CONVOLUTIONAL NEURAL NETWORK FOR SPEECH EMOTION DETECTION

Ajay Meena
Electrical and Electronics Engineering
Koc University
Istanbul, Turkey

***Abstract*—Automatic speech recognition is an active field of study in artificial intelligence and machine learning whose aim is to generate machines that communicate with people via speech. Speech is an information-rich signal that contains paralinguistic information as well as linguistic information. Emotion is one key instance of paralinguistic information that is, in part, conveyed by speech. Developing machines that understand paralinguistic information, such as emotion, facilitates the human-machine communication as it makes the communication more clear and natural. In this report we implement a real-time Convolutional Neural Network model for speech emotion detection. This work utilizes Ryerson Audio-Visual Database of Emotional Speech and Song dataset comprised of 3 different emotions: *Happy*, *Sad* and *Angry*. The implmented model achieves an overall accuracy of 83%. This report also provide an indept model visualization and analysis of CNN filters. On the frequency domain the CNN filters distribute throughout all the spectrum range, with higher concentration around the average pitch range re-lated to that emotion. Each filter also activates at multiple frequency intervals, presumably due to the additional contri-bution of amplitude-related feature learning.**

***Index Terms*—Emotion Recognition, Convolutional Neural Network**

I. Introduction

Multimedia pattern recognition is an emerging technology that can extract and analyze large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Automatic emotion recognition has a direct application in the space of medicine and therapy. For those that have social communication disorders like Alexithymia, social-emotional agnosia, or even autism, emotions are very complex to understand and can often feel out of reach. Their inability to detect emotions limits their interactions with familiar people and they are commonly at risk of severely damaging interpersonal relationships. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in media. However, recognizing emotions from speech is a very challenging problem because people express emotions in different ways, and the features are unclear to distinguish the emotions. Actually, the paralinguistic problem is challenging even for humans.

In recent years people have delegated the role of learning emotional models to deep neural networks, which have superseded the state-of-the-art methods. Several neural network variants were developed that take as input traditional prosodic features , spectrograms or directly raw audio samples. The latter are the most promising, as they entirely eliminate all the overhead required for the feature extraction step, while yielding equally good or superior performance. Deep learning from raw audio is now replacing traditional feature-based learning in all speech-related tasks, with Automatic Speech Recognition the most prominent field.

In this report we implement a real-time, lightweight CNN model, able to process speech segments in a few hundred milliseconds on a low-end consumer notebook. Fast processing of speech signal is extremely important for the development of machine dialog systems able to instantly react to the user inputs, either acoustic, textual or visual. We then conduct an in-depth analysis of the model, showing where our emotion model activates in frequency. Such analysis represents another important step towards our goal to build a empathetic robot able to feel and react to emotions like a human would do. The architecture is programmed in Python using Keras and TensorFlow backend. The network was trained and subsequently tested with English language samples.

II. Relevant Work

While deep learning is being applied to many different tasks, often superseding former state-of-the-art methods, researchers are somehow ignoring the issue of what the model is actually learning. Some timid attempts to visualize the neural network activation have been proposed for very well-known tasks in image recognition, Natural Language Processing and ASR. Emotion detection is one of those tasks, where although people have replaced shallow classifiers with deep learning, there have not been enough attempts to understand what happens inside the DNNs. Being able to give proper interpretations is nevertheless an important challenge to tackle, in order afterwards to develop better and faster learning models

A. Dataset

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing 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, with an additional neutral expression. For the purpose of this implementation since we are only interested in classifying emotional state of happy, sad and angry dataset corresponding to the three emotional state are taken.

III. The Implemented Method

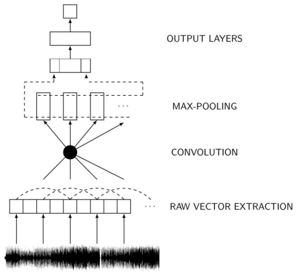

Fig. 1. Convolutional Neural Network architecture

The architecture is a lightweight framework and consist of one convolutional layer, one max pooling layer and one fully-connected layer. Due to the simplicity of framework the prediction of trained model is of the order of milliseconds and hence can be used in real time emotion prediction. In the following we would discuss about the implementation an functionality of each layer.

A. 1D Convolutional layer

The first layer is convolutional layer which consist of 1-dimensional 200 filters of length 200 each. All filters takes input as raw audio sampled at 22 kHz of arbitrary length. A convolution layer is run directly on the audio sample x. The output of convolutional layer with audio signal x as input is:

$$x_i^C = f(W_C x_{[i,i+v]} + b_c)$$

where v is the convolution window size and function f used is a Rectified Linear Unit(ReLU). We move the convolutional window with step size of 50. The role of this layer is to extract the features for each frame, and evaluate the differences among overlapping frames.

B. Max pooling layer

The convolutional output is passed to Max pooling layer. The max-pooling allows to select the contributions from the most significant frames, and to combine them into a fixed size vector.

$$x_j^{MP} = max_i(x_{i,j}^C)$$

where i is the window index, and j the vector index within each convolution window.

C. Fully Connected layer

The output of maxpooling layer is flattened so that it could be used as input for fully connected layer. The fully connected try to learn the high level features of speech signals. The output of this layer is given to a final softmax layer to perform the actual classification.

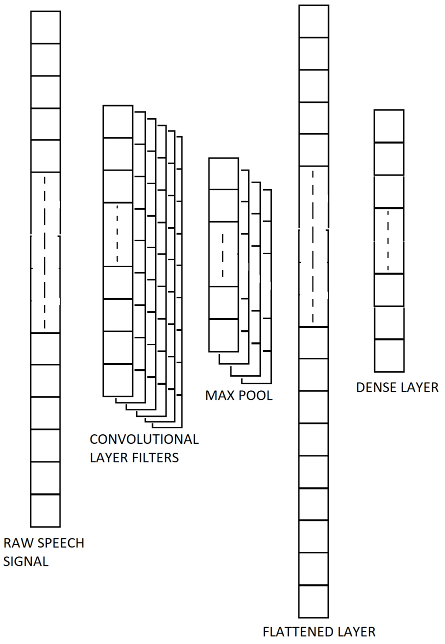

Fig. 2. CNN layers

IV. Experiment and Results

The network is trained for 50 epochs, using standard back-propagation, with momentum set to 0.9 and initial learning rate of 10^4.

Emotion Class	CNN accuracy in %	SVM accuracy in %
Angry	87.5	78.12
Sad	93.7	65.6
Happy	68.7	84.3
Overall	83.3	76.0

Table 1

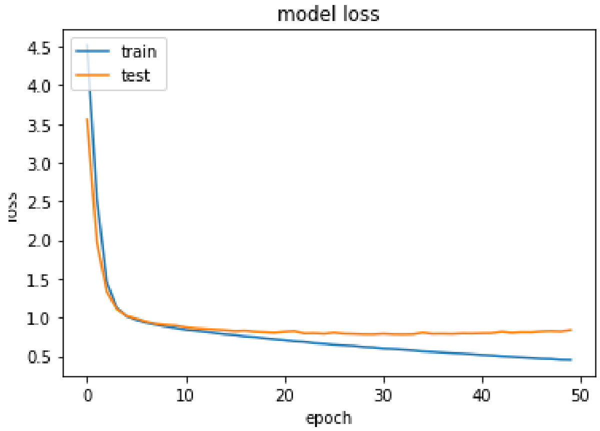

Fig. 3. Loss function

Average and class wise results for implemented SVM and CNN classifier are shown in Table 1. The average accuracy is

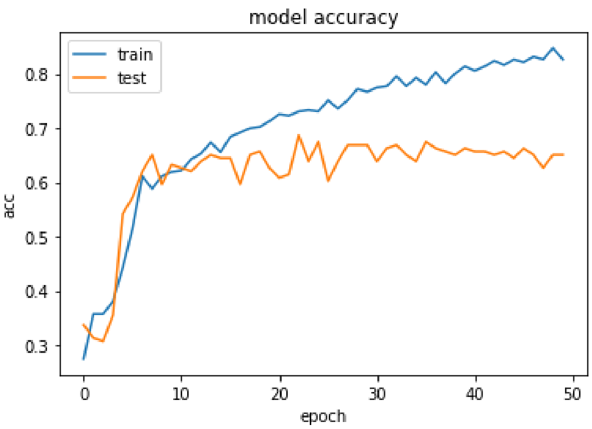

Fig. 4. Accuracy

around around 83% which is more than accuracy mentioned in paper, this may be due to different dataset which is used for training as well as evaluation. But the achieved accuracy is quiet good given the simple architecture and small dataset. The confusion matrix for SVM and CNN model are shown in figure 5 and 6 respectively. The graph for loss function and accuracy over epochs are shown above. It was observed that the result are affected by relative difference in the number of training samples of each emotional state.

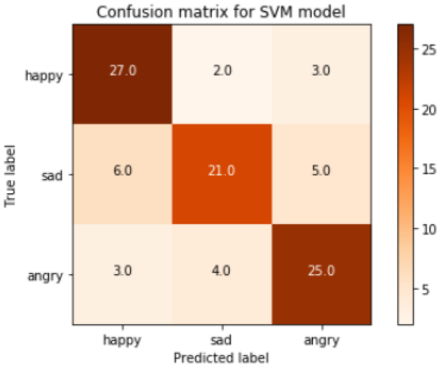

Fig. 5. Confusion Matrix for SVM model

V. Network Analysis

A. Frequency analysis

One of the aspect of visualization is where the network activates in the frequency domain. The first layer of our CNN is dedicated to features extraction and learning. Each row of the parameter matrix W^c is a filtering function which is applied to each convolution window. The contributions of all the filters (200 in our model) are then summed together. Each filter element $W^C_{i,j}$ is a time factor, spaced of the interval between one audio sample and the following of the discreet time input signal. Thus a filter can be easily converted to a frequency spectrum, taking the absolute values of the **FFT**:

$$F(W^C_i) = |FFT(W^C_i)|$$

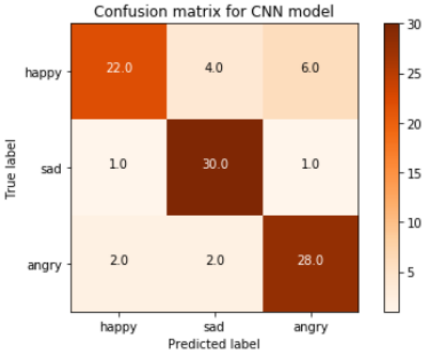

Fig. 6. Confusion Matrix for CNN model

$$A(i,j) = 20log_{10}(F(W^C_i)$$

Frequency response of some of the filters are in figure 8 From the frequency response we can observe that each filter act as band pass filter and try to learn low level features of speech segments.

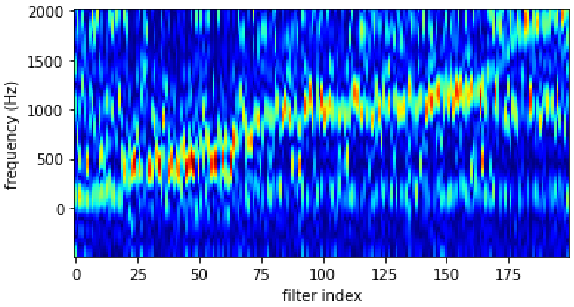

Fig. 7. Spectrum of the estimated filters in the convolutional layer

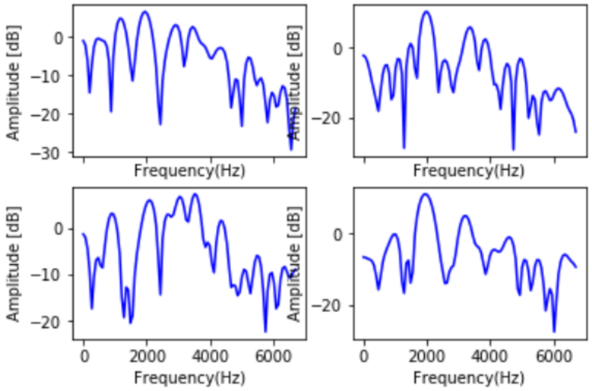

Fig. 8. Frequency response of filters

VI. Conclusion

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formalizations of machine learning problems driven by practical applications. However, recognizing real time emotions from speech is still a challenging problem. In this paper, we implemented the CNNs, SVM decision tree based networks without using any traditional hand-crafted features to classify emotional state of speech.

We compared the results of SVM decision tree model and CNN model. The CNN accuracy comes out to be 7% greater than SVM accuracy. The CNN model implemented uses light framework consisting of only a convloutuional, max-pooling and fully-connected layer, which reduces prediction time. We also provided a deeper analysis of the model activation in time and frequency. Each filter also activates at multiple frequencies in order to learn features. An accurate real-time emotion detection framework will be an important component of many speech-related application, in particular related to human-machine dialog systems.

REFERENCES

[1] Dario Bertero, Pascale Fung ”A First Look Into a Convolutional Neural Network For Speech Emotion Detection”

[2] Che-Wei Huang, Shrikanth S Narayanan ”Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition”

[3] Jongpil Lee, Jiyoung Park ”End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification”

[4] Livingstone, Steven R., Russo, Frank A. ”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)”