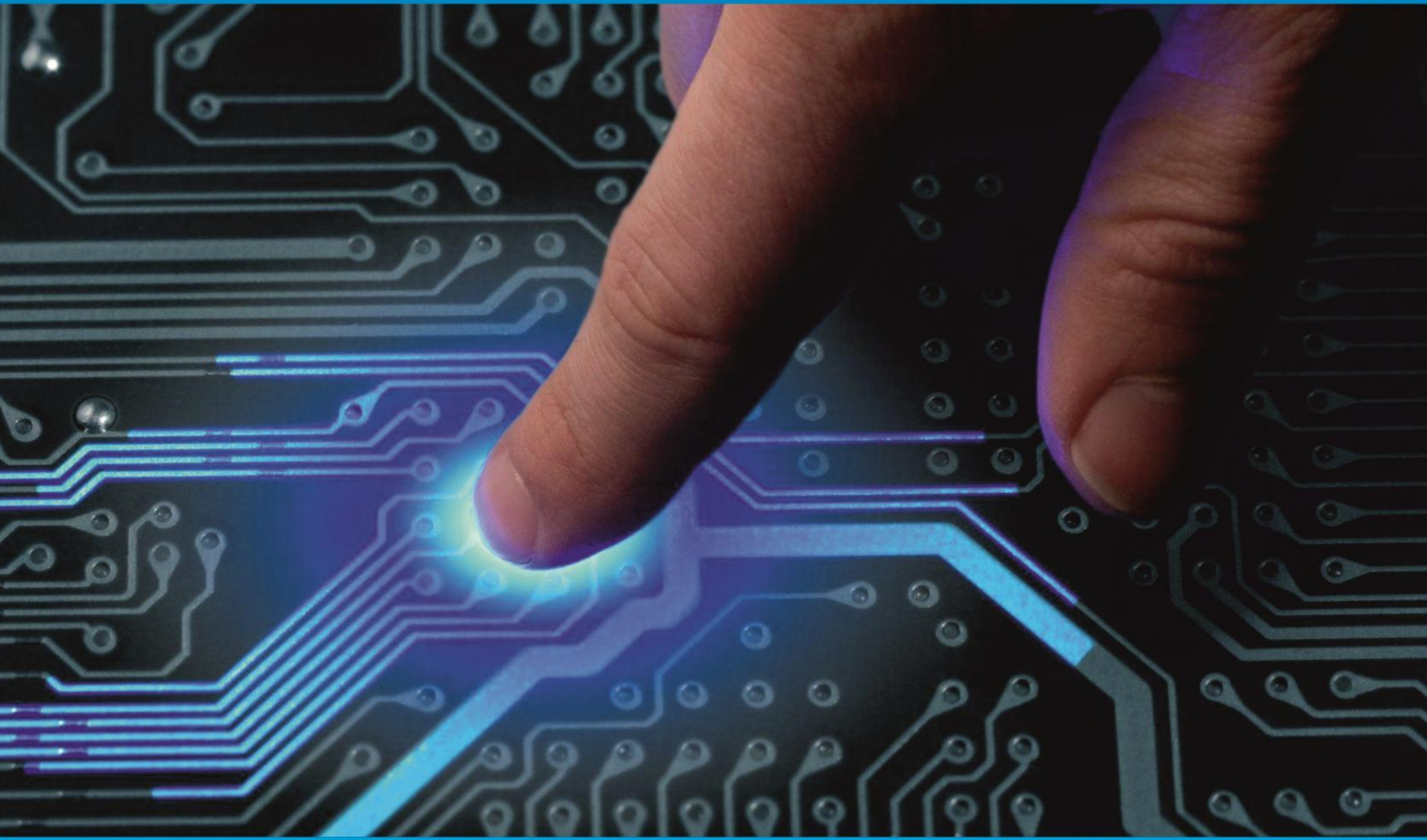
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 12, December 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

Emotion detection in speech using CNN

Pranav S,Yadhukrishna Madhu, Sachin Govind, KI Ajay Menon

UG Students, Dept. of Electronics., Model Engineering College, Thrikkakara, Kochi, Kerala, India

ABSTRACT: Speech is considered as the widest and most natural medium of communication. Speech is an information-rich signal that contains Para-linguistic information as well as linguistic information. Speech can convey a plethora of information regarding one's mental, behavioural, and emotional traits. Research on speech-emotion recognition exploiting concurrent machine learning techniques has been on the peak for some time. Numerous techniques like Recurrent Neural Network (RNN), Deep Neural Network (DNN), Spectral feature extraction and many more have been applied on different datasets. This paper presents a Convolution Neural Network (CNN) based speech emotion recognition system. A model is developed and fed with speech signals from a specific data set for training, classification, and testing with the help of high end Graphics Processing Unit (GPU).

KEYWORDS: Speech Analysis, Artificial Neural Networks, Convolution Neural Network, Mel Frequency Cepstral Coefficient

I. INTRODUCTION

Emotion recognition from a human speech is an attractive field of speech signal processing. It is drawing more attention in the applications where emotion recognition eases the speaker identification and mental status. It can be applied in fields such as criminal investigation, intelligent assistance and health care. Extracting the emotional state of a speaker from their speech is called speech emotion recognition. The speech emotion recognition involves analysis of the speech signal to identify the appropriate emotion based on its features. For feature extraction and testing of a speech signal a good number of algorithms have been formulated. Few of them are Artificial Neural Networks (ANN), Linear Prediction Cepstral Coefficients (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) and the Support Vector Machine (SVM). Most of the traditional machine learning algorithms and deep learning networks used for speech emotion recognition can only accept data with fixed dimensions as input. Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively minimized data rate for subsequent processing and analysis.

Human beings express their feelings, opinions, views and notions orally through speech. The speech production process includes articulation, voice and fluency. It is a complex naturally acquired human motor abilities. This is a task categorized in regular adults by the production of about 14 different sounds per second. There have been several successful attempts in the development of systems that can analyse, classify and recognize speech signals. Speaker recognition is the capability of a software or hardware to receive speech signal, identify the speaker present in the speech signal and recognize the speaker. Speaker recognition executes a task similar to what the human brain undertakes. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions.

The body of work on detecting emotion in speech is quite limited. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together. For a machine to understand the mindset or mood of the humans through a conversation, it needs to know who are interacting in the conversation and what is spoken. In virtual worlds, emotion recognition could help simulate more realistic avatar interaction. This paper presents an algorithm to identify various emotions present in a set of sound samples[1-2].

II. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network can do a lot of good things if they are fed with a bunch of signals, for instance, to learn some basic signals such as frequency and amplitude changes. A CNN network is shown in figure 1, since they are multi neural networks, the first layer is fed with this information and the second layer is fed with some recognizable features. To illustrate this, a signal of two-dimensional array of pixels is considered, it is a check board with each square on the board is either light or dark colour. By observing the pattern CNN decides whether it is a signal with frequency change or amplitude change.

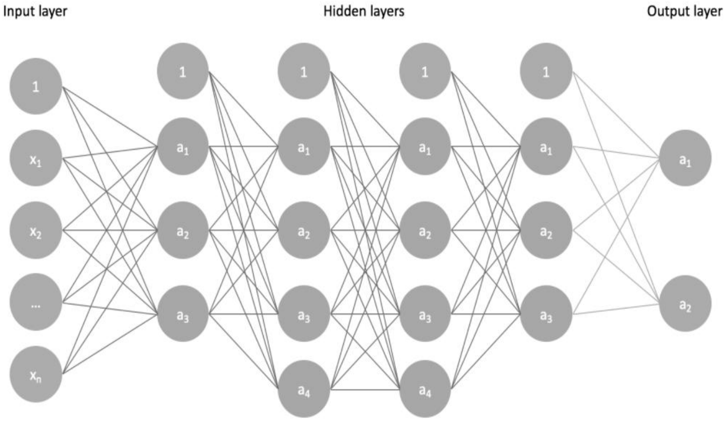


Figure 1

The convolutional neural network match the parts of the signal instead of considering the whole signal of pixels as it becomes difficult for a computer to identify the signal when the whole set of pixels are considered. The mathematics behind matching these is filtering. The way this is done is by considering the feature that is lined up with this patch signal. Then, one by one pixel are compared and multiplied by each other and then adds it up and divides it with the total number of pixels. This step is repeated for all the pixels that are considered. The act of convolving signals with a bunch of filters, a bunch of features which creates a stack of filtered images is called as convolution layer. It is a layer because it is operating based on stack that is in convolution one signal becomes a stack of filtered signals.

The next big part is called as pooling that is how a signal stack can be compressed. This is done by considering a small window pixel which might be a 2 by 2 window pixel or 3 by 3. On considering a 2 by 2 window pixel and pass it in strides across the filtered signals, from each window the maximum value is considered. It is then passed through the whole signal. At the end it is found that by considering only the maximum values the size of the filtered signal is reduced. The third part is normalization, in this if a pixel value is negative then the negative values are replaced with zeros. This is done to all the filtered signals. This becomes another type of layer which is known as a rectified linear unit, a stack of signals which becomes a stack of signals with no negative values. Now the three layers are stacked up so that one output will become the input for the next. The final layer is the fully connected layer.

The standard feed forward fully connected Neural network (NN) is a computational model composed of several layers. An input to a particular unit is outputs of all the units in the previous layer (or input data for the first layer). The unit output is a single linear regression, to which output value a specific activation function is applied. Convolution neural network (CNN) is a type of NN where the input variables are related spatially to each other. To take into account very important spatial positions, CNNs were developed. Not only they are able to detect general spatial dependencies, but also are capable of specific patterns recognition. Shared weights which represents different patterns, improve the convergence by reducing significantly the number of parameters. CNN recognize small patterns at each layer, generalizing them (detecting higher order and more complex patterns) in subsequent layers. This allows detection of various patterns and keeps the number of weights to be learnt very low [3].

III. PROPOSED ALGORITHM

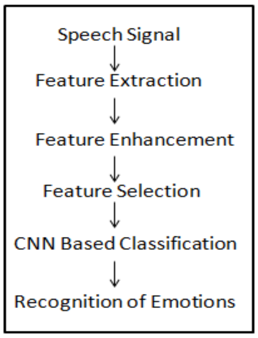


Figure 2

The proposed Algorithm is given in figure 2. The various components of the algorithm are,

A. *Speech Signal Dataset*

We have used RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contain 7356 files. The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry and fearful emotions. Each expression is produced at two levels of emotional intensity (normal and strong), with an additional neutral expression.

B. *Feature Extraction*

The time domain representation of sound is very complex, and in its original form it does not provide very good insight into key characteristics of the signal. Because of the characteristic of sound signal, we map this time domain representation into more telling features. The most straightforward technique involves determining the average energy of the signal. This metric, along with total energy in the signal indicates the volume of the speaker. Duration also offers insights into emotion, as do statistics like the maximum, minimum, range, mean, and standard deviation of both the signal and spectrum. These may indicate fluctuations in the volume or pitch that can be useful in determining emotion. For both the signal and spectrum, we also derive skewness, the measure of departure of horizontal symmetry in the signal, and kurtosis, the measure of height and sharpness of central peak, relative to a standard bell curve. In this method Librosa library in Python is used to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building 3 blocks necessary to create music information retrieval systems.

C. *Mel Frequency Cepstral Coefficient*

Mel frequency cepstral coefficients are computed on the basis of human hearing ability. In Mel frequency cepstral coefficients (MFCC) method two types of filter are used. Some filter is spaced linearly at low frequency below 1 kHz and other are spaced logarithmically at high frequency above 1 kHz. The block diagram of MFCC is shown in figure 3.

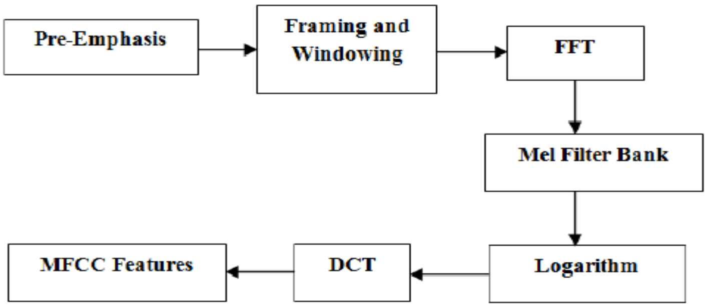


Figure 3

MFCC feature extraction process consists of a few steps as discussed below,

i. *Pre-emphasis*

Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increases the energy of signal. This increment of energy level gives more information. Framing, in this process, speech sample is segmented into 20-40 ms frames. The length of human voice may vary, so for fixing the size of speech this processes is necessary. Although the speech signal is non-stationary in nature (frequency can be changed over the time period), but for a short duration of time, signal behave like a stationary signal. Windowing, after framing process the windowing process is performed. Windowing function reduces the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a 10ms span. That means each frame contains some overlapping portion of previous frame.

ii. *Fast Fourier Transform (FFT)*

FFT is used to generate the frequency spectrum of each frame. Each sample of each frame converted from time domain to frequency domain by the FFT. It is used to find all frequencies present in the particular frame.

iii. *Mel scale filter bank*

This is a set of 20-30 triangular filters applied to each frame. The mel scale filter bank identify how much energy exists in a particular frame. The mathematical equation to convert the normal frequency ‘f’ to the Mel scale ‘m’ is as follows[4],

$$m = 2595 \log [1 + (f /700)]$$

iv. *Log energy computation*

After getting the filter bank energy of each frame, log function is applied to them. It is also inspired by human hearing perception. A human does not listen to loud volume on a linear scale. If the volume of the sound is high, human ear cannot recognize large variations in energy. Log energy computation gives those features for which human can listen clearly.

v. *Discrete Cosine Transformation (DCT)*

In the final step DCT is calculated of the log filter bank energies. In this proposed method, there are 25ms frames with 10ms of sliding. Also, 26 band pass filters. From each frame we computed 13 MFCC features. Energy of each frame should be calculated for further steps. After getting 13 MFCC features, in this process there are 13 velocity components and 13 acceleration components calculated using time derivatives of energy and MFCC.

D. *Energy*

The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific, energy associated with short term region of

| DOI: 10.15680/IJIRCCE.2020.0812037 |

speech. By the nature of production, the speech signal consist of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy. The short time energy of speech signals reflects the amplitude variation. The amplitude of the speech signal varies appreciably with time. In particular, the amplitude of the unvoiced segment is generally much lower than amplitude of the voiced segment. The short time energy of the speech signal provides a convenient representation that reflects the amplitude variation and can be defined as[4],

$$En = \sum_{m=-\infty}^{\infty} [x(m)\ W(n-m)]\ \text{^}2$$

E. *Zero Crossing Rate*

Zero Crossing Rate determines the information about the number of zero crossings present in a given signal. The concept behind zero crossing is to calculate how many times the signal waveform crosses the zero amplitude line by transition from a positive to negative or vice versa in a specific time. In mathematical terms, a ‘Zero Crossing’ is a point where the sign of a function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. Spontaneously if the numbers of zero crossings are more in a given signal, the signal will be changed rapidly which implies that the signal contains the high frequency information. Like the similar way, if the numbers of zero crossings are less. The signal will be changed slowly denoting that the signal contains low frequency information. The zero crossing of a signal can be depicted by the Figure 4. The zero crossing is also a technique which can be used to estimate the fundamental frequency of speech. The number of zero crossings per second is equal to twice the frequency of the signal. Therefore, we can say that ZCR gives indirect information about the frequency of the signal. The zero crossing rates are relatively high in unvoiced sounds compared to the zero crossing rates in the voiced sounds. The zero crossing rate value for the silence region should be zero.

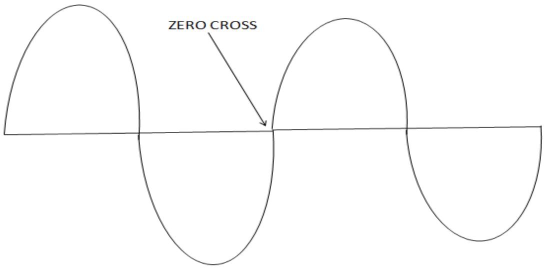


Figure 4

F. *Chroma*

Chroma feature, a quality of a pitch class which refers to the ‘color’ of a musical pitch, which can be decomposed in into an octave invariant value called chroma and a pitch height. Chroma features aim at representing the harmonic content (example: keys, chords) of a short-time window of audio. The feature vector is extracted from the magnitude spectrum by using a Short Time Fourier Transform (STFT) and Constant-Q transform (CQT).

G. *Feature Enhancement and Selection*

The extracted features are enhanced using windowing techniques, calculating the average energy. After processing the original sound signal to extract features, the high variance of the algorithm reveals the requirement to filter the many features to determine which contribute most to the classifier. The input speech signals are windowed, with approximately 72 windows per audio sample, and each of these windowed samples provides a total of 577 features. In total, 41,558 features can be extracted. This large number of features (much larger than the number of examples) results in a very high variance. The aim is to extract the most important features. Because of the large number of features, heuristics used to score each feature, rather than implement a brute force forward or backward search .

H. *Classifier*

A classification system is an approach to set each speech to a proper emotion class according to the extracted features from speech. There are different classifiers available for emotion recognition. Here Convolutional Neural Network (CNN)is proposed for classification. CNN The convolutional neural network (CNN) can be regarded as a variant of the standard neural network. Instead of using fully connected hidden layers, the CNN introduces a special network structure, which consists of alternating so called convolution and pooling layers. CNNs are current state of art models that are used to extract high level features from low level raw pixel information. CNN uses the numbers of kernels to extract high-level features from images and such features is used for training a CNN model to perform significant classification task. CNN architecture is a combination of three components; conventional layers, which contain some numbers of filters to apply on input.

Every filter scans the input using the dot product and submission method to produce the numbers of features maps in a single conventional layer. The second component is pooling layers, which is used for reducing or down-sampling the dimensionality of features maps. There are some schemes used for reducing dimensionality like max pooling, min pooling, mean pooling and average pooling. The last component is fully connected layers (FC) of CNN, which mainly used for extracting the global features that are fed to a SoftMax classifier to find out the probability for each class. A CNN arranges these all layers in hierarchical structure, convolutional layers (CL), pooling layers (PL), and then FC followed by the SoftMax classifier. CNNs have been shown, by extensive research, to be very useful in extracting information from raw signals in various applications such as speech recognition and image recognition.

IV. EXPERIMENT AND RESULTS

A. *Dataset*

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is used in this work. It consists of 2556 audio samples recorded by 24 actors (12 male actors and 12 female actors). It is an English audio dataset. The statements recorded are ‘Kids are talking by the door’ and ‘The dogs are sitting at the door’. These two statements are uttered in seven different emotions. Which are anger, fear, sad, disgust, surprise, happy and neutral. There are 344, 338, 264, 196, 209, 345, 356 audio samples pertaining to each emotions respectively.

B. *Basic Setup and Dependencies*

Google Colab can be used for performing the implementations. Google Colab is a free online cloud-based Jupyter notebook environment that allows to train and test machine learning and deep learning models on CPUs, GPUs, and TPUs. The dependencies used in this work are Python 3.7, Librosa, PyTorch, Keras and GPU. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. The librosa package is structured as collection of sub modules. PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing. Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible implementation. GPU is used to enable faster execution time.

C. *Data Preparation*

The entire dataset is divided into training set and test set by setting a split value. Here the split value is chosen as 0.8, which implies that of the 2556 audio samples, 80% will be taken as training samples and rest 20% will be test samples. Therefore the training set will have 2044 audio samples and the test set will have 512 audio samples. Next step is to specify the classes of emotions. Here it is taken as seven classes of emotions and they are labelled using numbers. The labelling is done as follows,

0 – Anger
1 – Disgust
2 – Fear
3 – Happy

4 – Neutral
5 – Sad
6 – Surprise

For the ease of operation the dataset is converted into CSV format. The CSV file contains the path of the audio files and their corresponding label.

D. *Data Pre-processing*

The feature extraction is done using the librosa package. The ‘get audio features()’ is used to extract the features. The features extracted are MFCC, pitch, energy and chroma. A total of 65 features can be extracted from each sample. The ‘get features data frame()’ is used to get the featured data frame. The labels are encoded using One Hot encoding technique. The one hot encoding technique converts categorical values into a one dimensional numerical vector. The resulting vector will have only one element equal to 1 and the rest will be 0.

E. *Model Creation*

The model type used is sequential, It is the easiest way to build a model in Keras. It allows building a model layer by layer. Our model has 4 layers, first layer consists of 256 neurons and the following layers consist of 128 neurons each. Size of the filter matrix is 5, which means the filter matrix will be of size 5x5. The activation function used is ReLu (Rectified Linear Activation). The activation is ‘softmax’, it makes the output sum up to 1 so the output can be interpreted as probabilities. The model will then make its prediction based on which option has the highest probability. RMSprop is the optimiser used to compile the model.

F. *Training and evaluation*

To train the model the fit() function is used with parameters, training data(x_traincnn), target data(y_train), batch size, number of epochs and validation data. The number of epochs is the number of times the model will cycle through the data. The more epochs the model runs, the more the model will improve, up to a certain point. After that point, the model will stop improving during each epoch. For this model the epoch is set as 370. After training, the test set was loaded to predict the emotions. It can be inferred that 6 out of 10 predictions are correct.

V. CONCLUSION

This work presents a novel, singular, convolution neural network based speech emotion recognition procedure that is a credible alternative to the other traditional methods. It can classify 7 types of emotion. It can be interpreted as a speech processing tool and can be trained to classify more types. Since it does not require any complex computational idea, it can be considered as user friendly and easily understandable. It is indeed a modern way to apply speech processing and convolution neural networks in practice for the betterment of technology.

REFERENCES

1. Srinivas Parthasarathy, Ivan Tashev, “CONVOLUTIONAL NEURAL NETWORK TECHNIQUES FOR SPEECH EMOTION RECOGNITION”, International Workshop on Acoustic Signal Enhancement (IWAENC2018), Sept. 2018, Tokyo, Japan.
2. Saad Albawi, Tareq Abed Mohammed, Saad Al-Zawi, “Understanding of a convolutional neural network”, IEEE Xplore, March 2018.
3. Nadia Jmour, Sehla Zayen, Afef Abdelkrim, “Convolutional neural networks for image classification”, IEEE Xplore, June 2018.
4. Saikat Basu, Jaybrata Chakraborty, Arnab Bag, Md. Aftabuddin, “A review on emotion recognition using speech”, IEEE Xplore, July 2017.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING