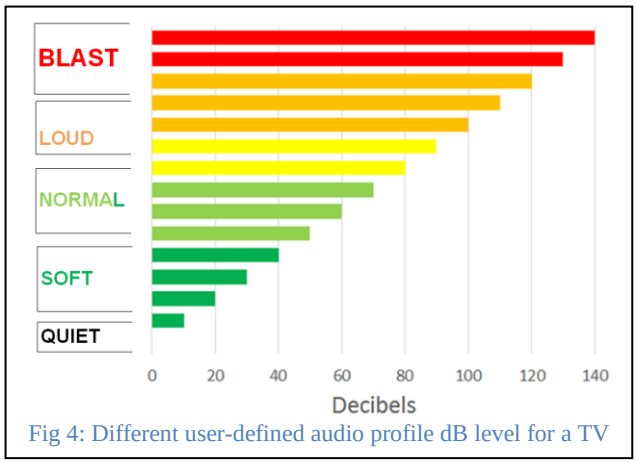

Fig 4: Different user-defined audio profile dB level for a TV

The proposed algorithm is designed for a Television viewer in a typical living room setup. The sensed acoustic information is assumed to be captured from the CPE device equipped with a microphone. The device itself can be imagined to be placed at a fixed location from a TV in a living room of dimensions say 20x20 square feet. The recorded acoustic information would consist of the living room noise and the noise generated from the Television itself. However, since the algorithm is designed to run on the CPE source device for the Television, it can estimate the sound decibel range outputted from the Television based on the current audio settings. Figure 4 shows the standard user defined audio profile and its associated dB range. The current active profile is one of the inputs for the algorithm to determine if the detected HAR event is from the TV noise or an external source. The HAR decision is made every 10 milliseconds. The HAR detection feature is programmed for a pre-configured threshold and is mapped to the corresponding user-defined function. IF differences > threshold for the detected HAR, return true; then mapped decision function is executed ELSE return false, and continue from the start of the flow.

D. *Detailed Pipeline flow*

The following section elaborates the detailed pipeline flow of execution in the following 6 steps.

STAGE 1, ACOUSTIC FEATURE EXTRACTION

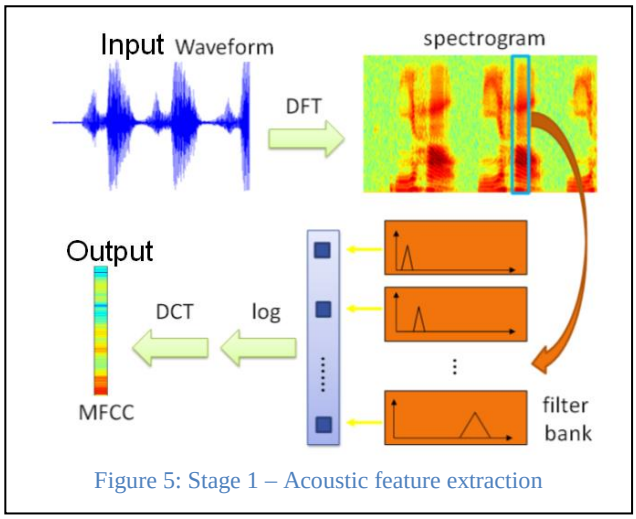

Figure 5: Stage 1 – Acoustic feature extraction

Although deep learning eliminates the need for hand-engineered features, we have to choose a representation model for our data. Instead of directly using the sound file as an amplitude vs time signal we use a log-scaled mel-spectrogram with 128 components (bands) covering the audible frequency range (0-22050 Hz), using a window size of 23 milli-seconds (1024 samples at 44.1 kHz) and a hop size of the same duration. This conversion takes into account the fact that human ear hears sound on log-scale, and closely scaled frequency are not well distinguished by the human Cochlea. The effect becomes stronger as frequency increases. Hence, we only take into account power in different frequency bands. The resultant audio output from this step is represented as a 128(frames) x 128(bands) spectrogram image. The Audio-classification problem is now transformed into an image classification problem

STAGE 2, HAR FEATURE EXTRACTION – DEEP LEARNING

Now that we have our audio in a format that's easy to process, we will feed it into a deep neural network. We use a convolutional Neural Network, to classify the spectrogram images. This is because CNNs work better in detecting local feature patterns (edges etc) in different parts of the image and

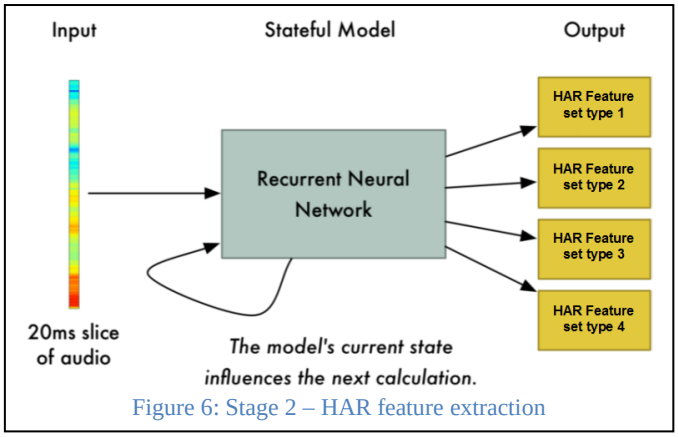

Figure 6: Stage 2 – HAR feature extraction

are also good at capturing hierarchical features which become subsequently complex with every layer. The input to the neural network will be the 23 millisecond audio chunks. For each little audio slice, it will try to classify to the trained HAR feature set. Figure 6 shows the recurrent neural network model that is used in the algorithm. Recurrent neural network - is a neural network that has a memory that influences future predictions.

STAGE 3, SOUND LEVEL EXTRACTION

In common usage, decibels are usually a way to measure the volume (loudness) of a sound. Decibels are a base 10 logarithmic unit, which means that increasing a sound by 10 decibels results in a sound that is twice as loud as the "base" sound. Several factors affect the noise level reading - The distance between the sensor and the source of the sound. The direction the noise source is facing, relative to the sensor. Noise reverberation based on surroundings etc., The sound level extraction stage is crucial for the proposed algorithm as this helps to determine if the identified HAR event is generated from the TV or an external source. The assumption here is that based on the extracted acoustic information from this stage the device is able to determine at any given point in time the dB level it is outputting and relate that to the measured instantaneous dB readings from its acoustic sensor.

STAGE 4, DECISION TREE CLASSIFIER

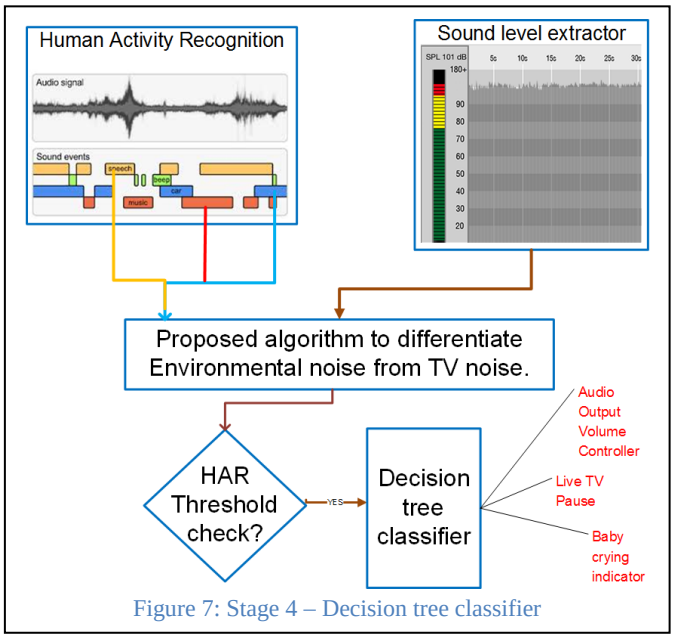

Figure 7: Stage 4 – Decision tree classifier

The Decision tree classifier stage makes the decision based on combined outcome of the threshold matching of the previous two extraction stages. For every HAR event identified the corresponding Sound dB threshold criteria must be satisfied, then branching decision is made to the corresponding execution flow for the associated HAR event feature set. Figure 7, represents logic flow for the decision tree classifier. An example branch flow for the user-defined automated control such as 'Auto volume control', 'Live TV pause' and a 'Baby crying indicator' is pictorially pointed out.

VI. APPLICATIONS

The applicability of the proposed algorithm can be used in a wide variety of fields that involves enhancing our machine interactions through voice which is current the highly researched topic. Our focus in this paper will be to demonstrate the same for an IPTV living home use-case scenario.

VII. EXPERIMENT AND RESULTS

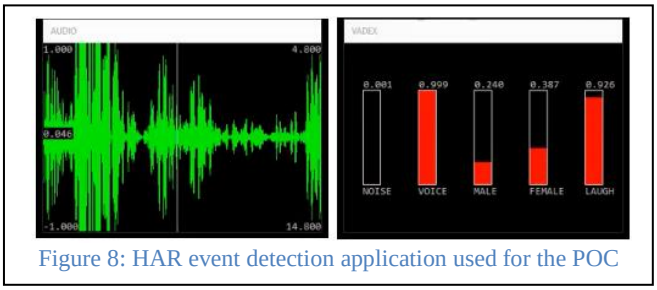

Figure 8: HAR event detection application used for the POC

As a proof of concept (POC), we tested the proposed technique to enable automatic volume control feature for an Android IP815 set top. The proposed algorithm was implemented using python script running on a laptop and acoustic data was captured from the laptop's in-built microphone. We experimented the use-case scenario for different audio content such as a music, talk show etc., The proposed user audio profiles– quiet, soft, normal, loud and blast and their associated dB levels were configured. The algorithm was successfully tested to automatically transition into different user profile based on the changing acoustic scenario. For ex: - A conversation detected event was mapped to 'quiet profile – 0dB' and for an externally induced noise such as the noise from dish washer, the transitioning happened to the 'loud'(80-100dB) and 'blast'(120-140dB) profiles respectively depending on the measured dB levels. Figure 8 shows the used application output identifying the different HAR events such as voice or noise, male or female voice, laugh etc., We also accounted the laugh indicator to make it extra smart so as to ignore the voice event if the laugh event is also identified for a conversation detected use-case scenario to prevent the false transitioning into the 'quite' profile.

VIII. FUTURE SCOPE AND CONCLUSION

The area of voice processing is undergoing intense research, but it is not yet perfected, however based on amount of the current research activity it is expected that the voice is going to be the preferred means of machine interaction in the near future.

IX. REFERENCES

[1] **HCMLAB/VADNET -** HTTP://OPENSSI.NET

[2] A comparative study of robustness of deep learning approaches for VAD - Sibo Tong, Hao Gu, Kai Yu

[3] Above the Din: Identification of Human Voice in Noisy Signals - A. Fried1, D. Stonestrom2, M. Stauber2